Doctoral Thesis in Speech and Music Communication

Spatially Grounded Communication in Embodied Agents

From Gesture Generation to Referential Understanding

ANNA DEICHLER

KTH ROYAL INSTITUTE OF TECHNOLOGY

Spatially Grounded Communication in Embodied Agents

From Gesture Generation to Referential Understanding

ANNA DEICHLER

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Monday, the 15th June 2026, at 10:00 in Sal F3, Lindstedtsvägen 26, Stockholm.

Doctoral Thesis in Speech and Music Communication
KTH Royal Institute of Technology
Stockholm, Sweden 2026

© Anna Deichler 2026

TRITA-EECS-AVL-2026:60
ISBN 978-91-8106-641-8

Printed by Universitetsservice US-AB, Sweden 2026

Abstract

When a person says “put that over there” while pointing at a shelf, the meaning depends on the spatial relationship between speaker, listener, and shared physical scene. Embodied agents that participate in such interactions must both produce spatially grounded gestures and interpret multimodal references. Yet these capabilities have largely been studied in isolation, with separate data, methods, and evaluation paradigms.

This thesis argues that gesture generation and referential grounding are two sides of the same communicative process, and that studying them jointly reveals structure that neither subfield uncovers alone. The argument is developed across seven papers. On the production side, contrastive speech-motion pretraining enables semantically aware co-speech gesture generation, while reinforcement learning with adversarial motion priors produces pointing gestures that are both spatially accurate and motorically natural, outperforming supervised baselines in a human referential identification study. A flow-matching architecture further combines semantic and spatial conditioning within a single generative system through distinct pathways.

On the comprehension side, the thesis introduces multimodal conversational datasets recorded in virtual reality and with wearable AR sensors, combining full-body motion, gaze, speech, and 3D scene context. Experiments show that state-of-the-art vision–language models fail on conversational references not for lack of perceptual capability, but because they cannot determine what is being referred to from underspecified language. A rewrite-based decoupling experiment isolates this bottleneck: once the referent is explicitly described, even simple detectors localize it accurately.

A central finding across both threads is that semantic reasoning, what is being communicated, and spatial reasoning, where it is directed, benefit from separate architectural treatment. On the production side, audio conditioning drives gesture timing while spatial targets determine direction; on the comprehension side, linguistic reasoning identifies the referent while visual perception localizes it. In both cases, architectures that maintain this separation outperform those that conflate heterogeneous signals into a shared representation. A shared data infrastructure, built incrementally across the papers, makes this parallel empirically testable: the same referential annotations that define conditioning targets for generation also define evaluation targets for grounding.

The thesis contributes methods, datasets, benchmarks, and evaluation protocols that support a unified view of spatially grounded communication in embodied agents, where producing and interpreting meaning are coordinated processes grounded in language, body, and shared physical space.

Sammanfattning

När en person säger “ställ det där borta” och samtidigt pekar mot en hylla beror betydelsen på den rumsliga relationen mellan talare, lyssnare och delad fysisk omgivning. Förkroppsligade agenter som deltar i sådana interaktioner måste både producera rumsligt grundade gester och tolka multimodala referenser. Trots detta har dessa förmågor till stor del studerats isolerat, med separata data, metoder och utvärderingsparadigm.

Denna avhandling argumenterar för att gestgenerering och referentiell grundning är två sidor av samma kommunikativa process, och att ett samlat studium av dem blottlägger struktur som inget av delfälten fångar på egen hand. Argumentet utvecklas genom sju artiklar. På produktionssidan möjliggör kontrastiv tal–rörelse-förträning semantiskt medveten generering av talackompanjerande gester, medan förstärkningsinlärning med adversariella rörelseprior producerar pekgester som är både rumsligt precisa och motoriskt naturliga och överträffar övervakade baslinjer i en perceptuell identifieringsstudie. En flödesmatchningsarkitektur kombinerar vidare semantisk och rumslig konditionering inom ett enda generativt system genom distinkta signalvägar.

På förståelsesidan introducerar avhandlingen multimodala konversationsdataset inspelade i virtuell verklighet, vilka kombinerar helkroppsrörelse, blickriktning, tal och 3D-scenkontext. Experiment visar att ledande bild–språkmodeller misslyckas med konversationella referenser inte på grund av bristande perceptuell förmåga, utan för att de inte kan avgöra vad som åsyftas utifrån underspecificerat språk. Ett omskrivningsbaserat frikopplingsexperiment isolerar denna flaskhals: när referenten beskrivs explicit lokaliserar även enkla detektorer den korrekt.

Ett centralt resultat som löper genom båda spåren är att semantiskt resonerande, vad som kommuniceras, och rumsligt resonerande, vart det riktas, gynnas av separat arkitektonisk behandling. På produktionssidan styr audiokonditionering gesternas timing medan rumsliga mål bestämmer riktningen; på förståelsesidan identifierar språkligt resonerande referenten medan visuell perception lokaliserar den. I båda fallen överträffar arkitekturer som upprätthåller denna separation dem som sammanför heterogena signaler i en delad representation. En gemensam datainfrastruktur, uppbyggd inkrementellt genom artiklarna, gör denna parallell empiriskt prövbar: samma referensannoteringar som definierar konditioneringsmål för generering definierar även utvärderingsmål för grundning.

Avhandlingen bidrar med metoder, dataset, riktmärken och utvärderingsprotokoll som stödjer en enhetlig syn på rumsligt grundad kommunikation i förkroppsligade agenter, där produktion och tolkning av mening är samordnade processer grundade i språk, kropp och delat fysiskt rum.

Popular Science Summary

Most people have experienced how AI language models like ChatGPT can hold a conversation in text. A natural next step is to give these systems a body, a robot or a virtual character that can see, move, and share a physical space with people. But conversation in a shared environment is fundamentally different from conversation on a screen. When someone says “put that over there” while pointing at a shelf, the meaning depends not just on the words, but on where the speaker is looking, what they are pointing at, and which objects surround them. This kind of spatial, gesture-rich communication is effortless for humans, yet remains surprisingly difficult for artificial agents. This thesis investigates how such agents can learn to both produce and understand it.

The work has two complementary parts. The first asks how a virtual character can generate natural-looking gestures, especially pointing, that accurately indicate objects in a shared environment. Although pointing appears simple, it requires precise coordination of timing, arm direction, and body posture. This thesis develops methods that learn such behavior directly from motion capture recordings of human interaction.

The second part asks how an agent can determine what a person is referring to during a conversation. In natural dialogue, people rarely use fully specified descriptions; instead of “the red mug on the second shelf,” they say “that one” or simply “it,” relying on shared context. This thesis shows that current AI systems often fail in these cases not because they cannot perceive the scene, but because they cannot resolve which object is intended. A simple experiment illustrates this: when a vague reference like “that one” is rewritten into an explicit description using the dialogue context, even basic object detectors can identify the correct target.

These two directions reveal a common structure. In both generation and understanding, reasoning about *what* is being communicated and reasoning about *where* it is located are distinct problems that benefit from being treated separately. The thesis supports this claim through new multimodal datasets recorded in virtual reality and with wearable smart glasses, capturing speech, body movement, eye gaze, and 3D scene context in a unified framework.

The results contribute toward agents that can participate naturally in spatial, gesture-rich interaction, understanding not just words, but the full context of pointing, gaze, and shared physical space in which meaning arises, and ultimately serving as intuitive, practical helpers in everyday environments.

Populärvetenskaplig sammanfattning

De flesta har upplevt hur AI-språkmodeller som ChatGPT kan föra en konversation i text. Ett naturligt nästa steg är att ge dessa system en kropp, en robot eller en virtuell karaktär som kan se, röra sig och dela ett fysiskt rum med människor. Men samtal i en delad miljö skiljer sig fundamentalt från samtal på en skärm. När någon säger "ställ det där borta" och pekar mot en hylla beror betydelsen inte bara på orden, utan på vart talaren tittar, vad de pekar på och vilka föremål som finns runt omkring. Denna typ av rumslig, gestbaserad kommunikation är enkel för människor, men förvånansvärt svår för artificiella agenter. Denna avhandling undersöker hur sådana agenter kan lära sig att både producera och förstå den.

Arbetet har två kompletterande delar. Den första frågar hur en virtuell karaktär kan generera naturliga gester, särskilt pekgester, som korrekt indikerar föremål i en delad omgivning. Även om det verkar enkelt att peka kräver det precis koordination av timing, armriktning och kroppshållning. Denna avhandling utvecklar metoder som lär sig sådant beteende direkt från rörelseinspelningar av mänsklig interaktion.

Den andra delen frågar hur en agent kan avgöra vad en person syftar på under ett samtal. I naturlig dialog använder människor sällan fullständiga beskrivningar; istället för "den röda muggen på andra hyllan" säger de "den där" eller helt enkelt "den," och förlitar sig på delad kontext. Denna avhandling visar att nuvarande AI-system ofta misslyckas i dessa fall inte för att de inte kan uppfatta scenen, utan för att de inte kan avgöra vilket föremål som avses. Ett enkelt experiment illustrerar detta: när en vag referens som "den där" skrivs om till en explicit beskrivning med hjälp av dialogkontexten kan även enkla objektdetektorer identifiera rätt mål.

Dessa två riktningar avslöjar en gemensam struktur. Inom både generering och förståelse är resonerande om *vad* som kommuniceras och resonerande om *var* det befinner sig distinkta problem som gynnas av att behandlas separat. Avhandlingen stödjer detta påstående genom nya multimodala dataset inspelade i virtuell verklighet och med bärbara smarta glasögon, som fångar tal, kroppsrörelse, blickriktning och 3D-scenkontext i ett enhetligt ramverk.

Resultaten bidrar till agenter som kan delta naturligt i rumslig, gestbaserad interaktion, som förstår inte bara ord utan hela kontexten av pekande, blickar och delat fysiskt rum i vilken mening uppstår, och i förlängningen kan fungera som intuitiva, praktiska hjälpare i vardagliga miljöer.

Acknowledgements

I would like to thank my supervisor, Jonas, for the support and for giving me the freedom to find my own research directions. This made it possible for me to find a topic I truly enjoyed working on throughout my doctoral studies.

I would like to thank Joakim, Simon, Johan, and Hedvig for providing me with feedback during my seminars. Joakim's enthusiasm for his work was an inspiration, as was Simon's deep knowledge of human motion and virtual agents. I am grateful to Iolanda for her support and feedback throughout my doctoral studies, and for introducing me to Irmak. I also thank Irmak for her perspective and knowledge of referential grounding, which shaped my thinking on this topic. Thanks to Éva, for being a mentor and a good friend.

Thanks to the coordinators of the courses I TA'd, especially André Holzapfel, for the opportunity to TA in the Human Perception for Information Technology course. I really enjoyed the freedom of creating fun projects with the students. I would like to thank all the students I worked with for their enthusiasm about their projects, which I found really motivating. Thanks also to all my collaborators, whose contributions were essential to this work, as well as everyone I met at conferences for the discussions and perspectives that shaped my thinking. Thanks to Rishabh and Anindita, I've had fun co-organizing the HSI workshop.

I would like to thank my office mates — Charlotte, Ghazaleh, Ayushi, and Lubos — we may not have successfully coordinated fika duties, but we managed to keep the office occupied around the clock. Thanks, Charlotte, for making me more neutral towards cats. Jim, for being a good friend and not noticing edquota. Livia, for the occasionally lengthy chats in front of our offices. Esther and Ambika, for their friendship and sharing creativity at *danskroki* sessions. Bahar, for making TMH's social life more lively. Bram, for his lectures on Central?/Eastern European culture and politics. Harm, for making people go to Max near dawn. Pablo and Julio, for always being ready to participate in experiments. Anya and Harsha, Palma was actually fun. Thanks to all for all the good times at the Doors.

And thanks to all the people at TMH (during my time): Agnes, Alireza, Ambika, Anders, André, Anna-Kaisa, Anya, Axel, Ayushi, Beatrice, Bahar, Birger, Bo, Bob, Bram, Carol, Carlos, Charlotte, Christina, David, Dimos, Dmytro, Elin, Erik, Esther, Éva, Fredrik, Gabriel, Ghazaleh, Giampiero, Gustav, Haotian, Harm, Harsha, Huanchen, Ilaria (tiszteletbeli), Jens, Jim, Joakim, Johan, Jonas, Joris, Juliana, Julio, Jura, Kajornsak, Katharina, Katya, Koji, Laura, Lia, Livia, Lubos, Luca, Marco, Nicolas, Nikita, Olov, Pablo O., Pablo P., Patrik, Per, Peter, Pieter, Rajmund, Ricardo, Ronald, Ruben, Shivam, Silvia, Simon, Siyang, Sten, Ulme, and Zofia.

Family and friends. I would like to thank my parents and Jutka for their constant support. Edina, for being the real doctor in the family; Timi, for finishing her DLA in May, solidifying my role as the perpetual csuporka of the three of us. Dóri, for the fun city trips to Stockholm and London (and Budapest). Dani, for all the feedback on all things visual. Rebeka, and others, for the sunny times. VV, for the support. SzCs, for grounding. Thanks to my hot yoga friends.

This work was supported by the Digital Futures project *Advanced Adaptive Intelligent Systems* (AAIS).

May 2026

Contents

Abstract v

Sammanfattning vii

Popular Science Summary ix

Populärvetenskaplig sammanfattning xi

Acknowledgements xiii

List of included papers xix

List of Abbreviations xxi

I Overview 1

1 Introduction 3
1.1 Research Questions 4
1.2 Thesis Overview and Paper Map 5
1.3 Contributions 6

Author contributions to the included papers 11

2 Background 15
2.1 Co-Speech Gesture 16
2.2 Referential Communication and Deixis 17
2.3 Pointing as a Foundation of Human Communication 18
2.4 Implications for Embodied Agents 19
2.5 From Communication to Computation 19

3 Modeling Human Motion 23
3.1 Motion Representations 24
3.2 Motion Tracking, Imitation and Reinforcement Learning 26
3.3 Adversarial Motion Priors 28

3.4 Diffusion and Flow Matching Models 29
3.5 Comparative Analysis 30

4 Conditioning Embodied Behavior 33
4.1 Conditioning Architectures 34
4.2 Post-training: Extending Pretrained models With New Conditioning Signals 36
4.3 Semantic Conditioning: Audio and Language 37
4.4 Spatial Conditioning: Target Position in Pointing Gesture Generation 41
4.5 Unifying Semantic and Spatial Conditioning 43
4.6 Semantic and Spatial Conditioning: MM-Conv-Flow 44
4.7 The Generality of the Conditioning Problem 50

5 Multimodal Grounding in Situated Dialogue 53
5.1 The Grounding Problem in Situated Dialogue 53
5.2 Vision-Language Models for 2D Grounding 54
5.3 Viewpoint, Gaze, and Spatial Representation 56
5.4 Dialogue Context and Nonverbal Cues 58
5.5 Decoupling Referent Identification from Localization 59
5.6 Connecting Production and Comprehension 62

6 Data for Situated Interaction 65
6.1 Existing Datasets and Their Limitations 65
6.2 The Pointing Gesture Dataset 70
6.3 Synthetic Augmentation of the Pointing Data 71
6.4 VR-Based Multi-modal Conversational Data Collection 73
6.5 Egocentric Data Collection with AR Glasses 80
6.6 From Data to Benchmarks 82

7 Evaluation 83
7.1 Evaluating Human Motion Generation Models 83
7.2 Evaluating Referential Grounding in Situated Dialogue 85

8 Conclusions and Future Directions 89
8.1 Summary of Contributions 89
8.2 Limitations 91
8.3 Future Directions 92

Bibliography 95

II Included Publications 105

Paper A: Diffusion-Based Co-Speech Gesture Generation 110

Paper B: Learning to Generate Pointing Gestures 122

Paper C: Spatial Awareness in Data-Driven Gesture Generation 138

Paper D: Grounded Gesture Generation 146

Paper E: A Benchmark for Referential Gesture Generation 160

Paper F: MM-Conv 170

Paper G: Look and Tell 188

List of included papers

This doctoral dissertation consists of two parts. The first provides an overview of the research field and a summary of my contributions. The second part comprises the included publications.

Paper A **Diffusion-Based Co-Speech Gesture Generation Using Joint Text and Audio Representation**, Anna Deichler, Shivam Mehta, Simon Alexanderson, and Jonas Beskow. In Proceedings of the 25th International Conference on Multimodal Interaction (ICMI), 2023

Paper B **Learning to Generate Pointing Gestures in Situated Embodied Conversational Agents**, Anna Deichler, Siyang Wang, Simon Alexanderson, and Jonas Beskow. In Frontiers in Robotics and AI, 2023

Paper C **Incorporating Spatial Awareness in Data-Driven Gesture Generation for Virtual Agents**, Anna Deichler, Simon Alexanderson, and Jonas Beskow. In Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (IVA), 2024

Paper D **Grounded Gesture Generation: Language, Motion, and Space**, Anna Deichler, Jim O'Regan, Téo Guichoux, David Johansson, and Jonas Beskow. In Humanoid Agents Workshop, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025

Paper E **A Benchmark for Scene-Aware Referential Gesture Generation**, Anna Deichler, Rishabh Dabral, Fethiye Irmak Doğan, Anindita Ghosh, and Jonas Beskow. In Workshop on Human–Scene Interaction at the European Conference on Computer Vision (ECCV), 2026

Paper F **MM-Conv: A Multimodal Dataset and Benchmark for Context-Aware Grounding in 3D Dialogue**, Anna Deichler, Jim O'Regan, Fethiye Irmak Doğan, Luboš Marcinek, Anna Klezovich, Iolanda Leite, and Jonas Beskow. In Proceedings of the Language Resources and Evaluation Conference (LREC), 2026
The version included here extends the LREC 2026 publication with VLM baselines and rewrite quality validation.

Paper G **Look and Tell: A Dataset for Multimodal Grounding Across Egocentric and Exocentric Views**, Anna Deichler and Jonas Beskow. In Workshop on SPACE in Vision, Language, and Embodied AI (SpaVLE), Neural Information Processing Systems (NeurIPS), 2025

Other contributions by the author not included in the thesis.

[1] **Scene-Aware Referential Gesture Generation via Spatial Flow Matching with Directional Refinement**, Anna Deichler, Rishabh Dabral, Fethiye Irmak Doğan, Anindita Ghosh, and Jonas Beskow. In Under review, 2026

[2] **PoseRefer: Pathway-Local Parameters for Semantically Grounded Reference Resolution**, Anna Deichler. In Workshop on Semantics for Reliable Robot Autonomy (SRRA) at IEEE International Conference on Robotics and Automation (ICRA), 2026

[3] **Fake it to make it: Using synthetic data to remedy the data shortage in joint multimodal speech-and-gesture synthesis**, Shivam Mehta, Anna Deichler, Jim O'Regan, Birger Moëll, Jonas Beskow, Gustav Eje Henter, and Simon Alexanderson. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024

[4] **Gesture evaluation in virtual reality**, Axel Wiebe Werner, Jonas Beskow, and Anna Deichler. In Companion Proceedings of the 26th International Conference on Multimodal Interaction, 2024

[5] **To smile or not to smile: The effect of mismatched emotional expressions in a Human-Robot cooperative task**, Ilaria Torre, Anna Deichler, Matthew Nicholson, Rachel McDonnell, and Naomi Harte. In 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2022

[6] **Towards Context-Aware Human-Like Pointing Gestures with RL Motion Imitation**, Anna Deichler, Siyang Wang, Simon Alexanderson, and Jonas Beskow. In Context-Awareness in Human-Robot Interaction: Approaches and Challenges, Workshop at ACM/IEEE International Conference on Human-Robot Interaction, 2022

[7] **Spatio-Temporal Priors in 3D Human Motion**, Anna Deichler, Kiran Chhatre, Christopher Peters, and Jonas Beskow. In Workshop on Spatio-temporal Aspects of Embodied Predictive Processing (StEPP) at IEEE ICDL, 2021

[8] **Mechanical Chameleons: Evaluating the Effects of a Social Robot's Non-Verbal Behavior on Social Influence**, Patrik Jonell, Anna Deichler, Ilaria Torre, Iolanda Leite, and Jonas Beskow. In Proceedings of SCRITA, Workshop at IEEE RO-MAN, 2021

List of Abbreviations

AdaLN	Adaptive Layer Normalization
AMP	Adversarial Motion Prior
CFG	Classifier-Free Guidance
CSMP	Contrastive Speech–Motion Pretraining
DDPM	Denoising Diffusion Probabilistic Model
DiT	Diffusion Transformer
ECA	Embodied Conversational Agent
FGD	Fréchet Gesture Distance
HMR	Human Mesh Recovery
HSI	Human–Scene Interaction
IoU	Intersection over Union
IK	Inverse Kinematics
MLP	Multi-Layer Perceptron
PFNN	Phase-Functioned Neural Network
PPO	Proximal Policy Optimization
RL	Reinforcement Learning
RoPE	Rotary Position Embedding
SGS-HSI	Single-target pointing dataset for HSI (§6.6)
SMPL-X	Skinned Multi-Person Linear Model, Extended
VAE	Variational Autoencoder
VLM	Vision-Language Model
VR	Virtual Reality

All photographs in this thesis are the author's own unless otherwise noted.

Cover illustration after Arany János's Toldi *(1846).*

AI language assistance (Claude, Anthropic) was used for text editing and proofreading during the preparation of this thesis. All scientific content, analysis, and conclusions are the author's own.

Overview

1 Introduction

Embodied agents operating in shared environments must produce and interpret communicative behavior that is grounded in physical space, not just fluent language. A request such as *put that over there* is meaningful only in relation to a speaker, an addressee, and a shared spatial context. In human interaction, speech is tightly coordinated with gaze, gesture, posture, and the structure of that scene. Throughout this thesis, this class of behavior is referred to as *spatially grounded communication*, multimodal communicative acts whose production is conditioned on, and whose interpretation requires, the spatial context of the shared environment.

The term *grounding* is used here in a deliberately broad sense. In dialogue and pragmatics literature, grounding refers to the process by which interlocutors establish common ground through mutual coordination [16]. In computer vision and natural language processing, visual grounding refers to linking linguistic expressions to regions or objects in a scene [17][1]. In motion and gesture generation, grounded behavior is motion that is anchored to specific spatial targets in a 3D environment. This thesis treats these as facets of a single underlying problem: anchoring communicative behavior in shared physical and interaction context. Production and comprehension both draw on the same grounding relation between language, body, and scene, and progress on either benefits from treating them as connected rather than separate.

Recent years have seen meaningful progress on several components of spatially grounded communication. Grounding systems have moved beyond flat images toward 3D representations, and multimodal dialogue systems increasingly account for spatial reasoning. However, these advances have largely developed along separate tracks. In gesture generation, early co-speech models focused primarily on rhythm and audio-driven motion, producing motion that is temporally aligned with speech but disconnected from semantic and spatial context. More recent work has begun to address semantic content, yet incorporating scene context into gesture generation remains an open problem. Grounding systems reason over visual scenes but rarely account for the embodied agent in the loop. Production and comprehension remain fragmented across subfields,

[1] Visual grounding can be seen as one operationalization of the broader symbol grounding problem [18], though this thesis does not engage with the philosophical question directly.

even though in human interaction they are two facets of the same communicative process. The same fragmentation is reflected in available data: no existing corpus provides synchronized speech, body motion, gaze, multi-turn dialogue, and 3D scene context within the same recordings.

This thesis argues that the communicative behavior of embodied agents should be studied through the lens of spatially grounded communication. The work therefore follows two connected threads, supported throughout by new data collected as part of the research. The first investigates how agents can *produce* gestures that are semantically appropriate and spatially anchored in a three-dimensional environment. The second investigates how agents can *understand* multimodal references by integrating language with gesture, gaze, and visual context. Together, these threads move from isolated gesture synthesis and disembodied reference resolution toward situated communication, where meaning is negotiated through the coordination of language, body, and environment. Addressing this fragmentation requires first advancing each thread on its own terms before the connection between them can be made concrete. The thesis therefore poses two production-side questions, one on semantic conditioning, one on spatial grounding, before turning to comprehension, and allows the cross-cutting patterns to emerge from the shared problem domain rather than imposing a joint framework from the outset.

1.1 Research Questions

This thesis is organized around three research questions.

RQ1: Semantic Gesture Generation

How can data-driven gesture generation models produce co-speech gestures that reflect the semantic content of speech?

Human gestures are not random accompaniments to speech; they are shaped by discourse structure, emphasis, rhythm, and meaning. Early models focused on prosodic alignment, producing motion largely disconnected from semantic content. Addressing this question requires models that align language and speech signals with motion in a shared representation. This line of work is primarily addressed in Paper A. Paper A is the sole contributor to this question because its role in the thesis is foundational rather than self-contained: the contrastive speech–motion representation it introduces provides the semantic conditioning mechanism that the spatially grounded generation work in Section 4.6 subsequently builds upon and extends with spatial targets.

RQ2: Spatially Grounded Gesture Generation

How can generative models produce pointing gestures that are both spatially accurate with respect to 3D targets and natural in their motion dynamics?

Pointing is a canonical communicative act because it links bodily movement to a specific target in shared space. This question is addressed through reinforcement learning, adversarial motion priors, synthetic data augmentation, and diffusion-based generation in Paper B–Paper E. Across these works, the central challenge is to preserve natural human motion while enabling precise spatial conditioning.

RQ3: Multimodal Referential Grounding

How do spatial representations and multimodal context affect the grounding of referring expressions in spontaneous situated dialogue?

The third question addresses the comprehension side of situated communication. Resolving referential ambiguity requires access to contextual visual information and to nonverbal cues such as gaze and pointing. This question is explored in Paper F and Paper G through multimodal datasets and grounding benchmarks built in egocentric and interactive 3D settings.

Taken together, these questions define a progression from semantic gesture generation, to spatially grounded gesture production, to multimodal referential understanding. The overall thesis claim is that these should not be treated as isolated subfields, but as parts of a unified research agenda on spatially grounded communication in embodied agents.

1.2 Thesis Overview and Paper Map

The thesis is structured in two main parts. The first part focuses on *generation*: how embodied agents can produce communicative motion conditioned on speech, language, and spatial targets. The second part focuses on *understanding*: how agents can resolve references in situated interaction by combining linguistic and nonlinguistic signals.

Paper A–Paper E trace the progression of the generation thread. Paper A introduces semantic conditioning for co-speech gesture generation through a contrastive speech–motion pretraining framework coupled with a diffusion model. Paper B investigates pointing behavior through reinforcement learning with adversarial motion priors, emphasizing physical plausibility and referential accuracy. Paper C addresses the data bottleneck by augmenting gesture corpora with synthetic spatial references. Paper D introduces a multimodal dataset and

framework that brings spatial targets and co-speech motion into a shared format, establishing the data foundation for unified grounded generation. Paper E brings these lines together at the level of benchmark formulation: it joins semantic and spatial conditioning into a single evaluation problem, when to gesture, where to point, and what to indicate, with a paired data release (MM-Conv and SGS-HSI), a two-track task definition, and a spatio-temporal grounding protocol that decomposes referential gesture quality into temporal alignment, spatial accuracy, and referent recall. OmniControl-PT (Paper E) and MM-Conv-Flow (Section 4.6, [8]) serve as reference baselines, instantiating the unification architecturally and establishing an initial performance reference for the challenge.

Paper F and Paper G extend the thesis toward referential understanding. Paper F introduces MM-Conv, a multimodal benchmark for context-aware grounding in 3D dialogue, recorded in VR with synchronized speech, full-body motion, gaze, and 3D scene geometry. The same paper introduces a two-stage grounding pipeline that decouples linguistic disambiguation from visual localization, and reports a human evaluation that quantifies the cost of missing context for conversational reference. Paper G presents Look and Tell, an egocentric multimodal dataset that captures speech, gaze, and visual context from wearable AR sensors, enabling analysis of grounding across egocentric and exocentric viewpoints. They shift the thesis from controlled generation experiments to naturalistic interaction data, introducing the multimodal corpora, annotation schemes, and stratified evaluation protocols needed to diagnose where current grounding systems fail and why. These papers show that understanding reference in situated dialogue requires reasoning over the same multimodal signals (language, gesture, scene context) that the generation chapters condition on. MM-Conv in particular bridges the two threads, a connection developed in Section 1.3.

Conceptually, the thesis moves from modeling communicative behavior independently of its physical setting to modeling it in relation to a shared environment. Methodologically, it moves from isolated motion modeling to structured multimodal conditioning. Empirically, it builds on a sequence of datasets, benchmarks, and evaluation protocols designed to study communicative behavior in increasingly realistic interactive settings.

Figure 1.1 summarises this problem space and the placement of the thesis contributions within it.

1.3 Contributions

This thesis makes four connected contributions, across modeling, data, evaluation, and the perspective that unifies them.

First, it contributes modeling approaches that reveal a structural separation between semantic and spatial conditioning of communicative motion. Con-

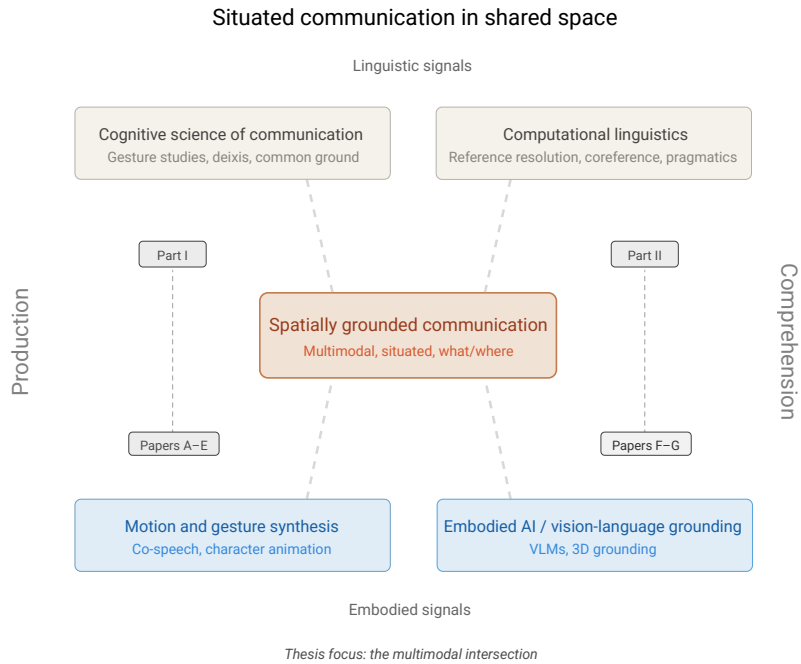


Figure 1.1: The problem space of spatially grounded communication. The thesis sits at the intersection of four research traditions—the cognitive science of communication, computational linguistics, motion and gesture synthesis, and embodied AI—organised along a linguistic/embodied axis (rows) and a production/comprehension axis (columns). The production side (Papers A–E) addresses gesture generation; the comprehension side (Papers F–G) addresses referential grounding.

trastive speech–motion pretraining (Paper A) shows that semantic conditioning of co-speech gesture benefits from a representation jointly aligned across language and motion rather than from acoustic features alone. Reinforcement-learning approaches with a phase-conditioned adversarial discriminator (Paper B) show that spatial conditioning for pointing can be implemented in a physics-based control loop where the simulator carries the corrective feedback, allowing the policy itself to remain architecturally minimal. The MM-ConvFlow system (Section 4.6) adopts this separation as a design principle, combining semantic and spatial signals in a single generative architecture through distinct conditioning pathways, with the spatial signal restricted to an additive residual stream that never modulates the audio-driven temporal computation. Across these papers, the recurring pattern is that semantic conditioning (what to gesture) and spatial conditioning (where to direct gesture) impose different

temporal and geometric demands and benefit from separate architectural pathways, a split that reappears on the comprehension side as a split between identification (*what*) and localization (*where*).

Second, the thesis contributes a data infrastructure that is designed as a sequence rather than as a set of independent corpora. The motion-capture pointing dataset (Paper B) provides the spatial targets absent from all existing co-speech corpora; the augmented pointing corpus (Paper C) scales and enriches it with naturalistic referring expressions through synthetic voice cloning, and the same synthesis framework produces the SGS-HSI single-target pointing clips that are later used in the Paper E benchmark. The unified gesture-and-scene format (Paper D) pairs spatial targets with co-speech motion trajectories; and MM-Conv (Paper F), recorded in VR with full-body motion, gaze, speech, and 3D scene geometry, closes the loop. Each dataset was designed to address the specific gap left by the previous one. MM-Conv in particular plays a dual role: its referential annotations simultaneously define conditioning targets for grounded gesture generation in Paper E and evaluation targets for grounding in Paper F itself, providing the empirical mechanism by which production and comprehension can be studied within a single research programme rather than across disjoint corpora.

Third, the thesis contributes evaluation paradigms that assess communicative function, not only motion quality, and the benchmarks that operationalize them. Standard distributional motion metrics are insufficient for spatially grounded behavior because they cannot tell whether a gesture is correct with respect to its intended referent. The VR-based referential game (Paper B) measures referential accuracy directly through identification by human observers, replacing quality ratings with a functional success measure. The spatio-temporal grounding protocol (Paper E) decomposes scene-aware referential gesture generation into temporal alignment, spatial accuracy, and referent recall, and packages these alongside MM-Conv and SGS-HSI as a community-facing benchmark and challenge. Expression-type stratification on MM-Conv (Paper F) makes the difference between full noun phrases, partitives, and pronominals visible where aggregate accuracy would conceal it; the rewrite-based decoupling experiment in the same paper separates the linguistic and perceptual components of grounding and identifies where current vision–language models actually fail. Look and Tell (Paper G) supports paired evaluation of spatial representation choices across egocentric and exocentric views. Each is a diagnostic that ties an evaluation number to a specific aspect of communicative success; together they form a benchmark series that turns the thesis's data infrastructure into resources the community can target directly.

Finally, the thesis contributes a research perspective that is grounded in a specific empirical observation: the same referential annotations in MM-Conv simultaneously define conditioning targets for gesture genera-

tion Paper E and evaluation targets for grounding Paper F, making the production–comprehension parallel empirically testable rather than rhetorical. The semantic–spatial split that organizes the generation work and the identification–localization split on the comprehension side are not two coincidentally similar findings but instances of the same underlying structure. The same architectural lesson holds in both directions: maintaining modality-specific computation outperforms forcing it through a shared representation. By developing both threads on this shared data infrastructure, the thesis supports the broader argument that embodied agents should be studied as systems that produce and interpret communicative behavior in coordination, not as separate generation and grounding pipelines.

Together, these contributions support a single overarching argument: that gesture generation and referential grounding are not isolated subproblems but two sides of the same communicative process, both requiring coordinated reasoning over language, body, and shared physical space. The progression from semantic gesture generation (Paper A), through spatially grounded production (Paper B–Paper E), to multimodal referential understanding (Paper F and Paper G), is not a sequence of unrelated projects but a cumulative case for this view.

Author contributions to the included papers

This section summarises my contributions to each of the included papers. Author initials are used throughout: S.A. (Simon Alexanderson), J.B. (Jonas Beskow), R.D. (Rishabh Dabral), A.D. (Anna Deichler), F.I.D. (Fethiye Irmak Doğan), A.G. (Anindita Ghosh), T.G. (Téo Guichoux), D.J. (David Johansson), A.K. (Anna Klezovich), I.L. (Iolanda Leite), L.M. (Luboš Marcinek), S.M. (Shivam Mehta), J.O'R. (Jim O'Regan), S.W. (Siyang Wang).

Paper A

Diffusion-Based Co-Speech Gesture Generation Using Joint Text and Audio Representation. This paper introduces a contrastive speech–motion pretraining (CSMP) module that learns a joint embedding of speech text and motion, used as a conditioning signal for a diffusion-based co-speech gesture generation model.

A.D. idea formation, implementation, training models, evaluation, manuscript writing; S.M. idea formation, implementation, training models, evaluation, manuscript writing; S.A. training models, technical discussions, manuscript editing; J.B. supervision, implementation, manuscript editing.

Paper B

Learning to Generate Pointing Gestures in Situated Embodied Conversational Agents. This paper proposes a reinforcement learning approach for generating pointing gestures in a humanoid agent, combining motion imitation with a goal-directed reaching objective so that the policy produces natural pointing motions that are also referentially accurate. The approach is evaluated both geometrically and through a human referential identification study.

A.D. idea formation, implementation, training models, evaluation, manuscript writing; S.W. implementation, training models, evaluation, manuscript writing; S.A. technical discussions, manuscript editing; J.B. supervision, manuscript editing.

Paper C

Incorporating Spatial Awareness in Data-Driven Gesture Generation for Virtual Agents. This paper addresses the absence of spatial referential behavior in existing co-speech gesture corpora by augmenting a standard dataset with synthetically generated pointing references, and shows that a generative model trained on the augmented corpus produces more spatially appropriate gestures than one trained on the original data.

A.D. idea formation, implementation, evaluation, manuscript writing; S.A. manuscript editing; J.B. implementation, supervision, manuscript editing.

Paper D

Grounded Gesture Generation: Language, Motion, and Space. This paper introduces a multimodal dataset and a baseline approach for generating co-speech gestures that are grounded in 3D scene context, building on the OmniControl spatial control mechanism to inject sparse spatial targets into a pre-trained motion diffusion backbone.

A.D. idea formation, dataset design, training models, manuscript writing; J.O'R. data processing, manuscript editing; T.G. data processing, training models, evaluation; D.J. visuals and renderings; J.B. supervision, manuscript editing.

Paper E

A Benchmark for Scene-Aware Referential Gesture Generation. This paper establishes evaluation infrastructure for spatially grounded gesture generation, comprising a paired dataset release with pointing annotations, a two-track task formulation distinguishing target-aimed pointing from scene-aware referential gesture, and a spatio-temporal evaluation protocol decomposing gesture quality into temporal alignment, spatial accuracy, and referent recall.

A.D. benchmark design, implementation, evaluation, baseline design and implementation, manuscript writing; R.D. benchmark formulation, manuscript editing; F.I.D. benchmark formulation, manuscript editing; A.G. benchmark formulation, implementation; J.B. benchmark formulation, supervision, manuscript editing.

Paper F

MM-Conv: A Multimodal Dataset and Benchmark for Context-Aware Grounding in 3D Dialogue. This paper introduces MM-Conv, a multimodal conversational dataset recorded in VR using the AI2-THOR simulator, with synchronized speech, full-body SMPLX motion, gaze, and 3D scene graphs. The

dataset is designed to support research on situated reference and grounded multimodal communication, and the paper presents baseline analyses of referring expressions and grounding.

A.D. idea formation, VR data collection setup, data collection, data processing, human study design, running models, manuscript writing; J.O'R. language and dialogue annotation, running models, manuscript editing; F.I.D. technical discussions, data processing, manuscript editing; I.L. technical discussions, manuscript editing; L.M. data processing, running models; A.K. data collection, designing and running human study; J.B. data processing, supervision, manuscript editing.

Paper G

Look and Tell: A Dataset for Multimodal Grounding Across Egocentric and Exocentric Views. This paper introduces Look and Tell, a dataset for multimodal referential communication captured simultaneously from egocentric and exocentric viewpoints, supporting research on cross-view grounding in situated dialogue.

A.D. idea formation, data collection setup, data processing, running models, manuscript writing; J.B. supervision, manuscript editing.

2 Background

Embodied agents that interact with people in shared physical environments must do more than process language. They must interpret and produce communicative behavior that is anchored in a concrete spatial situation. An instruction such as *put that over there* is meaningful only relative to a shared scene, a speaker's gaze, and a history of prior reference; a pointing gesture toward a shelf means one thing when two objects sit on it and another when twenty do. To act on such instructions, an agent must resolve what *that* and *there* refer to in the current spatial configuration. To produce them, it must generate language and gesture that pick out the intended referent from a listener's perspective. Effective embodied communication therefore requires the ability to coordinate language with body and environment, on both sides of the interaction.

This challenge is not confined to any single application domain. Robots interpreting collaborative task instructions, executing spatial commands, and AR assistants mediating between a user's gaze and their surroundings all face versions of the same problem: how to produce and interpret behavior that is grounded in shared physical space. Clark and Brennan characterise co-present interaction through properties such as shared visibility, simultaneity, and access to a common physical environment, arguing that these affordances support the establishment of common ground in communication [16]. Addressing these challenges computationally requires bridging theories of human communication with machine learning and generative modeling. This bridge is what the technical chapters of the thesis provide.

This chapter surveys the communicative phenomena that underpin this challenge. Section 2.1 introduces co-speech gesture, the primary embodied signal studied on the production side of the thesis. Section 2.2 examines referential communication and deixis, Section 2.3 then focuses on pointing as a foundational deictic behaviour. Section 2.4 discusses the implications for embodied agents on both the production and comprehension sides. Section 2.5 then draws these threads together and bridges to the technical chapters that follow.

2.1 Co-Speech Gesture

Co-speech gesture refers to the spontaneous hand and arm movements that accompany spoken language. Classic work by McNeill [19] and Kendon [20] distinguishes several broad gesture types, including iconic gestures that depict actions or shapes, metaphoric gestures that represent abstract ideas, beat gestures that mark discourse rhythm, and deictic gestures such as pointing that refer to entities in space. Although these categories are analytically useful, real communicative behavior often combines multiple functions at once. A single gesture may be simultaneously deictic and iconic, directing attention to an object while depicting an action performed on it, a pattern illustrated in the data collected for this thesis (Figure 6.7).

Gesture is not merely decorative. It is tightly coupled to speech in timing and meaning. Gestures can emphasize prosody, disambiguate reference, and externalize spatial or action-related information. McNeill's growth point theory argues that speech and gesture are co-expressive channels of a single underlying thought unit, rather than separately generated streams that happen to co-occur [19]. Kendon [20] makes a similar argument: he treats visible bodily action as part of the utterance itself, which undermines a clean separation between "verbal" and "nonverbal" communication. From this perspective, what we call gesture and what we call speech are two aspects of a single process of utterance production.

The research tradition that first formalised this insight computationally is the work on embodied conversational agents (ECAs). Cassell et al. [21] argued that conversational interfaces should integrate verbal and nonverbal channels, speech, gesture, facial expression, and gaze, within a unified system grounded in models of face-to-face dialogue. Early ECA systems such as BEAT [22] demonstrated that timing, posture, gesture, and eye gaze shape how utterances are interpreted and how natural an interaction feels. These systems relied on rule-based generation, manually mapping discourse structure to animation parameters, and operated primarily in abstract conversational settings without spatial context. More recently, machine learning has shifted the field toward data-driven generation. Modern models learn gesture and motion directly from motion-capture data, speech, and text, improving naturalness and variability considerably.

Within this data-driven tradition, the bulk of work has concentrated on beat gestures, for good reason: acoustic features are cheap to obtain at scale, beat gestures align directly with the prosodic structure of speech, and the resulting supervised problem is tractable. The semantic and spatial varieties of gesture, iconic depictions, metaphoric projections, and deictic references, have been comparatively understudied, because they require conditioning signals, lexical meaning and scene geometry, that are not present in the audio stream. This split, between semantic content tied to *what* is being said and spatial content

tied to *where* the referent is located, with beat rhythm carrying neither, organizes the technical chapters: it recurs throughout Chapters 4 and 5 as the organizing tension of both gesture generation and referential grounding. For embodied agents operating in homes, workplaces, or collaborative environments, the traditional focus is limiting, particularly for deictic behavior such as pointing, which is the subject of the next section.

2.2 Referential Communication and Deixis

Reference is the process by which speakers identify entities for one another in discourse. In situated interaction, reference is inherently multimodal: a speaker may say *that cup*, glance toward a shelf, and extend a hand toward a target, relying on the addressee to integrate these signals. Deixis refers to expressions whose meaning depends on the context of the utterance, such as *this*, *that*, *here*, and *there*. These forms are especially common in face-to-face dialogue because they are efficient when interlocutors share a perceptual environment.

Figure 2.1: Pointing and joint attention in natural interaction. Left: an adult directs a child's attention to an overhead feature; the child's gaze follows the deictic cue. Right: a young child points toward a tree. Both instantiate the triadic speaker–addressee–referent structure discussed in Section 2.3.

Two properties of deictic expressions are particularly important for the computational work in this thesis. First, deictic expressions are anchored in the speaker's spatial perspective: their interpretation depends on a coordinate system centred on the speaker's body, location, and time of utterance [23, 24]. Demonstratives establish a direct referential link between language and world

by focusing the interlocutor's attention on entities in the surrounding situation [25]. Second, deictic and anaphoric reference are distinct strategies. While anaphoric expressions refer back to entities already introduced in discourse, deictic expressions reach outward into the spatio-temporal surroundings [24]. Both strategies are pervasive in natural dialogue, and speakers shift fluidly between them depending on what they take to be mutually accessible.

Underpinning both strategies is the notion of common ground. Clark and Brennan [16] characterise common ground as the accumulation of mutual knowledge, beliefs, and assumptions that interlocutors build and maintain over the course of an interaction. Speakers design their referring expressions against this shared record: an expression is chosen to be just specific enough to identify the intended referent given what both parties can be assumed to know. This is why pronominal and deictic forms are not defective expressions but efficient ones, *that one* is the rational choice when the referent is already established in the common ground, just as a full noun phrase is the rational choice when it is not. The efficiency of deixis depends directly on the richness and currency of the common ground: the more that has been established through prior discourse, gesture, and shared perception, the less language needs to do. For embodied agents operating in shared physical environments, this has a concrete consequence. Common ground is not built from language alone; it accumulates through gaze, pointing, and the jointly perceived structure of the scene. A grounding system that has access only to the current utterance is therefore working with a systematically impoverished representation of the information the speaker assumed was available. This theoretical point is the backdrop for the empirical findings in Chapter 5, where context-dependent expressions account for nearly half of all references in naturalistic data, and where the cost of missing that context is made concrete.

2.3 Pointing as a Foundation of Human Communication

Among deictic behaviours, pointing occupies a special position. It has been described as "a basic building block" of human communication [26] and is one of the earliest communicative behaviours to emerge in development, appearing in infancy well before the onset of speech [26, 27] (Figure 2.1). Developmentally, the earliest observed pointing serves a declarative function, sharing attention rather than requesting an object, suggesting that the joint-attention function is primary rather than derivative of instrumental use [28].

The communicative function of pointing is fundamentally triadic. Following Bühler's [23] account, language use involves a speaker, a hearer, and the entities talked about. Pointing makes this triadic structure physically explicit: it links the speaker's body to a target in the world while simultaneously directing the hearer's attention toward it. This is what makes pointing particularly effective for establishing joint attention in shared space. Speakers adjust their referential strategy depending on whether gesture is available: when pointing

is possible, they tend to produce shorter deictic descriptions rather than longer, more explicit noun phrases [29].

These findings have important implications for computational modelling. They suggest that pointing is a core component of referential communication that shapes the structure of the utterance itself. A model of communicative gesture that omits deictic behaviour therefore misses one of the most functionally important aspects of human nonverbal communication.

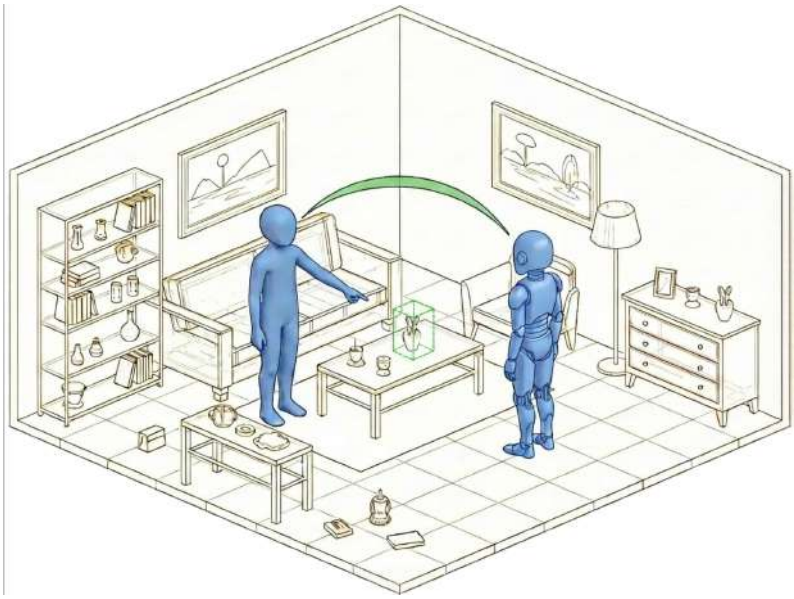
2.4 Implications for Embodied Agents

For embodied agents, pointing and referential communication create a dual challenge. On the production side, the agent must generate deictic behaviour that is spatially accurate, temporally coordinated with speech, and motorically natural. On the comprehension side, it must infer the intended referent by combining linguistic and nonlinguistic cues, including gaze, gesture, dialogue history, and scene geometry. Language alone is often underspecified: expressions such as *that one* or *put it there* require the hearer to recover the intended referent from a rich multimodal context. Figure 2.2 illustrates this dual challenge.

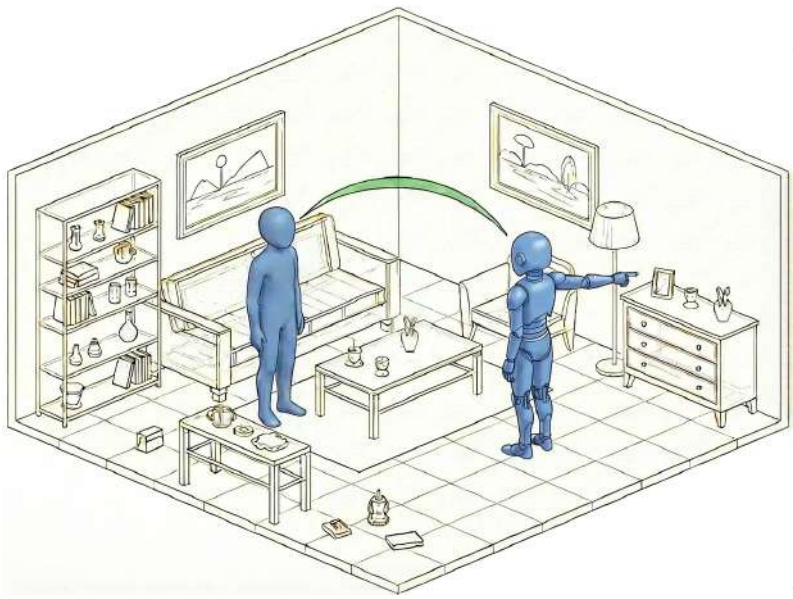
This thesis approaches these two challenges as connected aspects of spatially grounded communication. The generation work in Paper A–E investigates how agents can produce gestures that are both semantically and spatially appropriate, while the grounding work in Paper F–G investigates how agents can resolve references by integrating the same types of multimodal cues. Importantly, these challenges are not specific to virtual agents. A robot interpreting a deictic reference during a collaborative assembly task faces the same triadic coordination problem: it must integrate the speaker's language, gesture, and gaze with a model of the shared scene and the ongoing discourse to identify the intended referent. The shift from virtual to physical embodiment introduces additional challenges, noisy perception, real-time constraints, imprecise actuation, but the underlying communicative structure remains the same. The methods and representations developed in this thesis are therefore designed with this broader applicability in mind.

2.5 From Communication to Computation

Addressing these challenges computationally requires specific technical machinery. Chapter 3 lays the foundation by surveying the motion modeling paradigms, reinforcement learning, adversarial motion priors, diffusion and flow matching, that make it possible to generate natural human motion in the first place. Chapter 4 then addresses the conditioning problem: how to steer these generative models using heterogeneous signals including speech, language, and spatial targets, so that the resulting motion is not only natural but communicatively appropriate. Chapter 5 turns to the comprehension side, ex-



(a) The human produces a referential gesture toward a target object.



(b) The roles are reversed: the agent produces a pointing gesture that the human must interpret.

Figure 2.2: Situated communication between a human and an embodied agent. (a) The human produces a referential gesture toward a target object. (b) Roles are reversed: the agent points and the human must interpret. In both cases, surrounding objects create referential ambiguity that requires integrating gesture with scene layout.

amining how vision-language models, dialogue context, and multimodal cues can be combined to resolve the kinds of situated references described in this chapter. Together, these three technical chapters address the computational challenges that the communicative phenomena identified here demand. Chapter 6 describes the datasets and data collection strategies that support both research threads, and Chapter 7 discusses the evaluation paradigms used to assess communicative success and Chapter 8 concludes by synthesizing the contributions and identifying open questions.

3 Modeling Human Motion

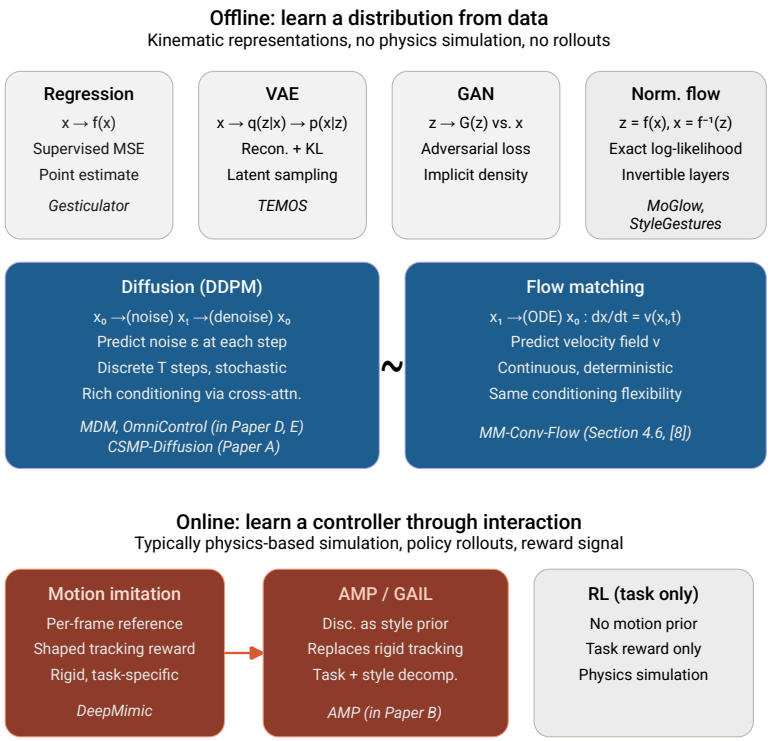


Figure 3.1: Generative model paradigms for human motion. Offline methods learn motion distributions from fixed data using kinematic representations. Online methods learn controllers through interaction, typically in physics-based simulation. Model acronyms appearing in the figure are defined where they first appear in Chapters 3–4.

Computational modeling of human motion serves multiple purposes: understanding the principles that govern how people move, generating realistic animation for virtual characters, and controlling embodied agents that must act in the world. Figure 3.1 provides an overview of the generative model paradigms covered in this chapter. In each case, the core challenge is to produce motion that is not only physically valid but also recognizably human. People move with characteristic dynamics, timing, and variability that are immediately apparent when violated. A model that generates motion outside this distribution, too stiff, too smooth, too symmetric, will be perceived as unnatural regardless of whether it satisfies other task requirements. Recovering and reproducing the statistical structure of natural human motion is therefore a foundational problem that underlies all of the generation work in this thesis.

Naturalness alone, however, is not sufficient. Generated motion must also serve a purpose. In locomotion, the body must reach a target location. In manipulation, the hand must contact an object. In communicative gesture, the motion must convey meaning, direct attention, or indicate a referent in space. This functional requirement introduces a conditioning problem: the generative process must be steered by external signals that specify what the motion should accomplish, without destroying the naturalness of the output. Different applications demand different types of conditioning, a speech signal, a spatial target, a task instruction, and the flexibility with which a model can accommodate such heterogeneous inputs determines what kinds of behavior it can produce.

These two requirements, faithful reproduction of the natural motion distribution and flexible goal-directed control, have shaped the progression of methods in the field. Reinforcement learning enables explicit optimization toward task objectives; when combined with motion imitation, it balances goal achievement with physical realism. More recently, diffusion models have proven highly effective for complex temporal sequence generation, with strong support for conditioning on diverse input signals. This chapter traces this methodological progression as it applies to the work in this thesis: from motion imitation and tracking (Section 3.2), through adversarial motion priors (Section 3.3), to diffusion and flow matching models (Section 3.4). Before turning to these paradigms, Section 3.1 discusses the representation choices that underpin all of them. The chapter concludes with a comparative analysis that frames these approaches in terms of their conditioning flexibility, a property that becomes central in Chapter 4, where the focus shifts from motion modeling to the structured conditioning mechanisms needed for communicative behavior.

3.1 Motion Representations

Human motion can be represented in multiple ways, and the choice of representation has important consequences for learning, generalization, and spatial control. The most common approach is to model the body as a kinematic skeleton, encoding motion through joint positions, joint rotations, or derived quanti-

ties such as velocities and contact labels. Skeletal representations are compact and well suited for generative modeling, but they differ in how they handle hierarchy, global orientation, and physical consistency.

Joint rotations preserve the kinematic chain and avoid bone-length violations by construction, but require careful parameterization: Euler angles suffer from gimbal lock, quaternions introduce a double-cover ambiguity. The continuous 6D rotation representations introduced by Zhou et al. [30] have become a popular alternative for stable gradient-based learning. Cartesian joint positions, by contrast, make spatial constraints easier to express directly but can produce skeletons with inconsistent bone lengths unless constraints or regularization are applied.

The HumanML3D representation [31] illustrates how they interact in practice. Rather than committing to a single view, it concatenates root velocities, joint positions, joint velocities, joint rotations, and foot-contact labels into a unified 263-dimensional vector per frame. This redundant encoding proved effective for the text-to-motion models that followed, including MDM [32] and many subsequent diffusion-based systems, because different learning objectives can exploit whichever subset of features is most informative. Recent work [33] has revisited this design, arguing that the redundancy may not be necessary: a compact product manifold of global translation and joint rotations suffices, provided the learning algorithm respects the underlying geometry. The question of minimal versus redundant representations remains open, but the trend in the field is toward representations that make the geometric structure of the motion space explicit rather than leaving it to the network to discover.

An alternative to skeletal formats is to use parametric body models such as SMPL [34], SMPL-X [35] or MHR [36], which provide a differentiable mapping from pose and shape parameters to a full body mesh. These models decouple body shape from pose and offer a common parameterization across capture setups, making them attractive when visual realism, cross-dataset consistency, or integration with vision pipelines is important. However, many motion generation systems, and most gesture generation systems specifically, continue to operate on skeletal formats, which are simpler to predict, faster to evaluate, and closer to the native output of optical motion capture, which produces marker trajectories typically solved to skeletal rigs. A practical consequence is that systems trained on different parametric backends (e.g., SMPL-X versus MHR) require explicit retargeting to share data.

Recent work has begun to address this fragmentation directly. SOMA [37] proposes a unified body layer that maps heterogeneous parametric models, SMPL, SMPL-X, MHR, to a single canonical topology and rig, enabling motion data from any supported backend to be consumed without custom retargeting. This reduction in cross-model friction is increasingly consequential as large-scale motion datasets drawn from multiple capture pipelines become the norm for training generative models.

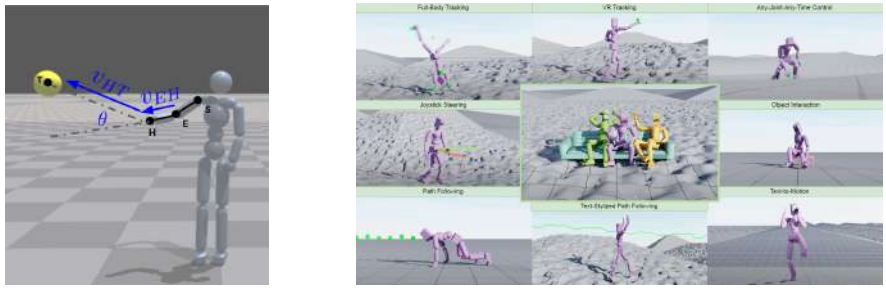
A final and particularly consequential design choice is whether motion is represented in local or global coordinates. Local representations, which express each frame relative to the character's root, simplify learning by factoring out absolute scene position and are standard in most unconditional or text-conditioned generation pipelines. However, they become problematic when the task requires explicit spatial targets in a shared world frame. OmniControl [38] demonstrated this tension directly: the underlying diffusion model (MDM) represents joint positions relative to the pelvis and pelvis motion relative to the previous frame, so global spatial targets cannot be imposed through simple inpainting because the relative pelvis positions needed for conversion are non-existent during denoising, and using intermediate predictions can yield unreasonable velocities or limb lengths that compound across steps. OmniControl's solution was to convert intermediate predictions to global coordinates and use analytic gradient-based guidance to enforce spatial constraints, combined with a learned realism term to keep the remaining joints coherent.

This local–global mismatch is particularly acute for communicative gestures such as pointing and referential indication, where the relevant quantity is the global geometric relationship between the hand, the body, and a target in the environment. Across the papers in this thesis, representation choices are driven by this need to balance expressive motion modeling with explicit control over spatial behavior: the RL-based work (Paper B) operates in global coordinates within a physics simulator where world-frame targets are natural, while the diffusion-based work (Paper A, Paper D, [8]) must contend with the fact that kinematic representations are typically root-relative and require additional mechanisms to incorporate scene-level spatial information. MM-ConvFlow (Section 4.6) sidesteps this conversion by expressing the spatial target in pelvis-relative coordinates, matching the motion representation's own reference frame. For single-target pointing, the pelvis-relative direction is the geometrically relevant quantity; extending to multiple simultaneous targets in a shared world frame would reintroduce the local–global conversion problem.

3.2 Motion Tracking, Imitation and Reinforcement Learning

Motion imitation methods learn policies that reproduce target movements by tracking reference trajectories. A highly influential example is DeepMimic [39], which showed that reinforcement learning can produce physically plausible character control by rewarding similarity to demonstration motion while respecting the dynamics of the body. This paradigm is attractive because it couples natural motion with a physically grounded control process.

These methods operate within a physics simulator, such as MuJoCo [40] or Isaac Gym [41], that numerically integrates the equations of motion for an articulated rigid-body system at each timestep. The simulator computes forward dynamics given applied forces or target joint angles, resolves ground and self-



(a) Pointing policy (Paper B,)

(b) Multi-objective humanoid control in Masked Mimic [42].

Figure 3.2: Physics-based humanoid control in Isaac Gym. (a) A pointing policy from Paper B learns to direct the arm toward spatial targets while maintaining natural posture. (b) Masked conditioning objectives in MaskedMimic [42] enable flexible multi-task control within the same framework.

contacts, and enforces joint limits and torque bounds. This means that physical consistency is not something the learning algorithm must discover from data, it is guaranteed by the environment. The simulator also defines the structure of the control problem: the state comprises joint configurations and velocities, actions correspond to torques or position targets at each actuated degree of freedom, and the transition function $\mathcal{T}(s_{t+1} \mid s_t, a_t)$ is determined by the physics engine rather than learned. What remains for the policy to acquire is which actions to take, a more narrowly scoped problem than learning the full distributional structure of natural motion from a fixed dataset.

DeepMimic's reward formulation is presented here as the canonical example because it establishes the template that later methods, including the Adversarial Motion Priors (AMP) framework adopted in Paper B, modify by replacing the handcrafted terms with a learned discriminator signal.

Mathematical framework. Character control through reinforcement learning (RL) is formulated as a Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma)$, where the objective is to learn a policy $\pi(a_t \mid s_t)$ that maximises the expected discounted return $\mathbb{E}_\pi\left[\sum_{t=0}^{T} \gamma^t\, r(s_t, a_t)\right]$. The state s_t encodes the character's joint configuration and velocities, actions a_t specify target joint angles or torques, and transitions $\mathcal{T}(s_{t+1} \mid s_t, a_t)$ are governed by the physics simulator. DeepMimic [39] defines the reward as a weighted reward function:

$$r_t = w^p\, r_t^p + w^v\, r_t^v + w^e\, r_t^e + w^c\, r_t^c, \tag{3.2.1}$$

where each term compares joint positions, velocities, end-effectors, or centre-of-mass via an exponential kernel over squared differences between simulated and reference motion, e.g.

$$r_t^p = \exp\Big(-\alpha^p \sum_j \|\hat{q}_t^j - q_t^j\|^2\Big).$$

The policy is optimised with proximal policy optimisation (PPO) [43], which maximises the expected discounted return while stabilising training by clipping the likelihood-ratio objective to prevent destructively large updates. This produces natural-looking motion but requires careful per-term tuning and close reference tracking, which limits flexibility when the desired behavior varies across contexts. This limitation motivated approaches that replace handcrafted tracking objectives with learned reward signals, decoupling task goals from motion style.

3.3 Adversarial Motion Priors

Generative adversarial imitation learning (GAIL) [44] offers a more flexible alternative: rather than manually designing a tracking objective, a discriminator learns to distinguish agent behavior from expert demonstrations and its output serves as the reward. Peng et al. [45] adapted this idea for physics-based character animation in Adversarial Motion Priors (AMP), where a discriminator $D_\phi(s_t, s_{t+1})$ yields a style reward. The original formulation uses $r_{\text{style}} = -\log(1 - D_\phi(s_t, s_{t+1}))$; practical implementations, including the one adopted in Paper B, often substitute a least-squares variant for training stability:

$$r_{\text{style}}(s_t, s_{t+1}) = \max\{0, 1 - 0.25(D_\phi(s_t, s_{t+1}) - 1)^2\}. \tag{3.3.2}$$

The total reward separates into a task component specifying what the agent should accomplish and a style component (Eq. 3.3.2) specifying how it should move. This decoupling substantially reduces the need for per-term reward shaping (cf. (3.2.1)).

The discriminator is trained on state transition pairs (s_t, s_{t+1}) drawn from two sources: rollouts of the current policy and transitions sampled from the reference motion capture data. A standard binary cross-entropy objective trains the discriminator to assign high values to reference transitions and low values to agent transitions, while the policy is simultaneously updated to maximise the resulting style reward. Because the discriminator operates on adjacent state pairs rather than full trajectories, it does not explicitly model global temporal structure of the motion and can be applied to reference clips of varying length without modification.

Training stability relies critically on Reference State Initialization (RSI): at the start of each episode, the character is initialized not from a fixed rest pose but from a state sampled uniformly from the reference motion. This ensures that the policy encounters the full range of the motion distribution during training rather than being confined to states reachable from a single starting configuration, which would bias learning toward early-phase behavior and make recovery from mid-motion states difficult to learn.

The AMP framework is architecturally lightweight by design: both the policy and the discriminator are simple multilayer perceptrons (MLPs), which have

proven sufficient for locomotion and general full-body tasks. Specific communicative tasks, however, may require injecting structured prior knowledge into the discriminator or the reward function, these adaptations are discussed in the context of pointing gesture generation in Section 4.4.

3.4 Diffusion and Flow Matching Models

Diffusion models have recently become a dominant paradigm for generative modeling across images, audio, and motion. Earlier generative approaches to human motion, including VAE-based models such as TEMOS [46], encode the full motion sequence into a single latent vector. This design is effective for single-signal conditioning but routes all conditioning through a shared bottleneck, making it difficult to incorporate heterogeneous signals. Diffusion models avoid this bottleneck by operating directly on the full-resolution sequence, allowing conditioning to be injected at multiple layers and time steps rather than compressed into a single latent representation (see Chapter 4). This architectural flexibility makes it substantially easier to incorporate heterogeneous signals through separate pathways (see Chapter 4). Among the first influential applications to human motion was MDM [32], which demonstrated that denoising diffusion models can generate high-quality motion under text and action conditioning.

Mathematical framework. In the standard formulation, a forward process gradually corrupts a data sample x_0 by adding Gaussian noise over T steps:

$$q(x_t \mid x_0) = \mathcal{N}(x_t;\ \sqrt{\bar{\alpha}_t}\, x_0,\ (1-\bar{\alpha}_t)\, \mathbf{I}), \tag{3.4.3}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ controls the noise schedule. Generation proceeds by learning to reverse this process: a network $\epsilon_\theta(x_t, t)$ is trained to predict the noise added at each step, minimising $\mathbb{E}_{x_0,\epsilon,t}[\|\epsilon - \theta(x_t, t)\|^2]$. At inference, iterative denoising from $x_T \sim \mathcal{N}(0, \mathbf{I})$ produces new samples. Because each denoising step is stochastic, the model can map a single conditioning input to many plausible outputs, which makes diffusion models well suited to the one-to-many nature of gesture generation.

Flow matching [47] recasts generation as learning a velocity field $v_\theta(x_t, t)$ that transports noise to data along a continuous-time path, trained with the objective

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,x_0,x_1}[\|v_\theta(x_t, t) - (x_1 - x_0)\|^2], \tag{3.4.4}$$

where $x_t = (1-t)\, x_0 + t\, x_1$ interpolates between noise x_0 and data x_1. This formulation often yields stabler training and permits deterministic sampling via ODE integration, often requiring significantly fewer steps than iterative diffusion denoising. Hu et al. [48] demonstrated this advantage for human motion, matching diffusion-based performance on text-to-motion benchmarks with ten sampling steps.

3.5 Comparative Analysis

The progression from motion imitation, to adversarial priors, to diffusion-based generation reflects a broader evolution in how human motion is modeled. Rather than a linear replacement, this progression represents a shift in where the modeling burden is placed. Figure 3.3 summarizes the structural contrast between the two paradigms that appear in this thesis.

In online, interaction-driven approaches (Figure 3.3a), the policy network itself can be simple, often a shallow MLP, because the complexity of the behavior emerges from the interaction loop. The physics simulator enforces skeletal constraints, contact dynamics, and Newtonian rigid-body dynamics, acting as an implicit prior that relieves the network of learning these regularities from data. The discriminator contributes an additional learned prior over motion style, and the task reward specifies what the agent should accomplish. Together, these components make RL-based methods comparatively data-efficient: a small corpus of reference clips can be sufficient to produce diverse, physically grounded behavior, because the policy learns through online rollouts rather than offline regression. Iteration loops are relatively fast, early learning signal is visible within hours, and reward terms can be evaluated through comparatively short training runs compared to retraining a full diffusion model, and the resulting behavior is physically consistent by construction. The cost is that reward shaping requires domain expertise, training can be unstable, the policy is tightly coupled to its simulation environment, and it is difficult to scale RL methods to leverage the large and diverse motion datasets that kinematic generative models exploit naturally.

In offline, data-driven approaches (Figure 3.3b), the network must capture the full variability of human motion from a fixed dataset. This demands deeper architectures, typically transformers, whose capacity is sufficient to model the rich distributional structure of human motion. The payoff is conditioning flexibility. New input modalities can be introduced through cross-attention, classifier-free guidance, or auxiliary encoders without fundamentally altering the denoising backbone, making it straightforward to progressively enrich the set of conditioning signals. The trade-off is that physical plausibility must largely be learned from data: generated motion is kinematic and may violate contact or dynamics constraints unless additional regularization is applied. Training is slower, and modifying the conditioning setup has traditionally required retraining on the full dataset, though the adapter-based strategies described in Section 4.2 increasingly mitigate this cost.

This data-dependence is directly visible in the pointing task studied in Paper B, which compared the two paradigms on the same task and dataset. Figure 3.4 shows the correlation between the normalizing-flow-based model (MoGlow's) pointing accuracy and training data density, estimated via kernel density estimation (KDE) over the training target positions. The strong positive correla-

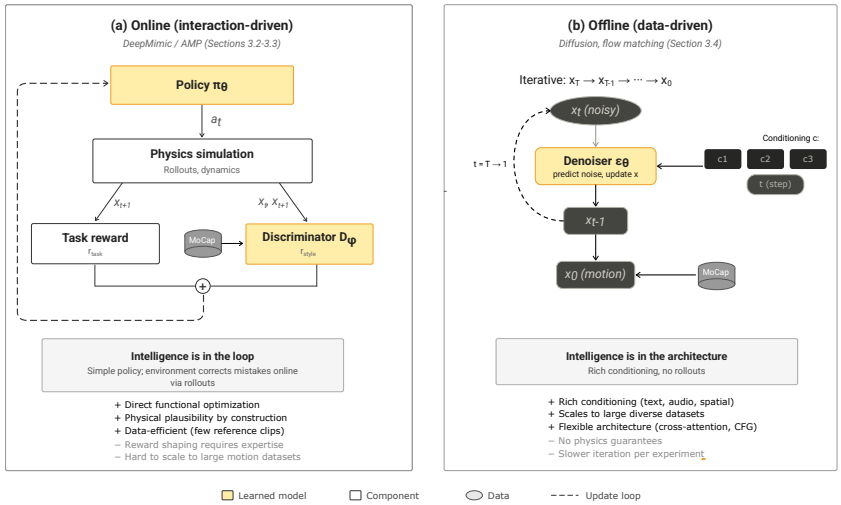


Figure 3.3: Structural contrast between the two generative paradigms used in this thesis. Online methods (a) employ a simple policy within a physics simulation loop, where the environment provides information; offline methods (b) employ a deep generative model trained on fixed data, where conditioning enters through architectural pathways (cross-attention, adaLN).

tion (Spearman $r = 0.718$, $p < 0.001$) provides strong evidence that the supervised model's accuracy is tightly bound to the training distribution: it points accurately only in regions densely covered by the motion capture data, and degrades sharply in sparse regions. The heatmaps in Figure 3.4B–C make this spatial dependence visible, with high-density training regions (Figure 3.4B) corresponding closely to high-accuracy regions (Figure 3.4C). The RL-based AMP models, by contrast, show weak or even negative correlations with training density, indicating that the simulation loop and task reward enable generalization to target positions far from those seen during training. This contrast concretely illustrates the data-efficiency advantage of online methods when the available dataset is small and task-specific, and it supports the choice of RL-based training for the spatial gesture work in Paper B.

These complementary strengths explain why both paradigms appear in this thesis. The RL-based work in Paper B exploits the efficiency and physical grounding of simulation-based training to study pointing behavior under precise spatial objectives, where rapid experimentation with reward structure was essential. The diffusion- and flow-matching-based work in Paper A and Paper D, [8] leverages the conditioning flexibility needed to model grounded co-speech gesture, where multiple heterogeneous inputs, speech, language, and spatial targets, must jointly shape a rich space of possible outputs.

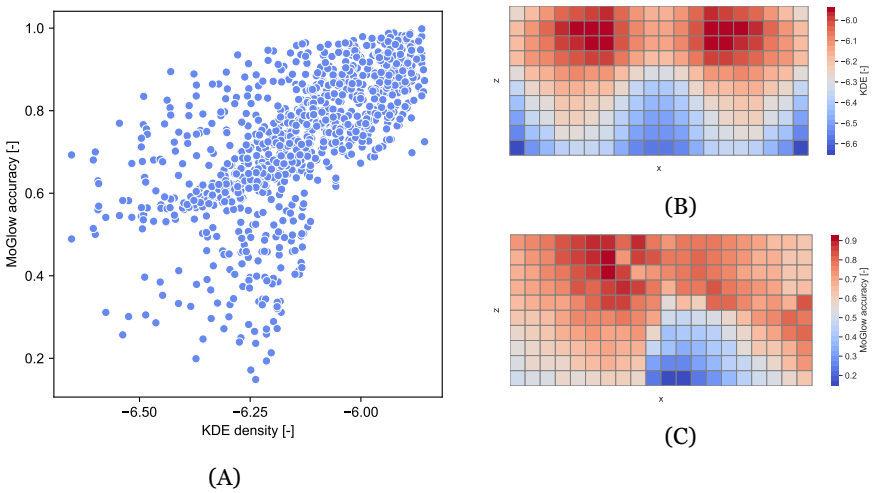


Figure 3.4: Correlation between MoGlow pointing accuracy and training data density (KDE), from Paper B. (A) Scatter plot showing strong positive correlation (Spearman $r = 0.718$, $p < 0.001$). (B) KDE density and (C) MoGlow accuracy projected onto the x–z plane by averaging over y, showing that high-density training regions correspond to high-accuracy regions.

Although this thesis uses the two paradigms separately, recent work has begun to combine them within a single system. CLoSD [49] maintains a closed-loop interaction between a text-conditioned autoregressive diffusion planner and a physics-based RL tracking controller: the diffusion model generates short-horizon kinematic motion plans from text prompts and spatial targets, while the controller executes them in simulation and feeds back the resulting state for re-planning. Each module uses the representation natural to its paradigm, delta-encoded HumanML3D features for the diffusion planner, global position and velocity for the physics controller, a design that directly reflects the local–global tension discussed in Section 3.1. MaskedMimic [42] approaches the same convergence from the RL side, showing that a single physics-based controller can accommodate diverse conditioning objectives through masked motion inpainting (Figure 3.2b), narrowing the conditioning gap that has traditionally favored data-driven methods. These hybrid architectures confirm the complementarity that motivates the present thesis while suggesting that the boundary between the two paradigms is becoming increasingly porous.

The conditioning problem itself, how to steer motion generation from heterogeneous multimodal signals, is the subject of the next chapter.

4 Conditioning Embodied Behavior

The previous chapter surveyed the learning paradigms used to model human motion, from reinforcement learning in physics simulation to diffusion and flow matching over kinematic data. That discussion focused on how motion distributions are learned, the choice of objective, the role of the simulator or dataset, and the trade-offs between online and offline training. This chapter shifts attention to a different question: how the output of a generative model can be made to depend on appropriate conditioning signals. For embodied agents, those signals include speech acoustics, linguistic content, spatial targets, and scene structure: inputs that differ in modality, temporal resolution, and representational format. The central challenge is not generation, but *conditioning*: steering the generative process so that the resulting motion is not only natural but communicatively appropriate.

Conditioning mechanisms can be characterized along three dimensions. The first is *when* conditioning is applied: at training time, where the signal is baked into the model's weights, or at inference time, where a frozen model is steered through gradient guidance or interpolation between conditioned and unconditioned outputs. The second is *how* the signal enters the model architecturally: through feature injection into residual blocks, normalization modulation via adaptive layer normalization, cross-attention over sequences of arbitrary length, or prefix tokens added to the input. The third is *what* type of signal is being incorporated: a global scalar or vector such as a spatial target, a temporally aligned stream such as audio features, or a variable-length sequential context such as text tokens or dialogue history.

These three dimensions interact: the architectural pathway constrains which signal types can be accommodated, and the training-versus-inference distinction determines what can be added to an existing model without retraining. The progression of architectures described in Section 4.1 focuses on the *how* and *what*, going from rigid single-stream injection toward flexible multi-stream conditioning. Section 4.2 focuses on the *when*, on extending pretrained models to new conditioning signals without retraining them. The remaining sections are organised around the types of conditioning signals. Section 4.3 discusses semantic conditioning from speech, while Section 4.4 addresses spatial conditioning from 3D targets and scene information. Sections 4.5 and 4.6 examine combining these streams within a single generative framework.

4.1 Conditioning Architectures

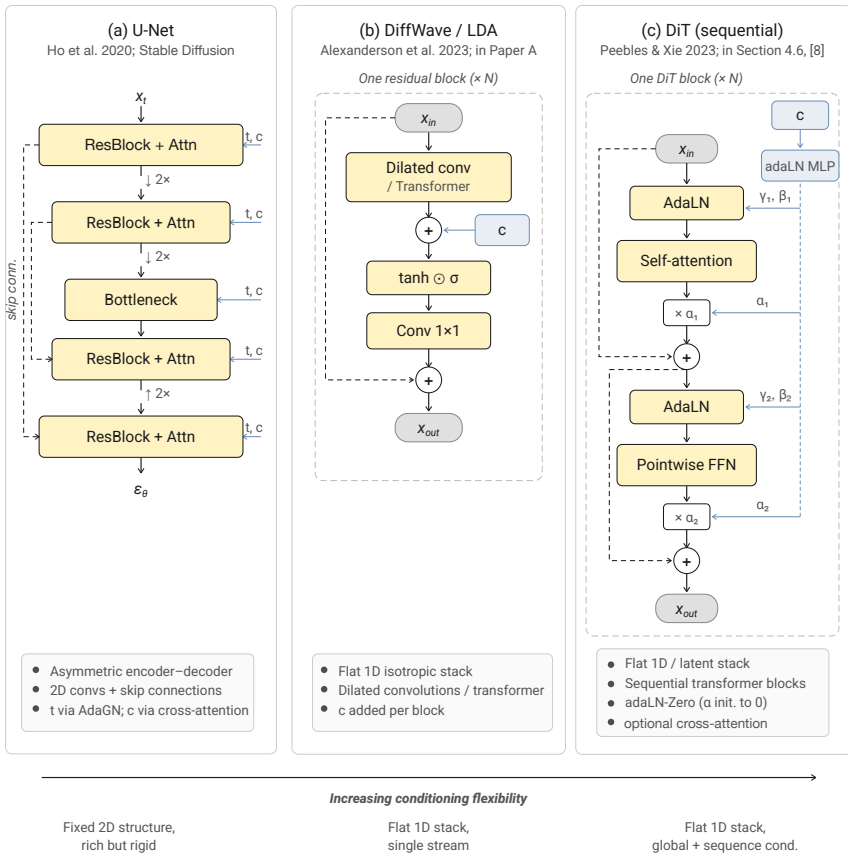


Figure 4.1: Denoising network architectures. (a) The U-Net uses an asymmetric encoder–decoder with 2D convolutions and skip connections; conditioning enters via adaptive group normalisation (AdaGN) (t) and cross-attention (c). (b) DiffWave / LDA replaces the encoder–decoder with a flat stack of residual blocks using dilated convolutions or transformers; a single temporally aligned conditioning stream is added at each block (Paper A). (c) The Diffusion Transformer (DiT) processes the sequence through sequential transformer blocks with adaLN-Zero conditioning and optional cross-attention, supporting both global and sequence-level signals (Section 4.6). The progression from left to right reflects increasing conditioning flexibility.

The conditioning mechanisms available to a generative model are constrained by its network architecture. In the image domain, the standard denoising network is a U-Net [50]: an encoder-decoder with skip connections, designed around two-dimensional spatial convolutions. Motion, however, is a temporal sequence with no grid structure, and the U-Net's spatial inductive biases do

not transfer naturally. DiffWave [51] addressed this mismatch for audio waveforms by replacing the U-Net with a stack of residual blocks built from dilated convolutions, inheriting the WaveNet [52] architecture that had proven effective for autoregressive audio synthesis. This was significant for motion generation because it demonstrated that diffusion models could operate on temporal sequences without the spatial assumptions of the U-Net. The listen-denoise-action (LDA) model [53] brought this design to human motion and further replaced the dilated convolutions within each residual block with translation-invariant transformers [54] to extend the temporal receptive field. Paper A builds on LDA, injecting the CSMP conditioning signal into the residual blocks in place of MFCC features. The architecture is summarised in Figure 4.2. In this paradigm, conditioning enters as a single temporally aligned stream added to the hidden state at each block, effective when the conditioning signal matches the output's temporal structure, but limiting when structurally different inputs need to be incorporated. Figure 4.1 summarises these three architectural families and the conditioning pathways available to each.

More recently, transformer-based denoisers have replaced the residual-block scaffolding entirely. MDM [32] pioneered this shift for human motion, processing the full sequence through self-attention with a CLIP text embedding prepended as a prefix token. The Diffusion Transformer (DiT) [55] crystallised this shift in the image domain, replacing the U-Net entirely with a stack of transformer blocks and introducing adaptive layer normalisation (adaLN) as the primary conditioning pathway. Unlike standard LayerNorm, which uses fixed learned scale and shift parameters, adaLN derives these from the condition signal, via a small MLP, making normalization a function of the condition. AdaLN-Zero initialises the output scale parameters to zero, so each transformer block begins training as an identity function and learns to modulate. This simple design already made frame-aligned conditioning optional rather than architecturally required and demonstrated the flexibility of transformers as denoising backbones.

Subsequent work in both motion and gesture generation has expanded the conditioning vocabulary further, introducing cross-attention to external sequences of arbitrary length, adaptive layer normalisation modulated by global variables, and multi-stream inputs that enter through distinct architectural pathways. Combined with classifier-free guidance [56], which interpolates between conditioned and unconditional model outputs at inference to control adherence to the conditioning signal, transformer-based denoisers now support the kind of structured multi-signal conditioning that spatially grounded gesture generation demands. Paper D and [8] adopt this broader paradigm, and the richer conditioning mechanisms it provides are what make it possible to move from the single-stream semantic conditioning of Paper A toward the unified spatial-and-semantic framework explored in Section 4.6.

4.2 Post-training: Extending Pretrained models With New Conditioning Signals

A pretrained generative model defines a prior over natural motion that can be reused across tasks. Post-training conditioning exploits this by treating the backbone as fixed and training only a lightweight adapter that introduces a new conditioning signal (spatial targets, semantic labels, or scene context) into the denoiser's feature space. Rather than baking all conditioning signals into a single training run, this approach decomposes conditioning into layers: the backbone handles motion naturalness, and adapters handle task-specific signals independently. This modularity is especially attractive when the new signal is available in limited quantities, as is the case for spatially annotated gesture data, since only the adapter needs to see it.

ControlNet-style adapters. To maintain whole-body coherence without discarding the pretrained prior, several works introduce a trainable copy of the denoiser's encoder connected to the frozen original through zero-initialised projections [57]. Because the projections initialise to zero, the adapter begins as an identity and gradually learns residual feature corrections as training proceeds, adding new conditioning without destabilising the prior. OmniControl [38] combines inference-time analytic spatial guidance, which converts generated motion to global coordinates for gradient-based refinement, with a ControlNet-inspired realism guidance module that outputs learned feature corrections through a trainable copy of the denoiser's transformer encoder connected via zero-initialized linear layers. SMooDi [58] applies a similar ControlNet-style adapter for style conditioning on top of MLD [59], complemented by both classifier-free and classifier-based style guidance.

Parameter-efficient and staged adaptation. When the target conditioning signal is genuinely limited, two strategies avoid the data demands of full backbone finetuning. Parameter-efficient adaptation, exemplified by LoRA-MDM [60], inserts low-rank updates into the backbone's attention projection matrices while leaving base weights frozen, minimising forgetting and preserving the motion manifold at the cost of adapter capacity. Staged proxy-task adaptation, used by SceneAdapt [61], addresses cases where fully paired training data does not exist: the model is first adapted to a related task supervisable from existing data, inbetweening in that case, and then a separate conditioning module is injected and trained on the scarcer target signal while earlier weights remain frozen.

In Paper D, the OmniControl mechanism is adopted as a starting point and finetuned for the pointing gesture domain, where spatial targets are sparse 3D endpoint locations in a scene. Paper E applies the same post-training strategy in its OmniControl-PT reference baseline: a learned IK module (108K parameters) maps target coordinates to wrist and elbow positions at a phrase-

anchored apex frame, and the finetuned OmniControl backbone generates full-body motion from these spatial hints. The baseline is scored under the spatio-temporal grounding protocol described in Section 7.1.3. Section 4.6 presents MM-Conv-Flow [8], a more substantial alternative that replaces OmniControl with a flow-matching DiT and adds gated spatial conditioning as a post-block residual.

Table 4.1 summarises the conditioning methods used in each paper and where they are described in the remainder of this chapter.

Table 4.1: Methods and their role in the thesis papers. The first group covers the generative backbone (Chapter 3); the second covers the conditioning mechanisms (Chapter 4). Sections refer to where each method is described in the kappa.

Method	Paper(s)	Section	Role
Generative backbone (Ch. 3)			
AMP (RL + discriminator)	B	3.3	Physics-based motion prior
DDPM (residual denoiser)	A	3.4	Offline kinematic generation
Flow matching (DiT)	[8]	3.4	Offline kinematic generation
Conditioning mechanism (Ch. 4)			
CSMP + diffusion	A	4.3	Semantic conditioning
Target coords + pointing reward	B	4.4	Spatial conditioning (RL)
MoGlow [62]	B	4.4	Supervised spatial baseline
OmniControl (fine-tuned)	D	4.2, 4.5	Spatial control adapter
OmniControl-PT	E	4.2	Spatial baseline (learned IK)
MM-Conv-Flow (DiT + gated)	[8]	4.6	Semantic + spatial conditioning

4.3 Semantic Conditioning: Audio and Language

A large body of work in co-speech gesture generation conditions motion on audio, text, or both, and the choice of input modality directly shapes which gesture types a system can model [63]. Acoustic features capture rhythm, emphasis, and prosodic structure, making audio-driven systems effective at generating beat gestures that reinforce the temporal contour of speech. Self-supervised speech models such as HuBERT [64] and WavLM [65] have further enriched these representations, providing frame-level features that go beyond hand-crafted spectral descriptors. However, generating representational gestures, iconic gestures that visually depict attributes of a referent, requires access to semantic content that acoustic features alone do not reliably encode. Text-based conditioning addresses this gap by providing lexical and semantic information, and pretrained language models offer strong encodings for this

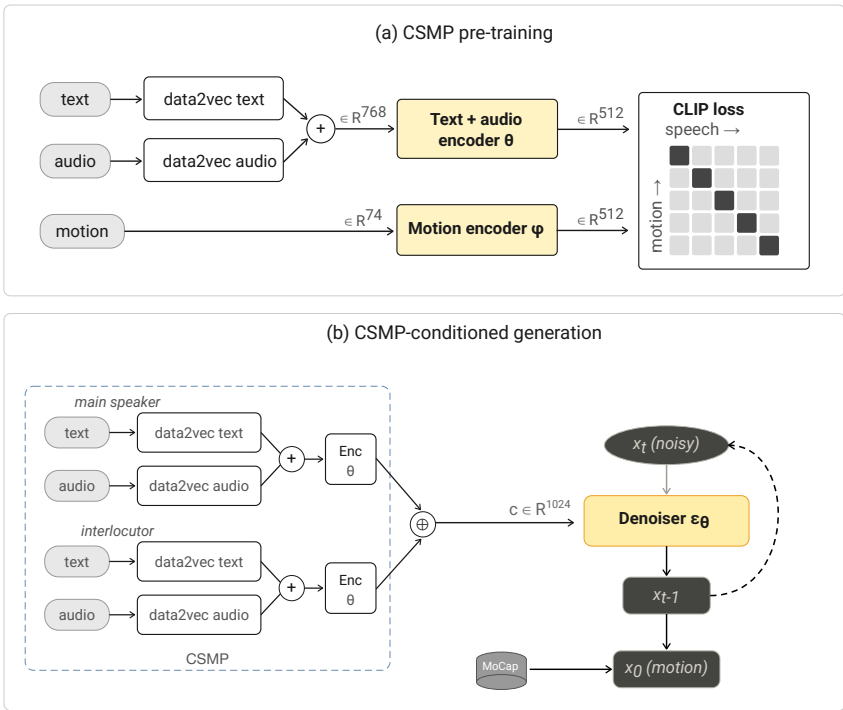


Figure 4.2: Architecture of the Contrastive Speech and Motion Pretraining (CSMP) module and its integration into the diffusion-based generation pipeline (Paper A). (a) During pre-training, data2vec representations of text and audio are summed and encoded by a transformer-based speech encoder θ, while a motion encoder φ processes raw pose vectors. The two encoders are trained with a symmetric contrastive (CLIP) loss. (b) At generation time, the pre-trained encoder θ produces embeddings for both the main speaker and the interlocutor, which are concatenated into a conditioning signal $c \in \mathbb{R}^{1024}$ that guides the classifier-free diffusion denoiser.

purpose. The challenge, then, is to combine acoustic and semantic signals in a way that preserves both the temporal alignment needed for beat synchrony and the meaning needed for representational gestures.

StyleGestures [66] used normalising flows conditioned on mel spectrograms to learn a distribution over motion sequences, demonstrating that probabilistic models can reproduce the versatility and expressivity of human gesture. By sampling different realisations from the learned distribution, the system could generate varied motion for the same input, avoiding the over-smoothing that plagues deterministic regression approaches. However, without access to lexical content, StyleGestures was limited to beat-level alignment with the speech signal. Gesticulator [67] addressed this gap by adding text conditioning

through BERT embeddings alongside mel spectrogram audio features. This combination showed that semantic input is essential for generating representational gestures such as iconic and deictic forms. However, its deterministic regression architecture (which used FiLM conditioning [68] to integrate autoregressive pose history) meant that the system could only produce a single output for a given input, constraining the variability and naturalness of the generated motion.

The listen-denoise-action (LDA) model [53] adopted denoising diffusion probabilistic models (DDPM), a stronger class of probabilistic generative model than normalising flows, and achieved state-of-the-art naturalness in perceptual evaluations. Diffusion models learn to reverse a gradual noising process, which allows them to capture complex, multimodal distributions over motion sequences. Yet LDA conditioned only on MFCC audio features and, like StyleGestures, lacked semantic awareness. All three systems also shared two further assumptions: they operated on single-speaker data and in a monologue setting, with no representation of dialogue context or interlocutor behaviour.

In parallel, work in the broader text-to-motion generation community began leveraging CLIP-based latent spaces to bridge language and motion. Contrastive Language–Image Pre-Training (CLIP) [69] learns a joint embedding space for images and text by training paired encoders with a contrastive objective that pulls matching image-text pairs together and pushes mismatched pairs apart. The resulting representations encode rich semantic structure and have been widely adopted as a conditioning signal in generative models across domains. Several works adapted this idea to the motion domain. MotionCLIP [70] learned a motion embedding space aligned with CLIP's latent space, enabling semantically expressive text-driven motion generation. The Motion Diffusion Model (MDM) [32] used CLIP text embeddings as conditioning information in a classifier-free diffusion framework, achieving strong results on text-to-motion benchmarks. Ao et al. [71] learned a gesture–text joint embedding using contrastive learning and a CLIP-based style encoding module within a diffusion-based gesture synthesis model. These works demonstrated the potential of contrastive pretraining for aligning linguistic and motion representations, but were primarily designed for text-to-motion tasks rather than speech-driven gesture generation in conversational settings.

The remainder of this section describes the approach taken in Paper A, which addresses the semantic gap by introducing a contrastive pretraining module that aligns speech and motion representations before generation. Paper A builds on LDA, replacing the MFCC conditioning with a Contrastive Speech and Motion Pretraining (CSMP) module (Figure 4.2) that learns a joint embedding of speech and gesture in order to provide the diffusion model with semantic information beyond what raw acoustics carry. CSMP adapts the CLIP [69] contrastive learning framework to the speech–motion domain. On the speech side, pre-trained data2vec [72] models extract self-supervised representations for

both audio and text: the audio encoder produces 768-dimensional embeddings at 50 Hz, resampled to 30 Hz to match the motion frame rate, while the text encoder produces 768-dimensional token embeddings that are time-aligned to the same rate using word-level timestamps. The summed audio–text representation is processed by a transformer encoder. On the motion side, a second encoder takes continuous motion features, 74-dimensional exponential map joint rotations, rather than discrete tokens, removing the need for an embedding layer. The two encoders are trained with a contrastive loss (InfoNCE [73]) that pulls corresponding speech–motion segments together and pushes non-corresponding pairs apart in a shared latent space, analogous to the image–text alignment in CLIP. Given a batch of N paired speech–motion segments $\{(\mathbf{s}_i, \mathbf{m}_i)\}_{i=1}^{N}$, the CSMP loss maximises the cosine similarity of matching pairs while minimising it for non-matching pairs:

$$\mathcal{L}_{\text{CSMP}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(\text{sim}(\mathbf{s}_i, \mathbf{m}_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{s}_i, \mathbf{m}_j)/\tau)} + \log\frac{\exp(\text{sim}(\mathbf{m}_i, \mathbf{s}_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{m}_i, \mathbf{s}_j)/\tau)}\right] \tag{4.3.1}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a learnable temperature parameter. The loss is symmetric: the first term matches each speech representation to its corresponding motion, and the second matches each motion representation to its corresponding speech.

To accommodate the longer temporal context required by continuous signals, the context length is extended from CLIP's 77 tokens to 500 frames, and training samples are extracted with a sliding window (hop length 250) so that each timestep appears at multiple relative positions, aiding generalisation of the positional encoding. At generation time, the CSMP embedding serves as the conditioning signal in a classifier-free diffusion guidance formulation: it is injected into each residual block as a temporally aligned stream, replacing LDA's MFCC features.

In the GENEA Challenge 2023 [74], the system achieved the highest human-likeness and speech appropriateness ratings among all submitted entries. The human-likeness score was statistically indistinguishable from the motion-capture reference condition, suggesting that the diffusion model's capacity to reproduce natural motion variability is not the limiting factor in co-speech gesture generation, the bottleneck lies in conditioning, not generation. Interlocutor appropriateness, however, ranked among the lowest, despite the system conditioning on both the main speaker's and the interlocutor's speech embeddings (concatenated into a 1024-dimensional signal $c \in \mathbb{R}^{1024}$). This result suggests that speech-level interlocutor conditioning without access to interlocutor motion is insufficient for social coordination, identifying dialogue-aware generation as the remaining open problem.

4.4 Spatial Conditioning: Target Position in Pointing Gesture Generation

Where semantic gesture generation conditions on speech features that unfold over time, pointing gesture generation conditions on spatial quantities derived from the agent's surroundings: in the simplest case, the three-dimensional position of a single referent; more generally, a structured scene representation such as a scene graph or object layout. The conditioning problem is structurally different in either case: the generated motion must achieve precise geometric alignment with an external target while retaining natural dynamics.

The spatial conditioning mechanisms described in the remainder of this section are drawn from Paper B, which investigated this problem through two contrasting paradigms: reinforcement learning with adversarial motion priors, and a supervised normalizing-flow baseline (MoGlow).

How this spatial information enters the system differs sharply between the two paradigms explored in Paper B and their baselines. In the reinforcement learning framework (Paper B), the target position participates in two ways. First, the xyz coordinates of the target are concatenated directly to the policy network input, giving the agent access to the goal at every decision step. Second, the target defines a task reward r_t^{Pt} that measures the angular alignment between the forearm extension and the hand-to-target direction (also see Figure 3.2a(a)):

$$\hat{\theta} = 1 - \frac{\angle(\vec{V}_{HT}, \vec{V}_{EH})}{\pi}, \qquad r_t^{\text{Pt}} = \frac{e^{\hat{\theta}} - 1}{e} \tag{4.4.2}$$

where $\vec{V}_{HT}$ is the vector from the hand joint to the target and $\vec{V}_{EH}$ is the elbow-to-hand vector. The target participates in two roles: as raw coordinates concatenated to the policy observation, and as the quantity the task reward optimises, rather than being routed through a learned representational bottleneck; the policy itself remains a simple MLP, with no attention layers or learned target embeddings. More broadly, this illustrates a defining property of online methods: spatial conditioning does not require explicit representational capacity in the architecture but can instead be enforced through interaction, allowing simple policies to achieve precise geometric behaviour.

Even with this distributed conditioning, the standard AMP discriminator proved insufficient for pointing. The raising phase and the retraction phase of a pointing gesture follow similar trajectories in opposite directions, and a vanilla MLP discriminator could not reliably distinguish between them, resulting in policies that produced motion stuck between idle and pointing poses (Paper B). Replacing the discriminator with a phase-functioned neural network (PFNN) [75], which explicitly conditions its weights on a cyclic phase variable, resolved this ambiguity by giving the discriminator access to where in the ges-

ture cycle a transition occurs. This modification illustrates how spatial conditioning in the RL paradigm extends beyond the policy's observation and reward: the discriminator, too, must be adapted when the motion it evaluates has temporal structure that a generic function approximator cannot capture. The point is conceptual rather than implementational: the same geometric trajectory can correspond to different communicative phases (raising toward a target versus retracting from it) so spatial conditioning for gesture is inherently spatio-temporal, and the model evaluating that motion must incorporate temporal context explicitly.

MoGlow [62] serves as the offline baseline: an autoregressive normalising-flow model that generates motion as $p(\mathbf{x}_t \mid \mathbf{x}_{<t}, \mathbf{c})$, conditioned on a four-dimensional vector $\mathbf{c} = [x, y, z, b]$ combining target coordinates with a binary phase indicator b distinguishing pointing from idle frames. The discrete signal suffices because the model has access to motion history; AMP, lacking history, instead receives a continuous phase variable $p \in [0, 1]$ at the policy input. Unlike AMP, MoGlow has no external correction loop: all knowledge of the spatial relationship between body and target must be captured in the learned conditional distribution. This is the offline counterpart to the RL observation: where the simulator provided geometric feedback for free, the offline model must internalise the spatial relationship in its parameters, which makes it sensitive to the training distribution and limits generalisation beyond regions densely covered by demonstrations.

The empirical contrast from Paper B supports this reading. In a VR-based referential game with 39 participants and two distractors placed 20–40 cm from each target, AMP-pfnn reached 95.5% referential accuracy and a naturalness MOS of 3.57, compared to 45.9% and 2.61 for MoGlow; the motion-capture nearest-neighbour topline, GT-nn, reached 85.6% and 4.01. On an objective grid of 1,000 test points spanning the training target volume, the same ordering holds: AMP-pfnn attains a mean normalised pointing accuracy of 0.62 against MoGlow's 0.46. Removing the phase input to the AMP policy network drops accuracy from 0.59 to 0.56, isolating the contribution of the phase signal. All pairwise differences on both measures are significant at $\alpha = 0.01$ under a Wilcoxon signed-rank test with Holm–Bonferroni correction, with the single exception of AMP-base versus GT-nn on referential accuracy. Generalisation across target space further separates the paradigms: MoGlow's per-target accuracy correlates strongly with training-set density (Spearman $r = 0.718$, $p < 0.001$), whereas none of the AMP variants shows a correlation above $|r| = 0.15$, indicating that the AMP variants exhibit substantially weaker dependence on training density than the supervised baseline.

This contrast reinforces the pattern identified in Chapter 3: in online methods, even a minimal conditioning interface suffices when the interaction loop compensates; in offline methods, the conditioning pathway bears the full representational burden and must be designed accordingly. The VR-based referential

game through which these results were obtained is itself a methodological contribution, replacing subjective quality ratings with a functional communicative measure; its design is discussed in Section 7.1.1.

In both cases, however, the conditioning signal (target-directed reward for RL, coordinate vector for the supervised model) knows nothing about *what* is being pointed at: spatial accuracy is achieved at the cost of semantic emptiness, with no representation of the referent itself. This split between the geometric and the semantic motivates the unified conditioning architecture of Section 4.6 and reappears on the comprehension side in Chapter 5 as the gap between localisation and identification. More fundamentally, the contrast suggests that spatial grounding in embodied agents is not purely a representational problem but an interactional one: geometric correctness can emerge from closed-loop feedback rather than explicit encoding, an observation that becomes central when extending from spatial to semantic conditioning, where such feedback is no longer directly available.

4.5 Unifying Semantic and Spatial Conditioning

The two conditioning threads examined in the preceding sections develop largely in isolation. Semantic conditioning (Section 4.3) operates on speech and language signals but assumes no scene; spatial conditioning (Section 4.4) targets precise geometric alignment but ignores semantic content. In natural referential communication, these aspects are inseparable: a speaker may say *see that chair over there* while extending an arm toward a target, with the speech constraining *what* is being referenced and the gesture constraining *where*. Modeling this joint behavior requires data and architectures that bring both streams together. This section addresses the data side of this challenge: bringing spatial targets and co-speech motion into a shared format, with the target position provided as oracle input. The architectural integration of semantic and spatial conditioning within a single generative model is taken up in Section 4.6.

Paper D argues that progress on this unification is blocked as much by a data gap as by an architectural one. Existing co-speech gesture corpora provide rich communicative motion but no spatial targets, the pointing dataset provides spatial targets but only formulaic demonstrative speech and text-to-motion datasets provide scene-conditioned behavior but no concept of gesture or speech. Paper D addresses this by standardizing the synthetic pointing corpus (Section 6.3) and the MM-Conv dataset (Section 6.4) into a unified HumanML3D-compatible format, yielding over 7.7 hours of synchronized motion, speech, and 3D scene information across referential and non-referential segments. As a proof of concept, the model from Section 4.2, OmniControl [38], is fine-tuned on these splits with pelvis, left-wrist, and right-wrist joints as spatial control targets. Fine-tuning consistently improves both motion quality (FID) and spatial control accuracy (L2) relative to the pretrained baseline

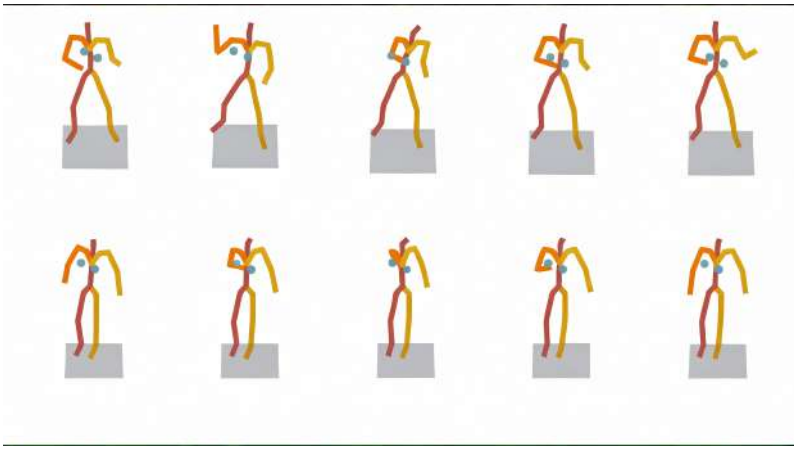


Figure 4.3: Qualitative effect of fine-tuning on pointing data. Top: the base OmniControl model reaches the target wrist position with a punching motion. Bottom: after fine-tuning on the pointing gesture corpus, the same spatial target elicits a natural pointing gesture. Both sequences are generated from the same conditioning input; only the training data differs.

across all training configurations and qualitatively produces pointing gestures where the base model produces arm-extension artifacts. The effect is visible qualitatively as well: Figure 4.3 shows a representative case in which the base OmniControl model produces a punching motion to reach the target wrist position, while the model fine-tuned on pointing data produces a natural pointing gesture toward the same target. The base model has no prior over communicative arm extension and defaults to the closest motion in its training distribution that achieves the positional constraint; fine-tuning on pointing data shifts this prior toward deictic behavior. Fine-tuning on the synthetic pointing split alone reduces wrist Control L_2 error from 0.11–0.12 to 0.06 and FID from 3.8–4.6 to 0.65–0.82, confirming that task-specific data substantially improves spatial control even when the backbone architecture remains unchanged.

4.6 Semantic and Spatial Conditioning: MM-Conv-Flow

This section presents MM-Conv-Flow, a conditional flow-matching system for controllable co-speech and referential gesture generation developed alongside the benchmark introduced in Paper E. Full experimental results, including ablations and a perceptual study, are reported in [8] [1]. Where the OmniControl-PT baseline (Section 4.2) adapts an existing backbone with a learned IK module and phrase-anchored timing, MM-Conv-Flow is built from a DiT backbone trained end-to-end on co-speech data, with spatial target control added through gated residual biases and late cross-attention. The design proceeds in two

[1] Under review; full experimental details are reported there.

phases: P1 establishes motion quality through semantic conditioning from a frozen cross-modal encoder, and P2 extends the architecture with spatial target conditioning via distinct injection pathways that preserve the co-speech prior.

The architectural details are presented here because the design decisions, particularly the factorization of semantic and spatial conditioning into distinct pathways, directly instantiate the thesis's central architectural claim. Paper E focuses on the evaluation protocol and benchmark infrastructure; [8] presents the method and experimental validation.

4.6.1 Semantic Conditioning

MM-Conv-Flow reuses the CSMP module introduced in Paper A (Section 4.3) as a frozen cross-modal encoder, rather than training audio and text encoders end-to-end with the generative model. End-to-end training would create chicken-and-egg dynamics between representation learning and generation, dilute the cross-modal alignment signal through the full generative loss, and require backpropagation through the audio encoder at every step. Freezing the encoder isolates representation learning from generation: CSMP is supervised directly by a contrastive retrieval objective on BEAT2 (100 epochs; 56.16% top-1 retrieval on validation), then provides stable features the generator can learn against without distribution shift.

The encoder provides two outputs to the generative model. The attention-pooled global embedding $c_{\text{csmp}} \in \mathbb{R}^{768}$ captures clip-level semantics and enters the AdaLN conditioning pathway, modulating normalization parameters uniformly across frames. The frame-level features $(B, T, 768)$ retain temporal structure and serve as context for cross-attention in the upper DiT blocks, allowing motion to align with audio prosody at fine temporal resolution. This dual pathway reflects the multi-scale structure of co-speech gesture: global style is uniform within a clip, while local timing must track speech rhythm.

4.6.2 MotionDiT Backbone

The denoising network is a 170.5M-parameter Diffusion Transformer [55] adapted for motion, trained via conditional flow matching on BEAT2.

Each block contains self-attention, cross-attention in layers 4–11, and a feed-forward MLP, with AdaLN conditioning before each sub-layer. Two motion-specific adaptations are used: RoPE in place of absolute position embeddings for length generalization, and windowed masking in cross-attention for beat alignment. The model operates on 300-dimensional motion vectors over the SMPL-X 25-joint skeleton, combining joint positions, 6D rotations [30], and joint velocities (75 + 150 + 75 dims).

Each DiT block receives a conditioning vector c that determines AdaLN scale and shift parameters. In P1, this composition is

$$c = c_{\text{base}} = t_{\text{emb}} + s \cdot \alpha_\ell \cdot c_{\text{csmp}}, \tag{4.6.3}$$

where $t_{\text{emb}} \in \mathbb{R}^{768}$ is the sinusoidal timestep embedding (processed through a 2-layer MLP), s is a learnable conditioning scale initialized to 0.8, $\alpha_\ell = 0.5 + 0.5 \cdot \ell/11$ provides layer-wise scaling from 0.5 in the first block to 1.0 in the last, and $c_{\text{csmp}} \in \mathbb{R}^{768}$ is the frozen CSMP global embedding. The layer-wise scaling reflects the intuition that early layers should focus on low-level motion structure with weaker global conditioning, while later layers should attend more strongly to semantics. A critical property is that c is uniform across all frames: frame-specific information enters only through cross-attention to audio context, not through AdaLN. The cross-attention layer range (4–11) also defines where P2 spatial conditioning is applied (Section 4.6.3).

The backbone incorporates two mechanisms for audio-gesture synchronization: adaptive temporal attention, which learns per-sample importance weights for beat-relevant audio frames by adding a learned bias to cross-attention scores before softmax, and windowed masking, which restricts each motion frame to attending only to local audio context ($\pm$100 ms), preventing long-range semantic leakage.

P1 is trained for 1100 epochs on a single RTX 3090 GPU, reaching FGD < 1.0 on BEAT2 validation. The jerk ratio (0.72) indicates mild over-smoothing characteristic of flow-matching models [76], and the variance ratio near 1.0 confirms the generator reproduces the training distribution's per-feature statistics. This checkpoint serves as initialization for P2. Training uses the conditional flow matching objective (Section 3.4) with timestep-weighted (t^2) auxiliary losses for bone length consistency, temporal smoothness, velocity matching, and jerk-distribution fidelity.

4.6.3 P2: Spatial Conditioning

P2 extends P1 to support 3D spatial target conditioning for pointing gestures while preserving co-speech motion quality. The design centers on a single architectural commitment: the spatial signal is added as a gated post-block residual to hidden states, never entering the AdaLN conditioning pathway. This keeps timestep and CSMP modulation identical to P1 and isolates spatial conditioning to an additive offset on the residual stream.

Design Rationale: Audio Handles Timing, Spatial Handles Direction Pointing gestures have two coupled but separable properties: *when* the apex occurs, and *where* the arm is aimed at that apex. P1 already learns the first, CSMP cross-attention carries the prosodic and lexical cues that trigger referential gestures. What P1 cannot do is aim the extension at a particular 3D location. The audio

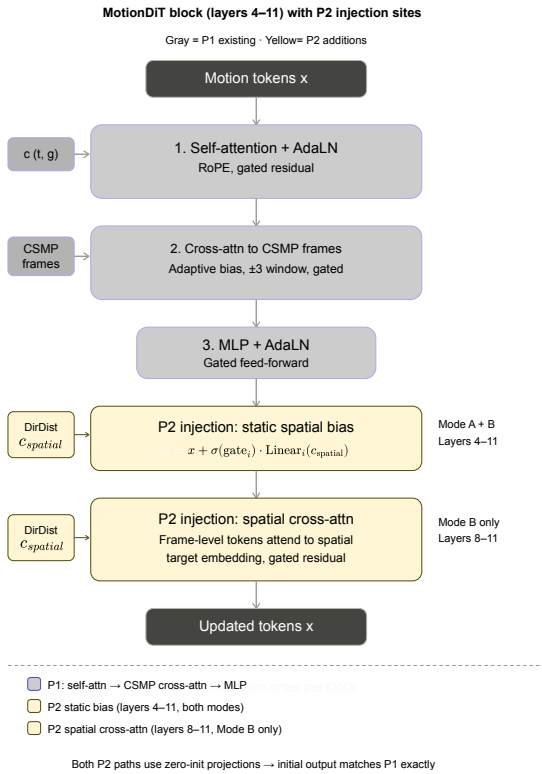


Figure 4.4: Single MotionDiT transformer block (layers 4–11) showing P1 co-speech components (gray) and P2 spatial injection sites (yellow). The spatial pathway adds a gated residual bias on layers 4–11 (Mode A and B) and single-token cross-attention on layers 8–11 (Mode B only). Both paths use zero-initialized projections, so the initial P2 output matches P1 exactly.

carries no information about the speaker's environment. P2's role is therefore strictly geometric: given that the audio is asking for a referential gesture, indicate the direction toward the target. The spatial signal is consequently constant across the sequence and requires no temporal envelope or peak-frame annotation at inference. When audio cross-attention drives arm extension, the spatial residual biases its direction; when it does not, no spurious pointing emerges.

Direction-distance encoding. The raw 3D target coordinate $\mathbf{t} \in \mathbb{R}^3$ (pelvis-relative) is encoded as:

$$c_{\text{spatial}} = \text{MLP}\left(\left[\frac{\mathbf{t}}{\|\mathbf{t}\|},\ \log(1 + \|\mathbf{t}\|)\right]\right) + \mathbf{e}_{\text{hand}} \in \mathbb{R}^{768}, \tag{4.6.4}$$

where the unit direction and log-distance are projected through a small MLP ($4 \to 256 \to 768$), and a learned hand embedding (left/right) is added to the output. This factored representation gives the encoder explicit access to direction and distance rather than requiring it to infer geometric structure from raw coordinates.

Spatial injection. The central design question is whether the pretrained backbone must be modified or whether a frozen adapter suffices. Two injection modes test this, both starting from the P1 checkpoint with zero-initialized new parameters so the initial output matches P1 exactly.

Mode A (frozen adapter): The P1 backbone is frozen. In layers 4–11, a per-layer linear projection maps c_{spatial} to a residual bias, gated by per-layer sigmoid scales:

$$\mathbf{x} \leftarrow \mathbf{x} + \sigma(g_\ell) \cdot W_\ell \, c_{\text{spatial}}, \tag{4.6.5}$$

where g_ℓ is initialized to -2.0 ($\sigma \approx 0.12$) and W_ℓ is zero-initialized. This tests whether pointing can be added without modifying the pretrained motion prior. Mode A achieves $\theta_{\text{eff}} = 36.2°$, confirming that the frozen backbone does not expose sufficient degrees of freedom for spatial pointing [8].

Mode B (unfrozen + cross-attention): The backbone is unfrozen. In addition to the static bias (Eq. 4.6.5), late transformer layers (8–11) receive single-token spatial cross-attention:

$$\mathbf{x} \leftarrow \mathbf{x} + \sigma(g_\ell^{\text{xattn}}) \cdot \text{CrossAttn}(\text{LN}(\mathbf{x}),\ c_{\text{spatial}}), \tag{4.6.6}$$

with zero-initialized output projections. Mode B has higher capacity and nearly doubles spatial accuracy relative to Mode A ($\theta_{\text{eff}} = 19.5°$), at the risk of degrading the co-speech prior if spatial gradients dominate. As shown in Section 4.6.4, this risk does not materialize: BEAT2 FGD is identical between P1 and P2 Mode B. All subsequent results use Mode B unless noted otherwise.

See Figure 4.4 for an architecture overview of P2.

Three properties of this formulation are load-bearing. First, the AdaLN conditioning vector c is identical to P1: it sees only timestep and CSMP, not the spatial signal. Second, the spatial contribution is a *post-block residual*: it is added after the block has run, with learned per-layer gating. Third, c_{spatial} is a single 768-dimensional vector broadcast identically across all N frames; the spatial offset is direction-constant across the sequence, a direct consequence of the rationale in Section 4.6.3. For inference-time controllability, c_{spatial} is zeroed with probability $p_{\text{uncond}} = 0.2$ during training, enabling classifier-free guidance over the spatial signal at inference. When no spatial target is present, c_{spatial} is zero and the post-block residual contributes nothing to any layer, so the model reduces precisely to P1.

Relational pointing losses. Spatial supervision uses relational losses rather than direct position regression, applied only to samples with spatial targets (SGS-HSI and MM-Conv pointing clips):

$$\mathcal{L}_{\text{align}} = 1 - \cos(\mathbf{d}_{\text{forearm}},\ \mathbf{d}_{\text{target}}), \tag{4.6.7}$$

$$\mathcal{L}_{\text{extend}} = \max(0,\ r_{\text{rest}} - r_{\text{current}}), \tag{4.6.8}$$

where $\mathbf{d}_{\text{forearm}}$ is the elbow-to-wrist direction, $\mathbf{d}_{\text{target}}$ is the elbow-to-target direction, and r measures arm extension relative to rest pose. These losses are weighted by t^2 (emphasizing near-clean timesteps) and masked by per-sample spatial weights (zero for BEAT2 samples in mixed batches). Without the relational loss, the flow-matching objective alone has no mechanism to reward directional accuracy.

Training recipe. P2 is initialized from the final P1 checkpoint and fine-tuned on a mixture of 70% pointing data (SGS-HSI + MM-Conv) and 30% BEAT2 co-speech data for prior preservation. For BEAT2 samples, c_{spatial} is zeroed at the embedding level and the relational loss is skipped; for pointing samples, c_{spatial} is computed from the target position and the relational loss provides a directional gradient on the wrist trajectory. Training uses audio dropout ($p_{\text{uncond_audio}} = 0.3$), explicit smoothness losses, and layer-wise learning-rate decay to protect the co-speech prior.

4.6.4 Evidence for Spatial-Temporal Separation

The architecture's design rests on the spatial and temporal pathways operating as separable channels: the CSMP audio pathway controls gesture timing, while the spatial pathway controls pointing direction. Three pieces of evidence support this separation.

First, an inference-time directional refinement that steers pointing direction during flow-matching sampling improves spatial accuracy by $\sim 8^\circ$ with $|\Delta\text{Score}_{1a}| \leq 0.012$ (temporal alignment; defined in Section 7.1.3) across three of four model variants [8]. This orthogonality is only possible if spatial corrections do not propagate into the model's temporal decisions. The refinement is gated by the model's own arm extension: at each Euler step, the per-frame correction is scaled by the current extension ratio $\|w - s\| / \max_f \|w - s\|$, so corrections can redirect an ongoing extension but cannot initiate one. The gating therefore leverages the channel separation rather than imposing it: directional adjustment is possible only because the temporal decision has already been made elsewhere in the architecture. The fourth variant's larger temporal change ($\Delta\text{Score}_{1a} = -0.071$) reflects apex-detector frame-shift on a model trained with cleaner, more consistently extended pointing data, rather than direct temporal influence from the spatial pathway [8].

Table 4.2: Co-speech preservation on BEAT2 validation (no spatial targets). P1: pretrained backbone; P2: spatially-conditioned model evaluated without spatial input. FGD computed against BEAT2 training distribution via EMAGE VAE latent.

Model	BEAT2 FGD ↓	BEAT2 Jerk ↓
P1 (co-speech only)	0.099	0.014
P2 Mode B	0.099	0.016

Second, spatial conditioning does not degrade co-speech gesture quality. Table 4.2 compares the P1 backbone and the P2 spatially-conditioned model on pure BEAT2 co-speech data (no spatial targets active). BEAT2 FGD is identical (0.099) and jerk increases only marginally (0.014 to 0.016), indicating that the spatial pathway does not overwrite the co-speech motion prior.

Third, attention diagnostics on trained checkpoints corroborate this picture: CSMP cross-attention in layers 4–7 shows peaked activation near the gesture apex, while spatial cross-attention in layers 8–11 is temporally flat, consistent with the spatial pathway operating as a frame-uniform directional bias rather than learning its own temporal structure.

Together these results support the thesis's central architectural claim: the semantic/spatial gap is tractable at the architecture level, with the two conditioning channels separable in a single generative model without functional cost. The CSMP pathway bundles lexical and prosodic content into a single speech-driven stream, so the timing–direction separation demonstrated here is the architectural manifestation of the broader semantic/spatial split discussed throughout the thesis. The same architectural lesson reappears on the comprehension side in Chapter 5: when linguistic identification and visual localization are routed through separate pathways, both improve. See Figure 4.5 for samples contrasting ground-truth motion versus MM-Conv-Flow generated pointing gesture at apex.

4.7 The Generality of the Conditioning Problem

This chapter has traced an architectural evolution in how conditioning enters the denoising process: from the U-Net's rigid encoder–decoder pathways, through DiffWave/LDA's flat residual stacks with single-stream injection, to the Diffusion Transformer's flexible combination of adaLN-Zero for global variables and cross-attention for sequential inputs of arbitrary length. Classifier-free guidance added a mechanism for balancing multiple conditioning streams at inference time. This trajectory, from fixed single-stream conditioning to flexible multi-stream conditioning is what made it possible to move from the audio-only generation of early systems to the joint semantic-and-spatial conditioning explored in Paper D and Paper E.

Figure 4.5: Qualitative comparison at the pointing apex on the held-out test room. Top: ground-truth motion capture. Bottom: MM-Conv-Flow. Each column shows the same target object. From [8].

The same architectural vocabulary has appeared in robotic action generation. Chi et al. [77] framed visuomotor control as conditional denoising over action sequences. More recent vision-language-action systems such as π_0 [78] and SpatialVLA [79] build on vision-language backbones to produce manipulation trajectories. Across these architecturally diverse systems, a structurally similar tension recurs: the vision-language backbone follows semantic instructions effectively but struggles with precise spatial specification when 3D structure is collapsed into flat visual features. SpatialVLA addresses this by adding egocentric 3D position embeddings to the vision encoder's semantic features so spatial information becomes an explicit input signal. S^2-Diffusion [80] makes the same point from the opposite starting condition: a diffusion policy conditioned on raw RGB overfits to instance-level appearance, and replacing the image with separated semantic masks and monocular depth restores category-level generalization.

The tension between semantic and spatial processing is not confined to generation. As Chapter 5 will show, it reappears on the comprehension side: current vision-language models for grounding conflate reasoning about *what* is being referred to with reasoning about *where* it is, and disentangling the two yields substantial gains. The next chapter turns from producing conditioned

behaviour to interpreting it: how agents can resolve the references that such behaviour is designed to convey.

5 Multimodal Grounding in Situated Dialogue

Chapters 3 and 4 examined how agents produce spatially grounded communicative behavior, focusing on the generation of gestures conditioned on linguistic and spatial context. This chapter addresses the complementary problem: how agents interpret such behavior in order to resolve references in situated interaction. The two directions are tightly coupled. Both rely on a shared set of underlying representations, discourse context, scene structure, and embodied signals such as gaze and gesture. The form of a referring expression is shaped by what a speaker assumes to be recoverable from context; interpretation, in turn, depends on precisely those contextual cues. A system that generates a pointing gesture toward a target and one that resolves a deictic reference by integrating gaze and dialogue history are engaged in two sides of the same communicative process. This chapter argues that grounding in situated dialogue is not a straightforward extension of visual object localization, but a problem whose structure reflects the multimodal, contextual nature of the communicative setting in which references arise.

The chapter moves from problem framing and a survey of current vision-language models (Sections 5.1–5.2) through the spatial and contextual factors that distinguish situated grounding from standard benchmarks (Sections 5.3–5.4), to the decoupling of referent identification from localization that emerges from the experiments in Paper F (Section 5.5). Section 5.6 then returns to the production–comprehension connection, showing that the identification–localization split on the comprehension side mirrors the semantic–spatial split identified in Chapter 4, and previews a pose-conditioned approach to referential grounding from concurrent work [9].

5.1 The Grounding Problem in Situated Dialogue

In the visual grounding literature, the task is typically formulated as a mapping from a referring expression to a region in an image. The expression is self-contained, written to be unambiguous in isolation; the visual input is a single static image; and the task is treated as one-shot retrieval with no notion of interaction history. In situated dialogue, these assumptions rarely hold. References emerge incrementally, shaped by what has been said, what is mutually

visible, and what participants can assume to be shared. A bare pronoun or a deictic phrase such as *that one* is not a defective expression, it is efficient precisely because the intended referent is recoverable from context.

The dominance of such expressions is not a marginal phenomenon. In the conversational data collected for this thesis, pronominal and deictic forms account for nearly half of all referring expressions. These forms are opaque in isolation: without dialogue history, scene context, or nonverbal behavior, there is no basis for determining what they refer to. Yet they are the dominant referential strategy in natural interaction, because speakers design their expressions against the common ground established over the course of the conversation (cf. Chapter 2).

Grounding in this setting cannot be reduced to perceptual matching. It requires access to the structure of the ongoing interaction, the dialogue so far, the current visual scene, and the nonverbal behavior of the interlocutors. This chapter examines how current vision-language models handle these requirements, and where they fall short when applied to conversational data.

5.2 Vision-Language Models for 2D Grounding

The dominant approach to visual grounding relies on vision-language models that jointly encode text and image to predict the location of a referred object. This section reviews how these models represent spatial information internally, since the encoding mechanism determines what kinds of spatial reasoning the model can perform, and where it will fail.

Open-vocabulary detectors. Earlier grounding models used separate text and image encoders whose representations are fused through cross-modal attention. Grounding DINO [81] combines a Transformer text encoder (BERT-like) with a Swin Transformer visual backbone and a DETR-style detection head. Spatial information enters only through the visual encoder's feature maps: the model has no explicit 3D coordinate representation, and spatial reasoning must emerge implicitly from the learned visual features. Cross-attention between text tokens and image features allows the model to attend to regions that match the linguistic description, but the spatial relations encoded in phrases such as *on the left* or *next to the lamp* must be recovered from the 2D feature geometry rather than from any explicit spatial encoding. This works well when the referring expression is descriptive and the target is visually distinctive, but provides no mechanism for reasoning about viewer-dependent relations or 3D scene structure.

Generative grounding architectures. More recent models unify visual and linguistic processing within a single sequence-to-sequence framework, encoding spatial output as discrete location tokens in the text vocabulary. Kosmos-2 [82]

represents bounding boxes as sequences of quantized location tokens interleaved with the generated text in a Markdown-link-like format, where each text span is paired with the location tokens describing its referent. Grounding is thereby reduced to next-token prediction over a vocabulary that combines text and discrete spatial tokens. Florence-2 [83] is a unified prompt-driven vision foundation model with an image encoder and a seq2seq multi-modality decoder, supporting tasks including detection, grounding, OCR, segmentation, and polygon/box outputs. These sequence-based architectures vary in how much conversational context they can accommodate. Among the models evaluated in this thesis, Qwen2.5-VL [84] is the most naturally suited to dialogue-history conditioning, as an instruction-tuned multimodal model with strong object localization, document and layout reasoning, and long-video and complex-input capabilities.

In all of these architectures, spatial information is encoded as 2D pixel coordinates, either implicitly in feature maps or explicitly as discretized (x, y) tokens. Depth, occlusion, and 3D spatial relations are not represented. The model sees a projection of the scene, not the scene itself. For self-contained expressions that describe visually salient objects (*the red chair*, *the person in the striped shirt*), this is often sufficient. For viewpoint-dependent references (*on the left*, *behind the vase*) or expressions that require reasoning about occluded or spatially ambiguous configurations, the 2D projection discards the information needed for disambiguation.

3D Vision-Language Grounding. A parallel line of work operates directly on 3D scene representations. Models such as ScanQA [85], 3D-LLM [86], and LL3DA [87] encode point clouds or scene graphs alongside text, giving them access to metric spatial relations that 2D projections discard. However, these models are also trained on single-turn, self-contained descriptions, ScanRefer queries, template-generated questions, and none is designed to handle the kind of multi-turn conversational reference found in MM-Conv, where pronominal and partitive expressions resolve against discourse history rather than being self-contained. The spatial representation gap and the dialogue context gap are largely orthogonal: 3D models address the first but not the second. Both the Look and Tell dataset (Paper G), with its reconstructed point clouds, and MM-Conv (Paper F), with its full 3D scene graphs and object meshes from AI2-THOR, provide resources for evaluating 3D grounding models on conversational language. The concurrent PoseRefer [9] work takes a step in this direction by ranking scene-graph candidates from 3D pose and utterance features, though a full evaluation of 3D VLMs on conversational data remains future work.

The benchmark mismatch. These models are trained and evaluated on datasets that reinforce the 2D, single-turn formulation. RefCOCO [17] and its variants consist of self-contained noun phrases written by annotators to be unambigu-

ous in isolation. 3D benchmarks such as ScanRefer [88] and ReferIt3D [89] provide richer spatial descriptions but remain single-turn, with no dialogue context or temporal dynamics. Models trained on such data learn to treat the input expression as a complete specification of the target. When faced with a pronoun such as *it* or a partitive [1] such as *the one on the left*, they lack mechanisms to recover the intended referent; the necessary information is in the preceding dialogue, the visual scene, or both. As the experiments in Paper F demonstrate, state-of-the-art VLMs achieve only 5–9% accuracy on pronominal references without dialogue context; even with context supplied in the input, performance remains far below human levels.

It is important to note that the grounding evaluation in this thesis is restricted to 2D vision-language models operating on single egocentric images. The end-to-end baselines evaluated in Paper F, GroundingGPT, Ferret, Kosmos-2, Florence-2, and Qwen2.5-VL, all take a 2D image as visual input. Grounding DINO appears only in the rewriting stage discussed in Section 5.5, where it serves as a pure detector without conversational capability.

5.3 Viewpoint, Gaze, and Spatial Representation

Most computational grounding work assumes a fixed, disembodied observer: a single image from a canonical viewpoint, with the model's task being to locate the referent within that frame. In practice, viewpoint shapes what counts as a referring expression, which cues are available for disambiguation, and whether a referent is even visible. A speaker who says *the jar on the left* while looking at a kitchen counter may be producing an unambiguous expression from their egocentric perspective, only one jar falls on the left of their visual field. From an exocentric camera on the wall, several jars may be visible with no reliable left–right ordering relative to the speaker. The spatial relation is anchored in the speaker's deictic center, and recovering its meaning requires knowing where the speaker is and how the scene projects from their point of view.

For embodied agents, this viewpoint dependence is not merely theoretical. An agent interpreting a user's instruction must understand the reference from the user's egocentric perspective, even when its own model of the world is constructed from a different vantage point. Yet few grounding benchmarks provide the paired viewpoints needed to study this.

The Look and Tell dataset (Paper G) addresses this gap by recording participants from both egocentric and exocentric viewpoints simultaneously. The egocentric stream, captured through wearable Aria smart glasses, provides a

[1] The label *partitive* is used throughout this thesis as a dataset-specific cover term for underspecified NPs that are neither full definite descriptions nor bare pronominals; including attribute-restricted references, meronymic references, and locative pro-forms. The usage is broader than the strict linguistic sense of *partitive* (e.g. *one of the boys*); see Section 6.4.4 for the full taxonomy.

(a) Egocentric RGB with ground-truth mask overlay (b) Instance segmentation mask (c) Metric depth map

Figure 5.1: A grounded data sample from MM-Conv. For the referring expression box, the dataset provides (a) the egocentric RGB view with the ground-truth referent highlighted, (b) the per-pixel instance segmentation mask, and (c) the corresponding depth map. Word-level timestamps link the expression to this exact simulator frame. Rendered in Blender.

first-person view including the speaker's gaze direction. The exocentric stream, from a stationary camera, provides a third-person view of the speaker and scene. Together, these streams support direct comparison: given the same utterance, does a grounding model succeed or fail depending on the viewpoint?

The choice of spatial representation interacts with viewpoint. A 2D image collapses depth and occludes objects behind foreground clutter; a 3D point cloud preserves metric structure and makes relations such as *behind* and *above* geometrically explicit. Look and Tell provides both: 2D video frames from both cameras and 3D point clouds reconstructed via the Aria MPS pipeline. This makes it possible to study whether gains from 3D structure interact with viewpoint, for instance, whether egocentric 3D representations are more informative than exocentric 2D ones for resolving viewpoint-anchored references.

Gaze adds a further modality. Speakers routinely fixate on a referent before or during the verbal mention, providing an attentional cue that listeners exploit for disambiguation. The gaze–speech synchrony analysis in Paper G quantifies this coupling across 2,504 linked mention–fixation pairs. Two regimes emerge. In 41% of cases, the gaze fixation fully preceded the verbal mention, ending before the speaker began to name the referent; in the remaining 59%, gaze and speech overlapped, with the speaker still fixating the referent as the mention began. The average overlap between a fixation and its corresponding mention was approximately 350 ms. Gaze does not deterministically predict the upcoming referent, but it provides a strong probabilistic cue that narrows the candidate set, particularly for pronominal and deictic expressions where language alone is insufficient.

5.4 Dialogue Context and Nonverbal Cues

The previous section argued that grounding is viewpoint-dependent. A closely related problem is that it is also context-dependent: the meaning of a referring expression is shaped by what has been said before, what is currently visible, and what both participants take to be mutually known. As noted in Section 5.1, context-dependent expressions dominate the MM-Conv data, but the empirical consequences of this dominance are worth examining in detail.

The human evaluation in Paper F makes the cost of missing context concrete. Across 1,940 stimuli evaluated by 78 crowd workers (three annotations each, after attention-check filtering), accuracy on full noun phrases without dialogue history was 73%, dropping to 48% on partitives and 37% on bare pronouns. For pronominal references without context, up to 51% of unanimous misses showed scattered clicks with no consensus among the three annotators; these expressions are not merely harder but effectively unresolvable from text and a single egocentric image alone. Providing the preceding five utterances raised pronoun accuracy to 55% and partitive accuracy to 61%, with tighter agreement and longer completion times suggesting active reasoning over the history. Notably, full-NP accuracy dropped from 73% to 62% with context added, indicating that for already explicit references, additional dialogue introduces distractors rather than disambiguation. Inter-annotator agreement (Krippendorff's α) ranged from 0.66 on full NPs without context to 0.42 on pronominals without context, the same gradient observed in machine performance. The pattern is clear: the more underspecified the expression, the more indispensable the dialogue context.

Nonverbal cues provide a complementary source of disambiguation. The human evaluation in Paper F provides indirect evidence: even without linguistic context, humans achieved 37% accuracy on pronominal references, expressions that are semantically empty in isolation. The egocentric images sometimes contain visible pointing gestures, and participants appear to have used them to infer the referent. In qualitative inspection of VLM outputs on the same images, we observed cases where models grounded to the hand itself rather than following the deictic gesture to its target, interpreting the pointing limb as the salient visual feature rather than as a directional cue. This pattern is suggestive rather than quantified, but it points to a structural asymmetry: humans treat a pointing hand as a communicative signal that directs attention outward, while current VLMs lack the inductive bias to do so.

This divergence connects to a broader point. When a speaker says *that one* while extending an arm toward a shelf, the gesture carries referential content that the pronoun alone cannot. Without the gesture, the pronoun is unresolvable; without the pronoun, the gesture lacks the discourse framing that signals referential intent. The two modalities are complementary in a strong sense, and any grounding architecture that ignores one channel, or fails to integrate

the two at decision time, will fail on precisely the expressions that are most common in natural dialogue.

Quantitative evidence for this complementarity comes from PoseRefer [9][2], which directly tests pose–language fusion for reference resolution. On MM-Conv's 55% pointing / 45% non-pointing split, pose-only scoring achieves 27.9% top-1 on pointing references but collapses to 7.7% on non-pointing expressions, while text-only scoring (frozen MiniLM utterance and category embeddings) shows the opposite pattern: 23.2% on pointing, 27.2% on non-pointing. Neither modality alone captures the task's full structure. A decoupled late-fusion architecture with pathway-local parameters—pose and text pathways share no learned weights and use independent per-object category embedding tables—reaches 31.9% aggregate, exceeding both singletons across every reference type and every gesture-strength tier. The fusion gate's learned weight α flips from 0.84 (trusting pose heavily) to 0.22 (trusting text) depending solely on whether the text pathway has per-object category access, a consistent result across cross-validation folds that confirms pathway decoupling makes ablations interpretable.

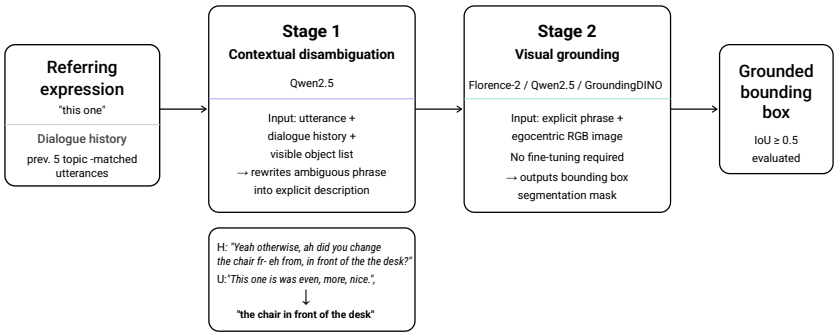
5.5 Decoupling Referent Identification from Localization

Table 5.1: End-to-end VLM grounding performance on MM-Conv, stratified by referring expression type and dialogue context condition (w/o vs. w/ preceding five utterances). Match rate at IoU 0.5 / 0.3 (M@.5 / M@.3); best M@.5 per row in **bold**. GGPT = GroundingGPT; Flor.-2 = Florence-2; Qwen2.5 = Qwen2.5-VL. Pronominal accuracy collapses across all models without context, and even with context remains far below the 55% human baseline reported in Section 5.4.

Expr. Type	Ctx	GGPT	Ferret	Kosmos-2	Flor.-2	Qwen2.5
Full NP	w/o	37.9/47.3	39.1/45.2	42.6/48.1	46.3/52.6	**51.8**/60.9
	w/	28.2/41.1	40.9/47.9	42.0/48.2	28.0/33.1	**53.2**/62.0
Partitive NP	w/o	14.2/16.3	14.7/16.7	16.5/17.9	16.3/18.8	**22.5**/26.4
	w/	14.7/20.9	18.3/21.6	19.7/22.2	21.6/23.6	**29.6**/34.2
Pronominal	w/o	4.7/6.9	5.1/7.4	6.3/9.1	**9.2**/10.1	5.9/8.3
	w/	10.7/16.0	12.3/15.4	9.3/11.4	25.8/28.5	**30.4**/38.0

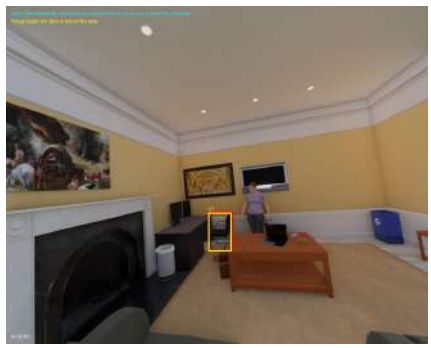
The preceding sections established that grounding in situated dialogue involves two kinds of reasoning: linguistic reasoning to determine *what* is being referred to, and visual localization to determine *where* it is in the scene. Current VLMs attempt both in a single forward pass. The experiments in Paper F show that this conflation is the primary source of failure on conversational language.

[2] Results are presented as supporting evidence rather than as a primary thesis contribution.



(a) Two-stage rewrite-based grounding pipeline.

(b) Kosmos-2 end-to-end: grounds to desk (red, IoU = 0.018) despite dialogue history.



(c) GroundingDINO with rewrite: correctly localizes the chair (IoU = 0.901).

Figure 5.2: Decoupling referent identification from localization. (a) The rewrite pipeline resolves "This one" to "the chair in front of the desk" before grounding. (b) Without rewriting, Kosmos-2 fails to exploit dialogue context. (c) With the rewritten query, GroundingDINO succeeds despite having no conversational capability.

Following the expression-type taxonomy used in the MM-Conv annotation pipeline (described in Section 6.4.4; see Figure 5.1 for a representative sample), results are reported separately for full noun phrases, partitive/attribute references, and pronominal references. These categories make different demands on a grounding system. Full noun phrases reduce to lexical-visual alignment of the kind RefCOCO and ScanRefer were designed to evaluate: the expression itself specifies the referent, and the task is matching it to a region. Pronominals lie at the opposite extreme, they carry almost no lexical content, and their resolution depends on context external to the expression itself, established in the preceding discourse or in the shared situation through gaze and gesture. Partitives sit between the two, carrying partial lexical content (a colour, a part name, a spatial term) that constrains but does not determine the referent. An

Table 5.2: Rewrite-based Grounding Results. Referring expressions are rewritten by Qwen2.5-VL then grounded. M@.5 shown with improvement over baseline in parentheses.

	Full NP		Partitive NP		Pronominal	
Model	Base	Rew.	Base	Rew.	Base	Rew.
Qwen2.5-VL	53.2	54.4 (+1.2)	29.6	40.8 (+11.2)	30.4	50.3 (+19.9)
Florence-2	28.0	49.1 (+21.1)	21.6	39.7 (+18.1)	25.8	48.9 (+23.1)
GroundingDINO	—*	**61.1**	—*	**49.5**	—*	**56.7**
Avg. Gain†		+11.2		+14.7		+21.5

*No conversational capability; baseline not applicable.

†Average gain for Qwen2.5-VL and Florence-2 only.

aggregate accuracy number averages across these distinct failure modes, stratification is what makes them visible.

Table 5.1 summarizes end-to-end VLM performance. With dialogue context, the strongest baseline (Qwen2.5-VL) reaches 30.4% on pronominal references, better than the 5–9% without context, but far below the 55% human baseline. The critical question is whether the bottleneck is perceptual or linguistic. Paper F disentangles the two through a rewriting experiment: each conversational expression is rewritten into an explicit noun phrase using only the dialogue history and no visual input (e.g. *it* → *the wall painting*), and the rewritten query is passed to a grounding model in place of the original.

The rewriting produces consistent improvements across models and reference types (Table 5.2). Qwen2.5-VL's pronominal accuracy jumps from 30.4% to 50.3%, Florence-2 from 25.8% to 48.9%. Grounding DINO, a detection model with no capacity for dialogue, achieves 56.7% on pronominal references with rewritten queries, nearly double the best end-to-end baseline with context. The same pattern holds for partitives. On successfully matched samples, all models achieve mean IoU of 77–89%, confirming that the perceptual capability is intact. Even on unmatched samples, predictions tend to land on plausible objects in the scene rather than scattering randomly, indicating that the failure is upstream of localization rather than a breakdown of visual perception. The challenge is identification, not localization, and the rewriting pipeline addresses identification directly. This mirrors the human-eval pattern from Section 5.4, context helps precisely on the underspecified forms (pronominals and partitives) and slightly hurts on full NPs, where it introduces distractors rather than disambiguation.

Validation shows 83.4% of rewrites correctly preserve the intended referent, with quality varying by expression type: 87.2% for pronominals, 84.6% for full NPs, and 67.4% for partitives. The gap between pronominal and partitive fidelity is informative about what each category demands of the rewriter. For

pronominals, the rewriter's task is to resolve the expression against available context; primarily dialogue history, though purely deictic uses (resolved in the original interaction through gaze or pointing) are not recoverable from text alone. The system must recover, for example, that *this one* in "This one was even more nice" [sic] refers to the chair that had been the topic of the preceding turns, and produce *the chair in front of the desk* (Figure 5.2). For full noun phrases, the task is often little more than name normalization: *that teapot* becomes *the kettle on the dresser* once the visible-object list confirms the speaker has misnamed the object. Both have a single correct answer that the dialogue context or scene inventory can supply. Partitives are heterogeneous: the category as defined here covers part-of references (*the cloud* in a painting, *the leg* of a table, *his hat* on a statue), attribute restrictions (*the green one*, *the yellowish one*, *the small one*), and spatial deixis (*there*, *up there*). These subtypes likely have distinct failure modes, segmentation-granularity ceilings for part-of references, attribute-matching errors for restrictions, deictic-resolution gaps for spatial expressions, but the validation set is too small for stratified analysis. Subtype-resolved evaluation of partitive grounding is left to future work. Downstream grounding accuracy nonetheless tracks rewrite quality: correctly rewritten expressions achieve 55.0% match rate at IoU 0.5, against 23.4% for incorrect rewrites, confirming that the rewriting step is causally responsible for the gain rather than an artifact of the evaluation.

The structural explanation is straightforward. These VLMs were trained on self-contained referring expressions, image–caption pairs and annotator-written queries, and have no learned mechanism for extracting referent identity from dialogue history, because their training data never required it. Providing context in the input is necessary but not sufficient without the inductive biases or training signal to reason over it.

This observation should not be read as an argument against end-to-end architectures in principle. The decoupling experiment reveals where current models fail, on linguistic reasoning under ambiguity, not that joint processing is inherently impossible. Whether future models can close this gap will depend on training data and objectives that explicitly reward contextual disambiguation alongside perceptual localization.

5.6 Connecting Production and Comprehension

In natural interaction, production and comprehension draw on the same representations, discourse state, scene structure, and embodied signals, and constrain each other: the form of a referring expression reflects what the speaker assumes is recoverable from context, and comprehension exploits exactly those cues.

This coupling is reflected concretely in the thesis data. The MM-Conv dataset supports both the grounding benchmarks in Paper F and the grounded gesture

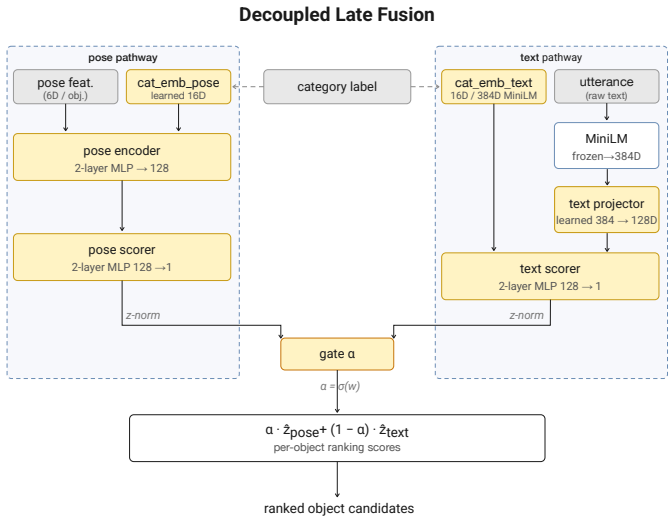


Figure 5.3: PoseRefer architecture. Pose and text pathways operate independently with no shared parameters. Each pathway has its own per-object category embedding table (learned 16-D for pose, frozen MiniLM 384-D for text). Per-pathway scores are z-normalized and combined by a learned scalar gate α. The decoupling makes pathway contributions directly ablatable: disabling category on one pathway leaves the other unchanged.

generation in Paper E, where the same referential annotations define conditioning targets for generation and evaluation targets for comprehension. The same data simultaneously serves both directions of the communicative process.

The architectural parallel is equally concrete. On the generation side, Chapter 4 showed that semantic and spatial conditioning serve distinct functional roles, speech shapes *what kind* of gesture is produced and *when* it occurs, while 3D targets shape *where* it is directed. On the comprehension side, the decoupling result in Section 5.5 reveals an analogous structure: linguistic reasoning determines *what* is being referred to, while visual localization determines *where* it is. In both directions, the two subproblems differ in computational character and benefit from distinct treatment in the architecture.

This architectural claim admits more than one realization. The rewrite-based pipeline of Section 5.5 is one response to the identification–localization split: a sequential, two-stage approach built on large pretrained models. Concurrent work [9] explores the opposite end of this design space. PoseRefer is a lightweight scoring network (73K parameters) that ranks scene-graph candidates on MM-Conv from 3D body pose and the utterance, addressing referential ambiguity through gesture geometry rather than through linguistic rewriting

against dialogue history. The pose-only variant of this architecture also serves as the Axis 2 learned ranker in the Paper E benchmark (Section 7.1.3).

The architecture maintains strict pathway decoupling (Figure 5.3): pose and text scoring pathways share no learned parameters, and per-object category is represented by independent embedding tables in each pathway (learned 16-D on pose, frozen MiniLM [90] 384-D on text). Pose features encode temporally pooled angular proximity between body channels (arms, head, torso) and object centroids over the utterance window; the text pathway receives only the utterance embedding and per-object category labels, with no access to world-frame geometry. A learned scalar gate combines z-normalized scores from both pathways. Results favor this late fusion (31.9% top-1) over both pose-only (18.8%) and text-only (25.0%) baselines, with gains consistent across all reference types and gesture-strength tiers.

The pattern is consistent with the rewrite results of Section 5.5: in both cases, maintaining modality-specific computation outperforms forcing it through a shared representation. The two architectures occupy different points in the deployment trade-off space: the rewrite pipeline prioritizes accuracy on underspecified expressions at the cost of multi-stage inference, while a pose-conditioned scoring network targets the latency budget of interactive systems. Together they begin to map out the architectural options the dataset makes available.

These parallels suggest that both sides of situated communication require a shared spatial-semantic environment representation, one that captures object identities, spatial relations, and attentional affordances without the full geometric fidelity of a reconstruction pipeline. Unlike manipulation, where contact surfaces and object geometry are critical, resolving a reference such as *the jar on the left* requires primarily relational and categorical spatial information. The datasets developed in this thesis, with their combination of 2D frames, 3D point clouds, and gaze data, provide resources for investigating what level of spatial detail communicative tasks demand, a question taken up in Chapter 8.

6 Data for Situated Interaction

6.1 Existing Datasets and Their Limitations

Research on spatially grounded communication in embodied agents draws on data from several established communities, each of which has developed resources around a different aspect of the problem: communicative motion, understanding visual scenes through language, or modeling how bodies move through physical environments. Understanding where these resources overlap and where they leave gaps is essential for motivating the data contributions of this thesis. Table 6.1 provides a structured comparison.

Co-speech gesture datasets. The first family is co-speech gesture datasets. Resources such as the *Trinity* [101, 91], *BEAT2* [92], and the *Talking with Hands* dataset [93] have enabled major advances in speech-driven gesture generation by pairing motion capture with synchronized audio and, in some cases, text transcriptions. These datasets capture the temporal and semantic coupling between speech and body movement, making them well suited for modeling beat, iconic, and metaphoric gestures. However, they record speakers in isolation or in minimal settings, with no representation of the surrounding scene. Gestures are produced and modeled as if the body were communicating in a void: there are no spatial targets, no object layouts, and no environmental geometry against which deictic or referential behavior could be grounded.

Human–scene interaction datasets. The second family comprises human–scene interaction datasets. Resources such as *PROX* [95], *HUMANISE* [97], *CIRCLE* [102], *GRAB* [96], and *ARCTIC* [103] capture spatially aware agent behavior within 3D environments, providing a combination of motion, together with scene meshes, object annotations, and contact information. These datasets have driven progress in locomotion, navigation, hand–object manipulation, and physically plausible motion synthesis. They model how bodies move through, act upon, and interact with environments, sitting, walking, reaching, grasping, rather than how bodies *communicate* within them. Conversational multi-party settings, spoken dialogue, and gesture–speech coordination are generally outside their scope.

Table 6.1: Comparison of dataset families relevant to spatially grounded communication. The first three groups summarize existing resources for gesture generation, human–scene interaction, and referential grounding, the fourth lists the thesis contributions.

Dataset	Sp/Tx	MoCap	3D	Tgt	Dial.	NV	Scale
Co-speech gesture (generation)							
Trinity II [91]	✓	✓	–	–	–	–	∼6 h
BEAT2 [92]	✓	✓	–	–	–	–	∼60 h
Talking w/ Hands [93]	✓	✓	–	–	–	–	∼100 h
Embody 3D [94]	✓	✓	–	–	✓	–	∼500 h
Human–scene interaction							
PROX [95]	–	✓	✓	–	–	–	100k frames
GRAB [96]	–	✓	✓	✓	–	–	∼3.8 h
HUMANISE [97]	✓*	✓	✓	✓	–	–	∼6.5k clips
Referential grounding (understanding)							
ScanRefer [88]	✓	–	✓	✓	–	–	51.6k utt.
ReferIt3D [89]	✓	–	✓	✓	–	–	125k utt.
YouRefIt [98]	✓	–	–	✓	✓	✓	4k ref. inst.
TEACh [99]	✓	–	✓	–	✓	–	3k dial.
ScanERU [100]	✓	–	✓	✓	–	✓‡	46k utt.
This thesis							
Pointing dataset (Paper B)	✓	✓	–	✓	–	✓	∼140 clips
Aug. pointing (Paper C)	✓	✓	–	✓	–	✓	∼1.2k clips
Look and Tell (Paper G)	✓	–	✓	✓	✓	✓	2.7k ref. / 3.67h
MM-Conv (Paper F)	✓	✓	✓	✓	✓	✓	4.2k ref. expr. / 6.7h

Columns: Sp/Tx = speech or text transcription; MoCap = full-body motion capture; 3D = 3D scene geometry; Tgt = spatial target annotations; Dial. = multi-turn dialogue; NV = nonverbal cues (gaze, gesture annotations). A dash indicates the property is absent.

* Language-conditioned motion; text instructions rather than conversational speech.

‡ Synthetic gesture annotations.

Multimodal referential grounding benchmarks. A third family of resources targets the comprehension side: resolving referring expressions in visual scenes. *ScanRefer* [88] and *ReferIt3D* [89] provide large-scale textual descriptions of objects in reconstructed 3D scenes, but the descriptions are single-turn, annotator-written, and designed to be unambiguous in isolation; there is no dialogue context, no embodied behavior, and no temporal dynamics. *YouRefIt*

records real-world human interactions with referring expressions accompanied by gesture and video context, but at limited scale. *TEACh* [99] captures multi-turn task-oriented dialogue in simulated 3D environments with embodied navigation and manipulation, but the interaction revolves around instruction execution rather than referential communication. *ScanERU* [100] augments 3D scene data with synthetic gesture annotations, enabling the study of gesture-aided grounding, but remains single-turn. Each of these benchmarks captures some of the properties needed for studying grounding in situated interaction, but none combines them all. The resources designed for large-scale evaluation lack embodiment and dialogue dynamics; those that include dialogue or gesture are limited in scale or scope.

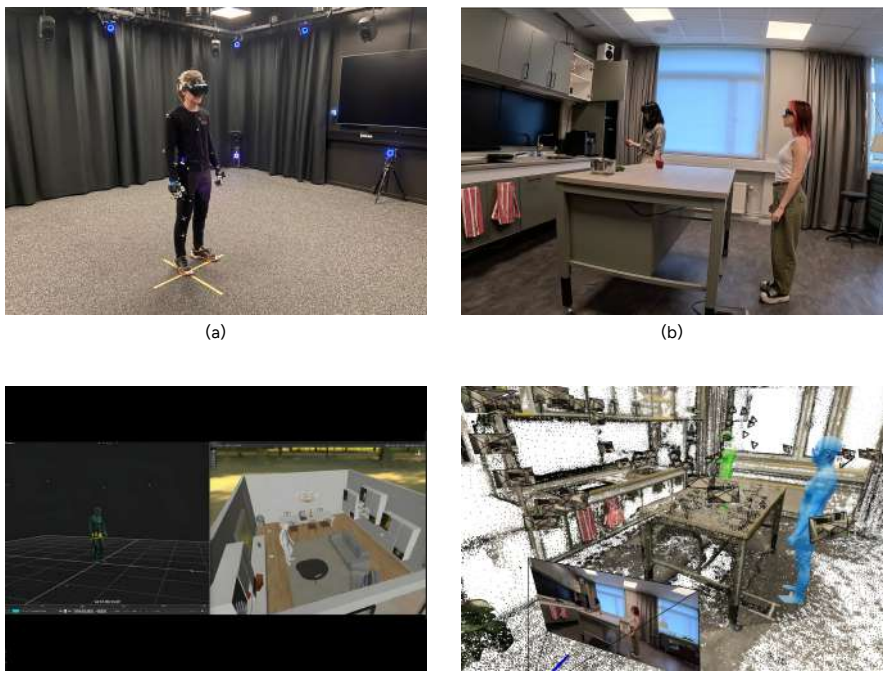


Figure 6.1: Data collection pipelines used in this thesis. **Left column:** Controlled motion caption studio recording with VR setup. (a) An actor wearing a motion capture suit and VR headset in an OptiTrack studio. (c) The captured skeleton is streamed into the AI2-THOR simulator via Motive, producing synchronized motion and scene data for the MM-Conv dataset (Paper F). **Right column:** Ecologically valid recording with ego and exo cameras. (b) Two participants engaged in a referential task, with the speaker wearing Meta Aria smart glasses. (d) COLMAP based 3D point cloud of the environment with registered body poses via sam-3d-body, forming the spatial backbone of the Look and Tell dataset (Paper G).

The combined gap. Taken together, these three families reveal a fragmented landscape. Co-speech gesture datasets provide communicative structure but lack spatial grounding. Human–scene interaction datasets provide spatial structure but lack communicative context. Referential grounding benchmarks provide visual reasoning tasks but vary widely in their coverage of dialogue, embodiment, and nonverbal cues, and none was designed to support both generation and understanding within the same data. In particular, to our knowledge, no existing resource provides synchronized speech, full-body motion, gaze, multi-turn dialogue history, 3D scene context, and referential annotations within the same dataset. The datasets introduced in this thesis are designed to bridge these gaps: by recording multimodal interaction in controlled yet naturalistic 3D environments, through VR, AR, and synthetic augmentation, the later papers provide data that supports the study of communicative behavior that is simultaneously speech-driven, spatially grounded, and accessible from both production and comprehension perspectives.

Recording modalities and ecological validity. The fragmentation described above is not only a matter of research focus; it is also a consequence of how motion data is collected, and the trade-offs that different capture pipelines impose. Optical motion capture (e.g. OptiTrack, Vicon) remains the gold standard for skeletal accuracy, particularly for fine-grained hand and finger tracking, but it constrains recording to instrumented studios with limited environmental complexity. This is the primary reason why co-speech gesture datasets capture rich communicative behavior in spatially impoverished settings: the studio that enables precise tracking is the same studio that prevents naturalistic scene interaction.

At the other end of the spectrum, monocular 3D human mesh recovery (HMR) has advanced rapidly. Single-image methods such as SAM 3D Body [104] can now recover full-body pose and shape, including hands and feet, from a single photograph, showing improved robustness to occlusion and challenging viewpoints relative to earlier methods. At the video level, recent work on joint human–scene reconstruction has begun to recover SMPL body meshes, camera trajectories, and dense scene geometry within a shared world coordinate frame. Methods such as SLAHMR [105], which combines visual SLAM with learned motion priors, and JOSH [106], which jointly optimizes human motion and scene point clouds using contact constraints, can estimate spatially grounded 3D motion data from in-the-wild footage. HSfM [107] has further shown that integrating human-body constraints into Structure from Motion improves both camera pose accuracy and scene reconstruction. Together, these advances open the door to ecologically valid recordings in unconstrained environments, but at a cost: estimated poses are noisier than marker-based capture, temporal motion is less precise than optical mocap, and the resulting annotations typically lack the synchronized speech and dialogue structure needed for studying communicative behavior. Resources such as *Motion-X++* [108] ex-

emplify this trade-off, achieving massive scale through video-based annotation pipelines while accepting reduced accuracy compared to studio capture.

A middle ground is emerging through wearable sensing devices and immersive platforms. Meta's Project Aria glasses, for instance, provide synchronized egocentric video, eye tracking, IMU data, and spatial mapping from a lightweight head-worn device, enabling data collection in natural environments with richer perceptual grounding than a fixed camera but without the instrumentation burden of a full motion-capture suit. VR-based recording offers a different compromise: participants interact in controlled but spatially rich 3D scenes, providing ground-truth scene geometry and object identity alongside tracked motion, at the cost of reduced ecological realism compared to real-world settings. The choice of recording modality involves a three-way trade-off between motion accuracy, environmental richness, and ecological validity. The datasets in this thesis exploit several of these modalities, optical mocap for the pointing dataset (Paper B, Paper C), VR-based capture with streamed mocap for MM-Conv (Paper F), and Aria-based egocentric recording with exocentric HMR for Look and Tell (Paper G), each chosen to match the specific requirements of the research question being addressed. Figure 6.1 contrasts the two primary capture setups used in this thesis.

The scale problem. Beyond the content gap, there is also a scale problem specific to multimodal synthesis. Training a joint model that produces both speech and gesture from text requires parallel data across all three modalities, text transcriptions, speech audio, and motion capture, recorded from the same speaker in the same session. Existing corpora that meet this requirement are small: the largest, *Talking with Hands* [93], contains approximately 100 hours, but with quality limitations [74]; the highest-quality resource used for joint synthesis, *Trinity Speech-Gesture Dataset II* [91], provides only around 6 hours. This is orders of magnitude less than the data available to state-of-the-art unimodal systems in speech synthesis or text generation.

Synthetic Data for Joint Speech-and-Gesture Synthesis Mehta et al. [10] address this bottleneck by chaining strong unimodal synthesizers, an LLM for text generation, a neural TTS system for speech, and the diffusion-based gesture generator from Paper A for motion, to construct a wholly synthetic corpus of approximately 38 hours of parallel speech-and-gesture data. The gesture-generation stage relies on the CSMP module introduced in Paper A, whose speaker-independent representations allow it to produce plausible motion from synthetic speech without domain-mismatch degradation. Pre-training a joint synthesis model on this material and fine-tuning on the target data yields significant improvements in both speech naturalness and gesture human-likeness.

This synthetic data strategy operates entirely within the co-speech domain: the

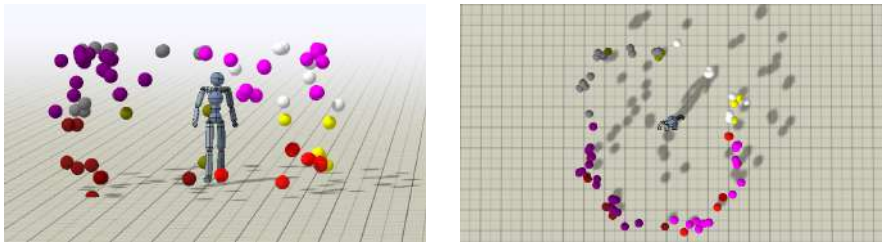


Figure 6.2: Target position distribution in the single-target pointing task, shown from front (left) and top (right) views. Colors indicate spatial quadrants (left/right, top/bottom, front/back). From Paper B, IsaacGym.

generated gestures are beat and iconic gestures produced without spatial context. The following sections describe how related strategies are applied to spatially grounded data, beginning with the collection of a pointing gesture dataset that provides the spatial targets absent from existing co-speech corpora.

6.2 The Pointing Gesture Dataset

While the synthetic augmentation described above addresses the volume of available co-speech data, it does not address the content gap identified in Section 6.1: the absence of spatial targets against which deictic behavior can be grounded. The pointing gesture dataset introduced in Papers B and C provides the first component needed to bridge this gap.

The dataset was recorded in an optical motion capture studio equipped with 16 OptiTrack Prime 41 cameras. A single actor wore a suit with 50 passive markers, a pair of Manus data gloves for hand capture, and a head-mounted iPhone 12 for synchronized face and voice recording. Crucially, the recording also tracked a physical pointing target, a rigid structure with 4 markers, whose 3D position was solved alongside the actor's skeleton using Motive 3.0. This co-registration of body motion and target location in a shared coordinate frame is what distinguishes this dataset from existing co-speech gesture corpora.

Three pointing tasks were recorded: single-target pointing, two-target selection, and two-target point-and-place. In each task, the target was moved to different positions around the stationary actor to achieve broad spatial coverage. The dataset contains ≈ 140 single-target pointing clips, with target positions distributed across front, back, left, right, top, and bottom quadrants (Figure 6.2). Both dominant and non-dominant hands are represented.

Analysis of the pointing movements reveals several properties relevant to downstream modeling. The velocity and precision profiles follow characteristic bell-shaped curves, reflecting the raise–hold–retract structure of pointing motion. The dominant hand exhibits higher peak and mean velocities than the non-dominant hand, as well as longer movement duration. Importantly, neither

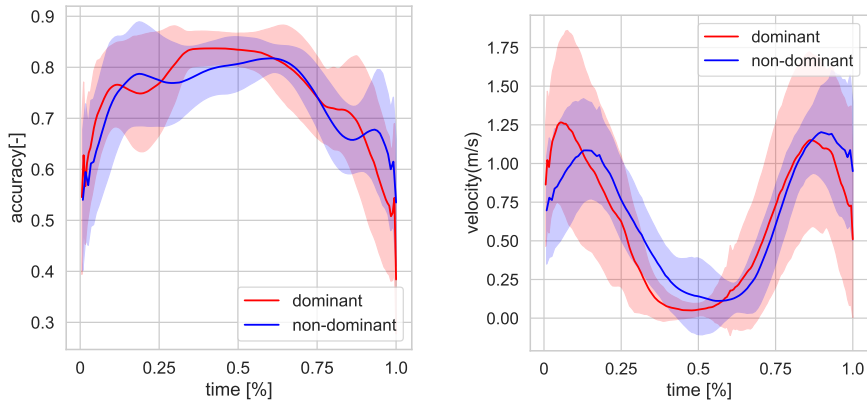


Figure 6.3: Averaged pointing accuracy (a) and hand velocity (b) profiles for front targets, shown for dominant (red) and non-dominant (blue) hands. Accuracy is defined as the inverse of the alignment angle between the forearm vector and the hand-to-target vector. Both profiles reflect the raise–hold–retract structure of pointing motion; neither hand reaches maximum accuracy. Shaded regions indicate ± 1 standard deviation. From Paper C.

hand achieves maximum alignment-angle accuracy: human pointing is inherently imprecise in geometric terms (Figure 6.3), a finding consistent with prior studies [109]. This observation has direct implications for how pointing accuracy should be defined and rewarded in generative models, as discussed in Section 4.4.

The dataset also includes synchronized speech, consisting of simple demonstrative expressions such as *"This one."* and *"That one over there."*. While minimal in linguistic complexity, these utterances provide temporal anchors for the onset of deictic reference and are later replaced with richer synthetic speech in the augmentation pipeline described in the following section.

To increase the effective size of the dataset, all motion clips were mirrored (left–right), doubling the available training material. The dataset has served as the foundation for reinforcement-learning-based pointing generation (Paper B), synthetic spatial augmentation (Paper C), and the unified generation models developed in Paper D and Paper E.

6.3 Synthetic Augmentation of the Pointing Data

The pointing gesture dataset described in the previous section provides spatially grounded motion with tracked 3D target positions, but its speech component is limited to short, formulaic demonstratives such as *"This one!"* or *"Give me that one over there!"*. These utterances lack the lexical and prosodic richness needed to train a speech-driven gesture generation model. Paper C addresses this limitation by applying the same synthetic augmentation philoso-

phy demonstrated for co-speech data in [10], but now in the service of spatially grounded generation. Large co-speech gesture corpora contain rich motion and speech but no spatial supervision; the pointing dataset provides spatial targets but minimal speech. These sources do not align directly, and Paper C uses synthetic augmentation to bridge them.

The goal is to pair each pointing motion clip with a more naturalistic referring expression, one that could plausibly accompany a deictic gesture directed at an object in a shared environment. The pipeline has three main stages: data extension, text-and-speech generation, and temporal alignment.

Data extension. The original dataset contains approximately 140 isolated single-target pointing clips, ranging from 0.5 to 9 seconds in duration. To increase effective size, all motion files were mirrored left-to-right and each clip was time-stretched at three rates sampled from the range $[0.7, 1.6]$, yielding a total of 1,160 augmented clips with a mean duration of 4.85 seconds.

Text and speech generation. GPT-4 [110] was used to generate exophoric referring expressions, utterances involving demonstratives (*this*, *that*) embedded in conversational sentences that describe or draw attention to an object. To encourage stylistic consistency with the co-speech gesture corpus that would later be combined with this data, the transcripts of that corpus were provided as contextual input to the language model. In total, 1,400 sentences were generated. Speech was then synthesized from these sentences using a commercial TTS engine, with SSML emphasis tags applied to the demonstrative words to produce prosodically appropriate stress patterns.

Temporal alignment. The synthesized speech segments were matched to the augmented pointing clips using the Hungarian algorithm. The cost function jointly penalized differences in total duration and misalignment of the demonstrative event, that is, it favored pairings where the stressed demonstrative in the speech coincided temporally with the pointing stroke in the motion. After matching, the speech segments were padded with leading and trailing silence to equal the length of the corresponding motion clips, ensuring that the demonstrative remained temporally aligned while the speech was fully contained within the animation.

The resulting dataset combines the pointing gesture corpus with an existing co-speech gesture dataset [91], producing a training set that contains both beat gestures driven by spontaneous monologue and pointing gestures driven by synthetic referring expressions. This combined corpus served as the first benchmark for models capable of generating both co-speech and deictic gestures from a unified framework, and was used to train the spatial gesture generation models described in Chapter 4.

Figure 6.4: Overview of the VR recording setup. Left: the Unity-based AI2-THOR simulator, where the actor's motion-capture data is streamed over a network connection. Right: the motion-capture studio with two actors engaged in conversation.

A key limitation of this approach is that the spatial grounding remains implicit: the synthetic utterances describe generic objects rather than specific items in a visible scene, and the target positions are abstract coordinates rather than locations of identifiable referents. The VR-based data collection described in the following section was designed to overcome precisely this limitation, by recording referential communication within richly annotated virtual environments.

6.4 VR-Based Multi-modal Conversational Data Collection

The synthetic augmentation described in the previous section enriched pointing gestures with naturalistic speech, but the spatial grounding remained implicit: target positions were abstract coordinates, and the accompanying utterances described generic objects rather than identifiable referents in a visible scene. For the grounding work in Part II, and for training generation models that must condition on scene context, what is needed is data in which referential behavior unfolds within a richly annotated environment, where every object has a known identity and position, and where the linguistic form of each reference can be linked to the spatial and discourse context from which it arose. Virtual reality provides a practical way to collect such data. It enables full-body tracking and synchronized capture of speech and gesture, while also providing precise access to the geometry and identity of every object in the scene. At the same time, it allows participants to interact in shared three-dimensional environments that support natural referential behavior. The *MM-Conv* dataset (Paper F) was developed with this goal in mind.

6.4.1 Interaction Design

Data was collected in dyads following an instruction-giver/follower paradigm common in cognitive studies of referential communication [111]. The main actor, equipped with a full-body motion capture suit, finger-tracking gloves, and a VR headset with gaze and facial tracking, described objects and spatial arrangements within simulated apartment environments rendered in the AI2-THOR simulator [112]. The interlocutor responded naturally but was instructed not to introduce new referents, in order to elicit a high density of spontaneous referring expressions and associated nonverbal behaviors from the main actor.

Five apartment environments were selected from AI2-THOR, each containing a diverse set of interactable objects (≈ 50 unique objects per room). For each room, two to three short role-play scenarios were designed to vary the discourse goals while maintaining ecological validity: showing off a new apartment to a friend, a landlord inspecting the space, and an interior designer offering suggestions. This variation encouraged a natural mix of referential strategies: explicit full noun phrases, underspecified partitive references, and context-dependent pronouns.

6.4.2 Recording Setup and Synchronization

The capture system integrated several hardware components to produce time-aligned multimodal data streams. Skeletal motion was recorded using an OptiTrack system with 16 Prime x41 cameras, tracking 50-marker skeletons for both participants. Finger movements were captured using MANUS Quantum Mocap MetaGloves, and a Meta Quest Pro headset provided binocular gaze tracking and 52 facial blendshapes for the main actor. The virtual scenes were rendered in real time in the AI2-THOR simulator running in Unity, and the actor's motion-capture data was streamed into the simulator over a network connection so that the virtual avatar moved in synchrony with the physical actor (see Figure 6.4).

Synchronization across all streams was achieved using SMPTE timecodes generated by Tentacle Sync E devices, injected into both the audio recordings and the Unity simulation. The headset pose was calibrated against the motion-captured head position to ensure alignment between the egocentric visual stream and the physical tracking data. This frame-accurate synchronization is what makes it possible to map from a specific word in the transcript to the exact simulator frame at which it was spoken, and from that frame to the RGB image, depth map, and object segmentation visible from the speaker's viewpoint at that moment.

Table 6.2: Modalities recorded in the MM-Conv dataset.

Modality	Details
Speech / Text	WhisperX transcripts with word-level timestamps
RGB	Egocentric frames rendered from main actor's viewpoint
Depth	Per-pixel metric depth maps
Segmentation	Per-pixel object IDs and ground-truth masks
Full-body motion	OptiTrack 50-marker skeletons (both participants)
Hands	MANUS MetaGloves finger tracking (main actor)
Face	52 facial blendshapes (main actor)
Gaze	Binocular gaze vectors from Meta Quest Pro
Scene graph	Per-frame object instances with 3D metadata

6.4.3 Recorded Modalities

The combination of egocentric vision, full 3D scene access, and dense nonverbal recordings distinguishes *MM-Conv* from prior referential benchmarks, which typically provide only text and RGB or lack continuous multimodal streams (cf. Table 6.1). Table 6.2 summarises the recorded modalities.

6.4.4 Annotation Pipeline

Transforming the raw recordings into a structured benchmark required a multi-stage annotation pipeline covering speech transcription, reference identification, and visual grounding. See Figure 6.5 for an overview of the annotations steps.

Speech transcription and alignment. Audio was transcribed using WhisperX [113], prompted with a manually transcribed portion of the recording to encourage retention of filled pauses, repetitions, and other disfluencies characteristic of spontaneous speech. The resulting transcripts were reviewed and corrected by human annotators using Label Studio, with audio chunks segmented by WhisperX's voice activity detection module. Non-speech events such as laughter and spoken noise were color-coded in the annotation interface for efficient triage. Because only one of the speakers was a native English speaker, annotators also checked for effects of accented speech on the ASR output. Corrected transcripts were then force-aligned at the word level using a CTC-based aligner, providing precise timestamps that link each word to its corresponding simulator frame.

Reference annotation. A two-stage pipeline using GPT-4o was used to identify and classify referring expressions. In the first pass, GPT-4o produced topic

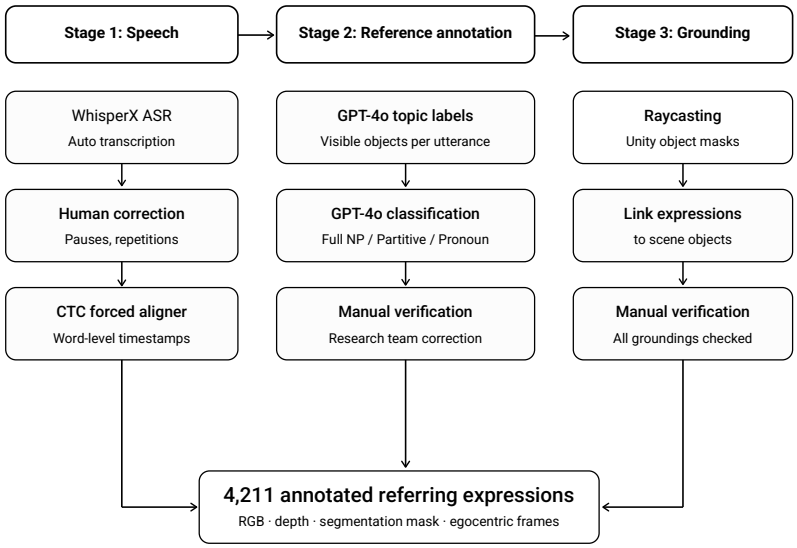


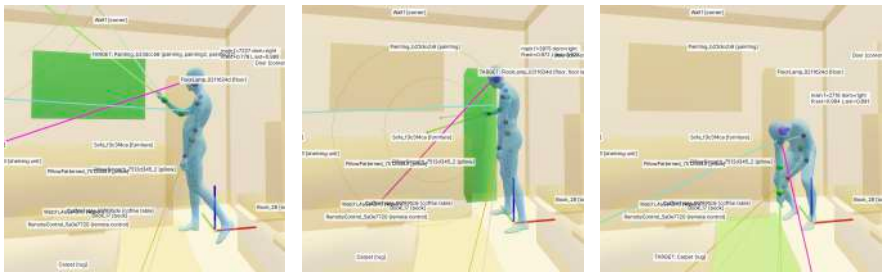
Figure 6.5: Annotation pipeline for the MM-Conv dataset. Each stage combines automated processing with manual verification.

annotations for each utterance based on the transcription and the set of visible objects in the Unity scene, tracking discourse focus and shifts in attention across the conversation. In the second pass, it parsed utterances to extract referring expressions and classify each into one of three categories drawn from the referring-expression literature [114]: full noun phrases as definite descriptions (explicit, uniquely identifying descriptions such as *the black sofa in the corner*), partitive/attribute noun phrases (underspecified references to parts, features, attributes, or spatial indicators, such as *the cloud in a painting*, *the yellowish one*, or *there*), and pronominal references (expressions like *it*, *that*, or *those* that depend on discourse context and shared perceptual attention). Together, these categories span a continuum from lexically explicit to contextually dependent reference. A terminological note: the label *partitive* is used in this dataset as a broader cover label for any underspecified NP that is neither a full definite description nor a bare pronominal, including attribute-restricted references (*the green one*), meronymic references (*the leg of the table*), and locative pro-forms (*there*). These subtypes have distinct linguistic structure and likely distinct failure modes (Section 5.5), but were grouped under a single label for annotation tractability. All GPT-4o classifications were manually verified by the research team.

Visual grounding. Each referring expression was grounded to its target object using raycasting and Unity-derived per-pixel object masks, then manually verified for correctness. Word-level timestamps were used to identify the exact simulator frame corresponding to each reference, from which synchronized egocentric RGB images, depth maps, and segmentation masks were extracted. Frames were rendered in Blender for high-quality egocentric imagery. The pipeline also supports alternative viewports, such as third-person views, for future tasks.

Gesture reference annotation. Pointing gestures are identified automatically by a random-forest classifier on 16 SMPL-X kinematic features (arm extension, wrist velocity, finger configuration, shoulder-wrist alignment), trained against manually annotated reference cases. The classifier separates clear pointing from no-gesture with AUC 0.97 on held-out speakers. Classifier scores stratify references into four confidence tiers: Tier 1 (clear, unambiguous pointing; $n = 1{,}344$), Tier 2 (moderate-confidence pointing with less precise alignment; $n = 558$), Tier 3 (ambiguous cases; $n = 153$), and Tier 4 (no-gesture or non-deictic motion; $n = 1{,}543$). A post-hoc Tier 5 ($n = 166$) captures two-arm pointing cases that the single-arm classifier underweights. Throughout subsequent analyses we use the binary partition *pointing* (T1 ∪ T2 ∪ T5, $n = 2{,}068$, 55%) versus *non-pointing* (T3 ∪ T4, $n = 1{,}696$, 45%).

To validate the automatic classifier and correct edge cases, we developed an interactive annotation tool using the Viser 3D visualization framework. The tool displays synchronized SMPL-X body meshes alongside ground-truth bounding boxes extracted from the scene entity data, computes real-time angular deviation between the pointing ray and the vector toward the target centroid, and allows annotators to correct object assignments, adjust gesture peak frames, and modify dominant-arm labels. This enables iterative refinement (Figure 6.6). The tier boundaries are not always clean, particularly when referential and iconic gestures co-occur within the same utterance. A recurring challenge is utterances in which the speaker narrates an action involving the referent, producing gestures that are simultaneously directed at the target and depicting the described action. For example, in one recording the speaker says: *"I put my good thoughts in the blue bin. And every morning I open the bin... or the window, sorry. I'm a bit tired, so I open the window and I pour out the bad thoughts, so every day I'm getting rid of my bad thoughts from the day before, and then I... do symbolic wash in the blue bin and cover myself in good thoughts."* Figure 6.7 shows three frames from this sequence. The first shows a clear deictic point toward the bin (Tier 1). In the second, the hand performs an iconic opening motion directed away from the target. In the third, the speaker leans forward to enact a washing gesture whose trajectory converges on the target location, producing motion that is simultaneously iconic in form and spatially grounded in the referent. The annotation scope was restricted to referential function precisely because attempting a full gesture-type taxonomy



(a) (b) (c)

Figure 6.6: Gesture reference visualizer in the Viser-based tool. Each panel shows the SMPL-X mesh with object bounding boxes (green). (a) Pointing toward a painting (Tier 1: clear alignment). (b) Pointing toward a floor lamp with the target bounding box highlighted. (c) Bending gesture toward the carpet, illustrating a less canonical pointing pose.

(a) (b) (c)

Figure 6.7: Referential–iconic gesture blending within a single utterance, in which the participant describes ritually washing themselves in a garbage bin. Green: target object (`blue garbage can`). (a) Clear deictic point toward the bin (Tier 1). (b) Hands drawn to the chest in a self-directed iconic gesture ("cover myself in good thoughts"). (c) Leaning forward with arms extended toward the bin, an iconic washing motion that remains spatially grounded in the referent.

would require adjudicating such cases frame by frame, a level of granularity that the present dataset does not support reliably.

6.4.5 Dataset Statistics

The dataset contains 4,211 referring expressions produced by the main speaker and 1,547 by the interlocutor, for a total of 5,758 (see Table 6.3). The main-speaker subset is the primary unit of analysis throughout this thesis and Paper F, since the interaction protocol was designed to elicit dense referential behavior from this role; unless otherwise specified, all subsequent statistics refer to this subset. After filtering cases where the target object was not visible at the moment of reference, 4,001 main-speaker expressions remain for analysis. A defining characteristic of the data is the prevalence of context-dependent

Table 6.3: Distribution of referring expressions by type and speaker role in MM-Conv. Single/-multiple indicates whether the expression refers to one or more objects.

	Main speaker			**Interlocutor**		
	Full	Part.	Pron.	Full	Part.	Pron.
Single	1,215	435	1,591	248	48	901
Multiple	348	135	487	137	44	169
Total	1,563	570	2,078	385	92	1,070

language: pronominal references account for 49.4% of all main-speaker expressions (2,078 of 4,211), and 69.2% of interlocutor expressions (1,070 of 1,547).

This distribution reflects the natural dynamics of situated dialogue, where speakers frequently use pronouns and deictic forms to refer to objects already established in the shared attentional context. The data also reveals systematic differences across discourse roles: main speakers, who introduced objects into the conversation, produced relatively more full noun phrases, while interlocutors relied heavily on pronominal and deictic forms to maintain common ground.

These statistics have direct consequences for evaluation. Nearly half of all references in the dataset are opaque without access to the preceding conversation, which means that any grounding system evaluated on this data must contend with genuine contextual ambiguity rather than the self-contained descriptions typical of existing benchmarks. The grounding experiments that exploit this property are discussed in Chapter 5.

6.4.6 Dual Role of the Dataset

The *MM-Conv* dataset serves two complementary roles. As a benchmark for multimodal grounding, its annotated referring expressions, egocentric imagery, and dialogue context support the evaluation of context-aware reference resolution in Paper F where the dataset's high proportion of ambiguous references directly exposes the limitations of current vision-language models (Chapter 5). As a data resource for grounded generation, its synchronized motion, speech, and scene data provide the situated recordings needed by Paper E, where it is combined with the augmented pointing corpus (Section 6.3) to train models that produce spatially grounded referential gestures conditioned on language, motion, and 3D scene context. The same referential annotations that define evaluation targets for the grounding task simultaneously define conditioning targets for the generation task, making the dataset the empirical bridge between the two research threads.

Figure 6.8: The Look and Tell recording setup. Left: a participant wearing Meta Project Aria smart glasses during a referential task in the KTH Interaction Lab. Right: COLMAP-based 3D point cloud of the kitchen environment with registered egocentric trajectory, forming the spatial backbone of the dataset (Paper G).

VR also proved useful beyond data collection. In the pointing evaluations of Paper B (Chapter 7), VR served as a perceptual evaluation environment, where participants judged not only whether a motion looked natural but also whether it successfully communicated a target in space. This dual use, as both data source and evaluation medium, reflects the broader argument of the thesis that spatially grounded communication should be studied in settings where the spatial context is perceptually available.

6.5 Egocentric Data Collection with AR Glasses

Paper G introduces the *Look and Tell* dataset, which shifts from the controlled virtual environments of *MM-Conv* to a real-world kitchen setting recorded with augmented-reality hardware. Twenty-five participants (18 female, 7 male; aged 22–37) were recruited at the KTH Interaction Lab. Each participant wore Meta Project Aria smart glasses, which simultaneously captured egocentric RGB video (up to 1408×1408 pixels at 30 fps), eye-tracking data (fixation events and gaze vectors), and speech audio. Stationary GoPro cameras provided complementary exocentric video, yielding a dual-view recording setup that pairs the participant's first-person perspective with an observer's third-person view of the same scene (see Figure 6.8).

The task was designed to elicit naturalistic referential communication. Each participant memorized a step of a recipe and then instructed a conversational partner to identify the corresponding ingredients among items laid out in the kitchen. Across five recipes per participant, the procedure yielded 125 sessions totaling 3.67 hours of synchronized recordings (396,208 RGB frames). Sessions lasted on average 1.8 ± 0.7 minutes, with individual sessions ranging from 0.7 to 4.3 minutes.

Annotations were produced through a multi-stage pipeline. Speech was transcribed at the word level using WhisperX [113]. Referential mentions, in-

gredient names, pronouns, coreferences, and spontaneously mentioned objects, were extracted via GPT-based prompting [110] constrained to a JSON schema, followed by deterministic post-processing for normalization, alias enrichment, and coreference chain construction. For visual grounding, the multimodal model Molmo [115] localized referenced objects in egocentric frames, and SAM2 [116] propagated segmentation masks across the full mention interval. In total, the dataset contains 2,707 annotated mentions ($\sim$22 per session). Ingredients account for 62% of mentions, while pronouns and coreferential forms make up 23%, reflecting the prevalence of indirect reference in spontaneous speech. The remainder comprises additional objects, distractors, and unresolved mentions. Coreference chains average 6.7 mentions, indicating that speakers frequently re-reference items once introduced.

Beyond the 2D annotations, the dataset includes reconstructed 3D point clouds of the kitchen environment, obtained via the Meta Aria MPS service and canonicalized into a shared coordinate system. Egocentric trajectories were registered to the COLMAP world frame through Sim(3) transformations estimated via hierarchical localization [117], achieving $\geq 70\%$ inlier correspondence for 87% of sessions with a mean alignment error of 9.1 cm. Ego and exo streams were temporally synchronized using audio cross-correlation. This alignment places gaze vectors, speech transcripts, segmentation masks, egocentric and exocentric video, and 3D scene geometry in a single spatially and temporally consistent reference frame.

A gaze–speech synchrony analysis revealed two distinct regimes. In 41.1% of mentions, the gaze fixation ended before the verbal mention began; in the remaining 58.9%, fixation extended past mention onset. The mean offset between fixation termination and mention onset was -189 ms (median -102 ms, SD $= 541$ ms), with the negative mean driven by a heavy tail of early-preceding fixations. The average temporal overlap was 352 ms. These findings suggest that speakers frequently begin an utterance while still fixating on the referent rather than consistently shifting gaze before speaking, a pattern relevant for models that use gaze as a predictive cue for reference resolution.

The *Look and Tell* dataset complements *MM-Conv* along two axes. First, it moves from virtual to real environments, introducing the noise, visual complexity, and physical constraints of an actual kitchen. Second, its dual-view design with 3D reconstruction enables systematic comparison of spatial representation choices, 2D versus 3D, egocentric versus exocentric, for grounding, a comparison that few existing datasets support. These properties make it a benchmark not only for gaze–speech analysis but also for evaluating how different spatial representations affect multimodal grounding performance, as explored in Chapter 5.

6.6 From Data to Benchmarks

Collecting data is only part of the challenge. To support progress in situated interaction, datasets must be turned into benchmarks with clear task definitions, train–test splits, and evaluation metrics. This is especially important when the same data serves both generation and comprehension, as different communities may otherwise adopt incompatible assumptions about context, candidate objects, or output format. The benchmarking efforts in this thesis span both directions.

On the generation side, Paper E introduces a benchmark for scene-aware referential gesture generation built on MM-Conv and SGS-HSI, hosted as a challenge at the 1st Workshop on Human–Scene Interaction at ECCV 2026. SGS-HSI contributes 1,135 clean single-target pointing clips recorded with the same five actors as MM-Conv in simplified setups without scene-graph distractors, providing supervised signal for the geometric primitive of pointing at an arbitrary 3D coordinate. The benchmark pairs a two-track task formulation, target-aimed pointing (Track A, given a coordinate) and scene-aware referential gesture (Track B, given a scene graph), with a three-axis spatio-temporal grounding protocol: temporal alignment (did the gesture peak at the communicatively appropriate moment?), spatial accuracy (did the pose point at the right place?), and referent recall (did the pose uniquely indicate the intended referent among scene-graph distractors?). This decomposition exposes a failure mode invisible to distributional metrics: a generator can produce geometrically accurate pointing that nonetheless fails to disambiguate the target from nearby objects, and the divergence between axes is diagnostic of where the system breaks down.

On the comprehension side, MM-Conv is framed not only as a dataset of multimodal interaction but as a benchmark for context-aware grounding in 3D dialogue (Paper F). The task setup specifies what counts as input context (utterance with or without preceding dialogue history), what the system must predict (a bounding box over the referred object), and how success is measured (IoU-based match rates stratified by expression type). This formulation helps expose where current vision–language models succeed and where they fail, especially on the pronominal and partitive references that dominate natural dialogue.

Together, these benchmarks define complementary problem spaces for the same underlying data: one asks whether a generated gesture communicates the right referent, the other asks whether an observer can recover the referent from language and vision. By constructing evaluation frameworks that jointly involve language, gesture, gaze, and scene context, the thesis encourages the field to move beyond simplified formulations and toward richer models of situated communication.

7 Evaluation

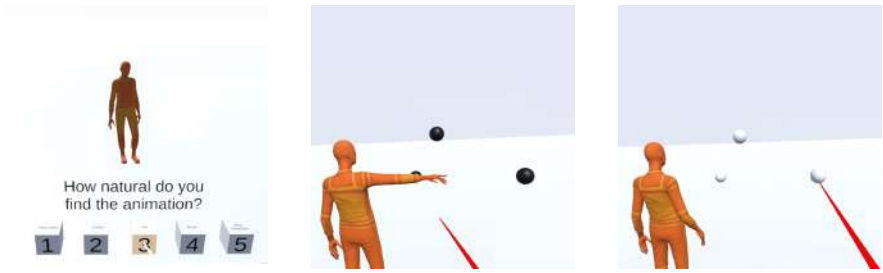
Evaluating spatially grounded communicative behavior exposes two shortcomings in current practice. First, perceptual evaluations of generated gesture are almost exclusively conducted in 2D: motion is rendered from a fixed viewpoint and presented as a flat video. For co-speech beat gestures this is adequate, but for referential behavior such as pointing, 2D projection destroys the information that defines success, the three-dimensional relationship between arm configuration, target position, and scene layout. Second, existing evaluation protocols ask quality questions (how natural? how appropriate to the speech?) rather than functional ones (did the gesture communicate the intended referent?). A gesture may score highly on naturalness while failing entirely as a referential act. This chapter briefly surveys the evaluation landscape for generated motion, then describes the contributions of this thesis that address these gaps.

7.1 Evaluating Human Motion Generation Models

Generated motion is typically assessed along two axes. Automatic metrics (Fréchet-based measures such as FGD, diversity scores, beat-alignment, and physical-plausibility statistics) are cheap and reproducible but correlate poorly with subjective human-likeness ratings, FGD being the main exception [118]. Human perceptual evaluation, formalised at scale by the GENEA challenges [74], remains the more reliable indicator. Both approaches share two limitations consequential for the work in this thesis: they rely on 2D video, and they assess quality rather than communicative function.

7.1.1 Evaluating Pointing in VR: The Referential Game

Paper B addresses both gaps by introducing a VR-based referential game for evaluating pointing gestures. The evaluation has two stages. In the first, participants rate the naturalness of pointing gestures without seeing the target object, so that judgments reflect motion quality alone. In the second, participants observe the gesture in the presence of one target and two distractors placed 20–40 cm apart and identify the pointed object. Figure 7.1 shows the participant's view in both stages. This design disentangles motion quality from communicative success: it can detect cases where a model produces natural-looking mo-



(a) (b) (c)

Figure 7.1: User view in the VR-based perceptual test. (a) Stage 1: the user rates the naturalness of the pointing gesture in isolation. (b) Stage 2: the agent points with one target and two distractors visible. (c) The user then selects the object they believe was indicated.

tion that nevertheless fails referentially, or where spatially accurate behavior is accompanied by unnatural dynamics.

The referential game yields a concrete, interpretable metric, identification accuracy, rather than a subjective rating. Results showed clear separation between models: AMP-PFNN achieved 95.5% referential accuracy compared to 45.9% for the supervised baseline (MoGlow), while ground-truth motion capture reached 85.6%. That a learned system exceeded ground-truth accuracy reflects the RL framework's direct optimization for target-directed behavior, though it also raises the question of whether a purely geometric reward fully captures communicative adequacy.

7.1.2 2D vs. VR Evaluation

That the evaluation medium itself has measurable effects was confirmed in a comparative study by Werner et al. [11], which evaluated three top-ranking submissions from the GENEA Challenge 2023, including the system from Paper A, alongside ground-truth motion capture, in both 2D and VR settings with 30 participants. VR ratings were slightly but significantly higher overall ($p = 0.0068$, Wilcoxon signed-rank test), and ground-truth motion was more clearly distinguished from generated systems in VR, suggesting that the immersive setting provides a more sensitive baseline. Importantly, relative system rankings were preserved across the two media, so VR evaluation does not distort model comparisons but increases sensitivity to the spatial aspects of motion quality. This sensitivity is especially consequential for referential behavior, where the 2D projection used in standard evaluations collapses precisely the spatial relationships that define communicative success. The VR setting introduced practical costs, hardware requirements, in-person participation, longer sessions, that limit scalability relative to web-based 2D studies, supporting its use for spatially grounded behavior while standard 2D evaluations remain adequate for

general co-speech gesture.

7.1.3 Spatio-Temporal Grounding for Scene-Aware Gesture Generation

The evaluation gap is not confined to the comprehension side. Paper E introduces a three-axis protocol for scene-aware referential gesture generation that decomposes communicative success into independently measurable components.

Axis 1a (temporal alignment) measures whether the system's gesture peak falls within the ground-truth hold region, identified kinematically from forearm extension. A logistic-regression apex detector, trained on the benchmark training split, recovers the hold on 85.5% of ground-truth clips, establishing a metric ceiling of Score_{1a} = 0.909.

Axis 1b (spatial accuracy) measures extent-aware angular error between the pointing ray and the target bounding box at the detected apex. The angular kernel width ($\sigma_{\text{angle}} = 20^{\circ}$) is calibrated to scene-graph ambiguity: the mean object angular diameter at target distance is $\sim 30^{\circ}$, so one standard deviation of pointing error corresponds to missing the target among typical nearby distractors.

The product of temporal and spatial scores yields the combined spatio-temporal grounding score (STG = $\text{Score}_{1a} \times \text{Score}_{1b}$), preventing accuracy inflation: a gesture that accidentally passes through the correct pointing vector at the wrong time scores near zero.

Axis 2 (referent recall) measures whether the generated pose uniquely indicates the intended referent among scene-graph distractors, reporting top-K recall over all candidate objects. The benchmark defines two rankers: a geometric Heuristic scorer (ray angle, head direction, distance; no learned parameters), used for official challenge scoring to ensure transparency, and a 73K-parameter learned MLP trained under leave-one-room-out cross-validation, reported as an analysis metric. The two can diverge on the same generated motion – accurate pointing that fails to disambiguate, or imprecise pointing that nonetheless ranks the target first – and the gap between them localises the failure mode.

Together, the three axes extend the diagnostic logic of the rewrite-based paradigm from comprehension to production: rather than asking "how good is the motion," the protocol asks "where does the communicative intent break down."

7.2 Evaluating Referential Grounding in Situated Dialogue

The evaluation paradigms described in the preceding sections address motion quality and communicative success in gesture generation. Evaluating referen-

tial grounding on the comprehension side poses a structurally different problem: the output of a grounding system is a spatial prediction, but the difficulty of producing that prediction depends on the linguistic form of the input expression in ways that aggregate accuracy scores conceal.

7.2.1 Match rate and expression-type stratification.

The primary metric used in Paper F is match rate at IoU thresholds of 0.5 and 0.3 (M@.5, M@.3), computed against ground-truth object masks derived from the MM-Conv annotation pipeline. The key methodological choice is to stratify results by expression type, full noun phrases, partitive noun phrases, and pronominal references, rather than reporting a single aggregate. This stratification is not merely a reporting convention; it is what makes failure modes visible. A model that achieves 50% aggregate accuracy may be performing near ceiling on full noun phrases while failing almost entirely on pronouns. Without the breakdown, the two cases are indistinguishable, and the nature of the deficit, linguistic rather than perceptual, goes undiagnosed. The distribution of expression types in MM-Conv, where pronominal references account for nearly half of all expressions, means that aggregate scores are dominated by the hardest case, making stratification essential for interpreting system behavior.

7.2.2 The rewrite-based decoupling paradigm.

The primary diagnostic contribution of Paper F is a method for separating the linguistic reasoning component of grounding from the visual localization component. The core procedure is simple: each conversational referring expression is rewritten into an explicit, self-contained noun phrase using only the dialogue history and no visual input, and the rewritten query is passed to a grounding model in place of the original. If performance improves substantially, the bottleneck was linguistic, the model could localize the referent once it was explicitly described, but could not resolve what was being referred to from conversational language. If performance does not improve, the bottleneck is perceptual.

This decoupling is methodologically general. It applies to any grounding system and any dataset where conversational expressions can be rewritten against dialogue context, and it yields a cleaner diagnosis than end-to-end evaluation alone. The validation step, checking what proportion of rewrites correctly preserve the intended referent, is necessary to interpret the downstream results: gains on incorrectly rewritten expressions conflate improved localization with changed targets. The 83.4% rewrite fidelity observed in Paper F, and the strong performance gap between correct and incorrect rewrites (55.0% vs 23.4%), confirm that the method is diagnostic rather than confounded.

The decoupling methodology generalizes beyond the rewrite-based approach. Concurrent work on pose–language fusion [9] demonstrates an architectural

form of the same principle: pathway-local parameters (no shared encoders, no shared category embeddings between pose and text) turn "what does category contribute in each pathway?" into a controlled experiment. When category is ablated on one pathway, only that pathway's learned representation changes, unlike shared-encoder architectures where ablating category simultaneously alters every downstream signal. This design choice revealed a confound invisible in prior work: the learned fusion gate commits to opposing policies (α = 0.84 vs. 0.22) depending solely on whether the text pathway has per-object category access, and category representation choice (learned 16-D vs. frozen MiniLM 384-D) changes fusion accuracy by 3.7 points while leaving text-alone accuracy statistically indistinguishable. Both findings are diagnostic rather than performance-competitive, but they illustrate the methodological value of architectural decoupling for localizing failure modes.

7.2.3 Human baseline design.

The human evaluation in Paper F is structured as a lower bound on multimodal grounding performance, not an upper bound on machine performance. Annotators saw an egocentric image and a referring expression under two conditions, without dialogue history and with the preceding five utterances, stripped of the gaze, gesture, and 3D scene access available to participants in the original interaction. The with/without structure mirrors the context conditions applied to the models, making human and model results directly comparable, and the gap to full multimodal access bounds the upside the dataset's additional modalities offer future evaluations. The finding that on 51% of pronominal expressions all three annotators missed and their clicks landed on different non-target objects establishes that these references are effectively unresolvable from image and text alone, regardless of model architecture.

Across both production and comprehension, the common thread is that evaluation must be designed to localize failure modes, not merely to rank systems. The VR-based referential game (Paper B) replaces quality ratings with a functional measure, whether the gesture communicates its intended referent. The 2D–VR comparison validates the medium itself, establishing that immersive evaluation preserves system rankings while increasing sensitivity to spatial aspects of motion quality. The spatio-temporal grounding protocol (Paper E) extends functional evaluation to scene-aware generation, decomposing referential gesture quality into temporal alignment, spatial accuracy, and referent recall. Expression-type stratification and rewrite-based decoupling (Paper F) then shift the diagnostic question from *how well* a grounding system performs to *where* it fails, separating linguistic from perceptual bottlenecks in a way aggregate accuracy conceals.

8 Conclusions and Future Directions

8.1 Summary of Contributions

This thesis examined spatially grounded communication in embodied agents across two complementary research threads: the generation of communicative behavior and the understanding of situated reference.

The three research questions posed in Chapter 1 can now be answered directly.

RQ1: How can data-driven gesture generation models produce co-speech gestures that reflect the semantic content of speech? Contrastive speech–motion pretraining (Paper A) demonstrates that semantic content can be injected into co-speech gesture generation through a joint text–motion embedding trained with a CLIP-style objective. The resulting representation, when used as a conditioning signal for a diffusion-based generator, achieved the highest human-likeness rating, statistically indistinguishable from motion capture, and the highest speech-appropriateness rating in the GENEA Challenge 2023. This establishes the semantic conditioning foundation that the spatially grounded work extends.

RQ2: How can generative models produce pointing gestures that are both spatially accurate with respect to 3D targets and natural in their motion dynamics? Paper B–Paper E show that the answer depends on the learning paradigm. In reinforcement learning with adversarial motion priors (Paper B), a phase-conditioned discriminator and a geometric pointing reward produce gestures that achieve 95.5% referential identification accuracy in a VR study. In offline generation, synthetic data augmentation (Paper C), unifying pointing and co-speech data within a shared format (Paper D), and a flow-matching architecture with gated spatial injection (Section 4.6) progressively close the gap between co-speech naturalness and spatial precision. The benchmark introduced in Paper E operationalizes the evaluation of this trade-off through a spatio-temporal grounding protocol.

RQ3: How do spatial representations and multimodal context affect the grounding of referring expressions in spontaneous situated dialogue? Paper F demonstrates that context-dependent expressions, which account for nearly

half of all references in naturalistic data, are the primary failure mode of current vision–language models. The rewrite-based decoupling experiment in Paper F isolates the bottleneck as linguistic rather than perceptual: once the referent is explicitly described, even simple visual detectors localize it accurately. Gaze provides a complementary disambiguation cue, preceding the verbal mention in 41% of instances (Paper G). Together, these results show that grounding in situated dialogue requires reasoning over dialogue history, nonverbal behavior, and scene structure, not just visual linguistic matching. The effect of spatial representation choice on grounding accuracy, 2D versus 3D, egocentric versus exocentric, is structurally supported by the Look and Tell dataset design but not yet systematically evaluated; this remains future work (Section 8.3).

Beyond the individual methods introduced in the appended papers, three findings cut across the thesis as a whole.

The semantic–spatial disconnect is a recurring architectural pattern. In Paper F, rewriting underspecified references into explicit noun phrases before grounding allows even a pure object detector with no conversational capability to outperform the best end-to-end VLM with full dialogue context, isolating the bottleneck as linguistic rather than perceptual. On the generation side, the architectural separation between CSMP cross-attention (timing) and gated spatial residuals (direction) in MM-Conv-Flow [8] preserves BEAT2 FGD (0.099/0.099) while halving angular pointing error. In both directions, maintaining modality-specific computation outperforms forcing it through a shared representation.

The choice between online and offline learning shapes what conditioning mechanisms are available. Chapter 3 framed this as a contrast between intelligence in the loop and intelligence in the architecture. Reinforcement-learning methods (Paper B) can afford simple policy networks because the physics simulator provides the corrective feedback that an offline model would have to encode in its weights. Diffusion-based methods (Paper A, Paper D, and the MM-Conv-Flow system in Section 4.6) must instead carry all spatial and semantic knowledge architecturally, which motivates the progression from single-stream conditioning toward the flexible multi-stream designs of modern transformer denoisers. By developing both paradigms on the same problem domain, spatially grounded gesture, the thesis makes this trade-off concrete rather than abstract, and shows that neither paradigm subsumes the other.

A unified data infrastructure makes the production–comprehension connection empirically tractable. The pointing dataset introduced alongside Paper B, the augmented pointing corpus (Paper C), MM-Conv (Paper F), and Look and Tell (Paper G) form a designed sequence in which each resource addresses a specific

gap left by the previous one. MM-Conv in particular plays a dual role in the thesis, serving both the generation and comprehension threads; the argumentative weight of this connection is developed in the fourth contribution below.

These three findings rest on a fourth contribution that is primarily methodological: the evaluation paradigms developed across the thesis consistently prioritize diagnosing *where* communicative intent breaks down over ranking systems by aggregate performance. The VR-based referential game (Paper B), the spatio-temporal grounding protocol (Paper E), expression-type stratification and rewrite-based decoupling (Paper F) each tie an evaluation number to a specific aspect of communicative success. Chapter 7 discusses these in detail; the point here is that this diagnostic orientation is itself a contribution, one that connects the modeling and data work to the broader argument by making failure modes visible rather than averaging over them.

This infrastructure has been operationalized into a series of community-facing benchmarks. Paper F formalizes context-aware grounding in 3D dialogue as a stratified evaluation over expression types. Paper E packages MM-Conv and SGS-HSI behind a spatio-temporal grounding protocol, temporal alignment, spatial accuracy, referent recall, and frames the result as a challenge at the 1st Workshop on Human–Scene Interaction at ECCV 2026. Look and Tell (Paper G) provides a paired ego/exo resource for evaluating how spatial representation affects grounding. Each commits to a task definition, a split, and a metric that the community can target without re-deriving the underlying annotations, turning the thesis's internal infrastructure into a resource for external use.

8.2 Limitations

Several limitations remain. First, much of the work is conducted in virtual or controlled environments rather than through deployment on physical robots. This makes it possible to study communication under precise conditions, but it leaves open questions about robustness under real-world sensing noise, embodiment constraints, and interaction variability. Second, the thesis focuses strongly on pointing and related referential gestures. While this is a principled choice, other spatially meaningful behaviors such as head orientation, torso movement, or more subtle deictic cues remain less explored. Third, the environments studied in the datasets and evaluations are still relatively constrained. Kitchen-like scenes and task-oriented interactions provide a good starting point, but real-world communication often involves more complex layouts, richer social dynamics, and longer interaction histories.

A further limitation concerns the scope of the production–comprehension connection itself. The thesis argues that generation and grounding are two sides of the same communicative process and shows that they share data, architectural patterns, and a common semantic–spatial split. It does not, however, build a

model that closes the loop directly: production and comprehension are studied within a shared infrastructure but remain architecturally separate. Whether a single model can jointly generate and interpret referential behavior, exploiting the symmetry the thesis identifies, is an open question that the present work motivates rather than resolves.

The absence of a unified formal model of grounding is intentional rather than incidental. Angular accuracy, IoU-based match rate, and the rewrite-based decoupling experiment constitute the operationalization of grounding in this thesis: each makes a specific aspect of the phenomenon measurable without presupposing a joint model that the empirical scope of the work does not support. A unified metric that captures communicative naturalness, semantic appropriateness, and spatial correctness together would substantially benefit the field, but developing one would require a theoretical commitment that the thesis deliberately holds open.

The contributions developed here also raise ethical considerations that warrant explicit acknowledgement. The VR and motion capture recordings were conducted with informed consent under institutional guidelines, but broader questions remain. Agents that produce convincing communicative behavior, including spatially accurate pointing and natural gesture timing, could be used to manipulate attention or fabricate nonverbal agreement. The same grounding capabilities that allow an agent to resolve a user's references also imply access to gaze, gesture, and dialogue history, raising privacy considerations for deployment in domestic or assistive settings. These concerns become more pressing as the methods developed here move from controlled research environments toward physical robots operating in shared human spaces.

8.3 Future Directions

Several directions follow naturally from the limitations identified above and from open questions that emerged across the thesis.

Spatial representations for production and comprehension. A recurring tension throughout this thesis is the mismatch between local and global coordinate representations. Kinematic generative models operate root-relative, which simplifies learning but makes global spatial targets difficult to impose; physics-based controllers operate in world frame, which makes spatial accuracy natural but limits conditioning flexibility. On the comprehension side, 2D vision-language models discard metric depth entirely, while 3D models recover spatial structure but lack dialogue context. What is missing is a systematic study of which spatial representations support both production and understanding within a unified framework. Scene graphs, which encode object identities, categories, and spatial relations without full geometric reconstruction, are a natural candidate: unlike manipulation, where contact surfaces and object geometry are critical, resolving a reference such as *the jar on the left* or generating a pointing

gesture toward it requires primarily relational and categorical spatial information. MM-Conv already provides ground-truth scene graphs, and the concurrent PoseRefer work [9] operates directly on this representation, but whether the same scene-graph abstraction suffices for both generation and grounding, and how it interacts with egocentric 3D representations of the kind explored in SpatialVLA and the Look and Tell dataset, has not been evaluated systematically. This connects directly to the vision-language-action setting, where the question of what spatial representation to condition on is equally open, and where the semantic–spatial separation identified in this thesis may inform architectural choices for embodied agents that must both act on and communicate about shared environments.

Unifying production and comprehension. This thesis treats gesture generation and referential grounding as separate but fundamentally coupled problems. In human interaction they are not separate: a speaker designs a referring expression against what they assume is recoverable from context, and a listener exploits exactly those cues. A natural next step is to model this coupling directly, for instance through shared latent representations of scene state and discourse context that jointly parameterise production and interpretation. The MM-Conv dataset, which supports both the grounding benchmarks in Paper F and the generation models in [8], provides a foundation for this, but a model that closes the loop has not yet been built. One concrete instance of this gap is communicative planning: the generation work in this thesis assumes the referent and its spatial target are given, but a fully situated agent must determine *what* to communicate based on its grounded understanding of the scene and the ongoing discourse. This gap mirrors the limitation of purely spatial conditioning identified in Section 4.4: the model learns *where* to point but not *what* to indicate. Closing it requires not a perception module upstream of generation, but a planning process in which language and gesture are jointly designed against the speaker's model of common ground. It is the production-side counterpart of the contextual reasoning that Chapter 5 identified as the bottleneck on the comprehension side.

From virtual agents to physical humanoids. The work in this thesis is conducted in virtual and controlled environments, which makes it possible to study communicative behavior under precise, reproducible conditions. As capable humanoid platforms become widely available, however, the question of how robots should communicate with people in shared environments becomes pressing in a way it was not when robots were confined to structured industrial settings. Recent work such as Kimodo [119] demonstrates that the same combination of text conditioning and kinematic constraints used for virtual character animation can generate demonstration data directly on humanoid robot skeletons, with generated motions feeding into physics-based tracking policies, closing the loop from kinematic generation to physical deployment. The methods

developed here are architecturally compatible with this setting: a humanoid assisting with a domestic task faces precisely the dual challenge described in Chapter 2, requiring gestures that are spatially accurate and motorically natural, and the ability to resolve the deictic references a human collaborator produces. The pointing gesture work in particular, which already operates in world-frame coordinates within a physics-based simulator, is well positioned for this transfer. Noisy perception, imprecise actuation, and real-time demands introduce additional challenges, but the underlying communicative structure remains the same.

Inference cost and automatic evaluation. The two-stage grounding pipeline (Paper F) and the CSMP-conditioned generation pipeline (Paper A) both incur latency costs that exceed the time scale of natural turn-taking. For interactive deployment, reducing these costs, whether through distillation, lightweight rewriters, or fewer denoising steps, is a practical priority. The modular, adapter-based design adopted throughout this thesis partially addresses this by limiting retraining to lightweight components. On the evaluation side, the Axis 2 learned ranker in Paper E (73K parameters) is a candidate for automatic evaluation of generated referential gesture, though its adequacy relative to human-judgement variance and rank-agreement with the VR-based protocol remain open empirical questions.

Richer interaction contexts and evaluation. The datasets and evaluations in this thesis focus on relatively constrained settings. Real-world situated communication involves longer interaction histories, more complex spatial layouts, and richer social dynamics than the controlled environments studied here. Extending the data collection and benchmarking efforts toward more naturalistic multi-turn interactions in unstructured scenes is an important direction. Alongside this, evaluation remains an open problem: no single metric currently captures naturalness, semantic appropriateness, and spatial correctness together in a way that retains diagnostic value, and developing such a metric, or a principled protocol for aggregating the existing axes, would substantially benefit the field.

Bibliography

[1] Anna Deichler, Shivam Mehta, Simon Alexanderson, et al. “Diffusion-Based Co-Speech Gesture Generation Using Joint Text and Audio Representation”. In: *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI)*. 2023, pp. 755–762. DOI: 10.1145/3577190.3616117.

[2] Anna Deichler, Siyang Wang, Simon Alexanderson, et al. “Learning to Generate Pointing Gestures in Situated Embodied Conversational Agents”. In: *Frontiers in Robotics and AI* 10 (2023), p. 1110534. DOI: 10.3389/frobt.2023.1110534.

[3] Anna Deichler, Simon Alexanderson, and Jonas Beskow. “Incorporating Spatial Awareness in Data-Driven Gesture Generation for Virtual Agents”. In: *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (IVA)*. 2024, pp. 1–4. DOI: 10.1145/3652988.3673923.

[4] Anna Deichler, Jim O’Regan, Téo Guichoux, et al. “Grounded Gesture Generation: Language, Motion, and Space”. In: *Humanoid Agents Workshop, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025.

[5] Anna Deichler, Rishabh Dabral, Fethiye Irmak Doğan, et al. “A Benchmark for Scene-Aware Referential Gesture Generation”. In: *Workshop on Human–Scene Interaction at the European Conference on Computer Vision (ECCV)*. To appear. Malmö, Sweden, 2026.

[6] Anna Deichler, Jim O’Regan, Fethiye Irmak Doğan, et al. “MM-Conv: A Multimodal Dataset and Benchmark for Context-Aware Grounding in 3D Dialogue”. In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*. 2026.

[7] Anna Deichler and Jonas Beskow. “Look and Tell: A Dataset for Multimodal Grounding Across Egocentric and Exocentric Views”. In: *Workshop on SPACE in Vision, Language, and Embodied AI (SpaVLE), Neural Information Processing Systems (NeurIPS)*. 2025.

[8] Anna Deichler, Rishabh Dabral, Fethiye Irmak Doğan, et al. “Scene-Aware Referential Gesture Generation via Spatial Flow Matching with Directional Refinement”. In: *Under review*. 2026.

[9] Anna Deichler. “PoseRefer: Pathway-Local Parameters for Semantically Grounded Reference Resolution”. In: *Workshop on Semantics for Reliable Robot Autonomy (SRRA) at IEEE International Conference on Robotics and Automation (ICRA)*. Non-archival. 2026.

[10] Shivam Mehta, Anna Deichler, Jim O’Regan, et al. “Fake it to make it: Using synthetic data to remedy the data shortage in joint multimodal speech-and-

gesture synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 1952–1964.

[11] Axel Wiebe Werner, Jonas Beskow, and Anna Deichler. "Gesture evaluation in virtual reality". In: *Companion Proceedings of the 26th International Conference on Multimodal Interaction*. 2024, pp. 156–164.

[12] Ilaria Torre, Anna Deichler, Matthew Nicholson, et al. "To smile or not to smile: The effect of mismatched emotional expressions in a Human-Robot cooperative task". In: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2022, pp. 8–13.

[13] Anna Deichler, Siyang Wang, Simon Alexanderson, et al. "Towards Context-Aware Human-Like Pointing Gestures with RL Motion Imitation". In: *Context-Awareness in Human-Robot Interaction: Approaches and Challenges, Workshop at ACM/IEEE International Conference on Human-Robot Interaction*. 2022.

[14] Anna Deichler, Kiran Chhatre, Christopher Peters, et al. "Spatio-Temporal Priors in 3D Human Motion". In: *Workshop on Spatio-temporal Aspects of Embodied Predictive Processing (StEPP) at IEEE ICDL*. Workshop presentation. Aug. 2021.

[15] Patrik Jonell, Anna Deichler, Ilaria Torre, et al. "Mechanical Chameleons: Evaluating the Effects of a Social Robot's Non-Verbal Behavior on Social Influence". In: *Proceedings of SCRITA, Workshop at IEEE RO-MAN*. 2021.

[16] Herbert H. Clark and Susan Brennan. "Grounding in communication". In: *Perspectives on socially shared cognition*. 1991.

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, et al. "ReferItGame: Referring to Objects in Photographs of Natural Scenes". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 787–798.

[18] Stevan Harnad. "The symbol grounding problem". In: *Physica D: Nonlinear Phenomena* 42.1-3 (1990), pp. 335–346.

[19] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[20] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

[21] Justine Cassell, Joseph Sullivan, Scott Prevost, et al., eds. *Embodied Conversational Agents*. Cambridge, MA: MIT Press, 2000.

[22] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. "Beat: the behavior expression animation toolkit". In: *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 2001, pp. 477–486.

[23] Karl Bühler. *Sprachtheorie. Die Darstellungsfunktion der Sprache*. Jena: Fischer, 1934.

[24] Leonard Talmy. "Targeting in Language: Unifying Deixis and Anaphora". In: *Frontiers in Psychology* Volume 11 - 2020 (2020). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2020.02016. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2020.02016.

[25] David Peeters, Emiel Krahmer, and Alfons Maes. "A conceptual framework for the study of demonstrative reference". In: *Psychonomic Bulletin & Review* 28 (2021), pp. 409–433.

[26] Sotaro Kita. *Pointing: Where language, culture, and cognition meet*. Psychology Press, 2003.

[27] Luigia Camaioni, Paola Perucchini, Francesca Bellagamba, et al. "The Role of Declarative Pointing in Developing a Theory of Mind". In: *Infancy* 5.3 (2004), pp. 291–308.

[28] George Butterworth. "Pointing is the royal road to language for babies". In: *Pointing*. Psychology Press, 2003, pp. 17–42.

[29] Adrian Bangerter. "Using Pointing and Describing to Achieve Joint Focus of Attention in Dialogue". In: *Psychological Science* 15.6 (2004), pp. 415–419.

[30] Yi Zhou, Connelly Barnes, Jingwan Lu, et al. "On the continuity of rotation representations in neural networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 5745–5753.

[31] Chuan Guo, Shihao Zou, Xinxin Zuo, et al. "Generating diverse and natural 3d human motions from text". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 5152–5161.

[32] Guy Tevet, Sigal Raab, Brian Gordon, et al. "Human Motion Diffusion Model". In: *International Conference on Learning Representations (ICLR)*. 2023. URL: https://openreview.net/forum?id=SJ1kSy02jwu.

[33] Fangran Miao, Jian Huang, and Ting Li. *Riemannian Motion Generation: A Unified Framework for Human Motion Representation and Generation via Riemannian Flow Matching*. 2026. arXiv: 2603.15016 [cs.CV]. URL: https://arxiv.org/abs/2603.15016.

[34] Matthew Loper, Naureen Mahmood, Javier Romero, et al. "SMPL: A skinned multi-person linear model". In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 2023, pp. 851–866.

[35] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, et al. "Expressive body capture: 3d hands, face, and body from a single image". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 10975–10985.

[36] Aaron Ferguson, Ahmed AA Osman, Berta Bescos, et al. "Mhr: Momentum human rig". In: *arXiv preprint arXiv:2511.15586* (2025).

[37] Jun Saito, Jiefeng Li, Michael de Ruyter, et al. "SOMA: Unifying Parametric Human Body Models". In: (2026). arXiv: 2603.16858. URL: https://arxiv.org/abs/2603.16858.

[38] Yiming Xie, Varun Jampani, Lei Zhong, et al. "OmniControl: Control Any Joint at Any Time for Human Motion Generation". In: *The Twelfth International Conference on Learning Representations*. 2024.

[39] Xue Bin Peng, Pieter Abbeel, Sergey Levine, et al. "DeepMimic: Example-Guided Deep Reinforcement Learning of Physics-Based Character Skills". In: *ACM Transactions on Graphics* 37.4 (2018), pp. 1–14.

[40] Emanuel Todorov, Tom Erez, and Yuval Tassa. "Mujoco: A physics engine for model-based control". In: *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE. 2012, pp. 5026–5033.

[41] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, et al. "Isaac Gym: High Performance GPU Based Physics Simulation For Robot Learning". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021.

[42] Chen Tessler, Yunrong Guo, Ofir Nabati, et al. “MaskedMimic: Unified physics-based character control through masked motion inpainting”. In: *ACM Transactions On Graphics (TOG)* 43.6 (2024), pp. 1–21.

[43] John Schulman, Filip Wolski, Prafulla Dhariwal, et al. “Proximal policy optimization algorithms”. In: *arXiv preprint arXiv:1707.06347* (2017).

[44] Jonathan Ho and Stefano Ermon. “Generative adversarial imitation learning”. In: *Advances in neural information processing systems* 29 (2016).

[45] Xue Bin Peng, Ze Ma, Pieter Abbeel, et al. “AMP: Adversarial Motion Priors for Stylized Physics-Based Character Control”. In: *ACM Transactions on Graphics* 40.4 (2021), pp. 1–20.

[46] Mathis Petrovich, Michael J Black, and Gül Varol. “Temos: Generating diverse human motions from textual descriptions”. In: *European conference on computer vision*. Springer. 2022, pp. 480–497.

[47] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, et al. “Flow Matching for Generative Modeling”. In: *The Eleventh International Conference on Learning Representations*. 2023.

[48] Vincent Tao Hu, Wenzhe Yin, Pingchuan Ma, et al. “Motion Flow Matching for Human Motion Synthesis and Editing”. In: *arXiv preprint arXiv:2312.08895* (2023).

[49] Guy Tevet, Sigal Raab, Setareh Cohan, et al. “CLoSD: Closing the Loop between Simulation and Diffusion for multi-task character control”. In: *The Thirteenth International Conference on Learning Representations*. 2025.

[50] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising Diffusion Probabilistic Models”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 6840–6851.

[51] Zhifeng Kong, Wei Ping, Jiaji Huang, et al. “DiffWave: A Versatile Diffusion Model for Audio Synthesis”. In: *International Conference on Learning Representations (ICLR)*. 2021. URL: https://openreview.net/forum?id=a-xFK8Ymz5J.

[52] Aäron van den Oord, Sander Dieleman, Heiga Zen, et al. “WaveNet: A Generative Model for Raw Audio”. In: *Proc. SSW 2016*. 2016, pp. 125–125.

[53] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, et al. “Listen, denoise, action! audio-driven motion synthesis with diffusion models”. In: *ACM Transactions on Graphics (TOG)* 42.4 (2023), pp. 1–20.

[54] Ulme Wennberg and Gustav Eje Henter. “The Case for Translation-Invariant Self-Attention in Transformer-Based Language Models”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, 2021, pp. 130–140. URL: https://aclanthology.org/2021.acl-short.18/.

[55] William Peebles and Saining Xie. “Scalable Diffusion Models with Transformers”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE. 2023, pp. 4172–4182.

[56] Jonathan Ho and Tim Salimans. “Classifier-Free Diffusion Guidance”. In: *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

[57] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. “Adding conditional control to text-to-image diffusion models”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 3836–3847.

[58] Lei Zhong, Yiming Xie, Varun Jampani, et al. “SMooDi: Stylized Motion Diffusion Model”. In: *ECCV*. 2024.

[59] Xin Chen, Biao Jiang, Wen Liu, et al. “Executing your commands via motion diffusion in latent space”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 18000–18010.

[60] Haim Sawdayee, Chuan Guo, Guy Tevet, et al. “Dance Like a Chicken: Low-rank stylization for human motion diffusion”. In: *arXiv preprint arXiv:2503.19557* (2025).

[61] Jungbin Cho, Minsu Kim, Jisoo Kim, et al. “SceneAdapt: Scene-aware Adaptation of Human Motion Diffusion”. In: *arXiv preprint arXiv:2510.13044* (2025).

[62] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. “Moglow: Probabilistic and controllable motion synthesis using normalising flows”. In: *ACM Transactions on Graphics (TOG)* 39.6 (2020), pp. 1–14.

[63] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, et al. “A Comprehensive Review of Data-Driven Co-Speech Gesture Generation”. In: *Computer Graphics Forum* 42.2 (2023), pp. 569–595.

[64] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, et al. “Hubert: Self-supervised speech representation learning by masked prediction of hidden units”. In: *IEEE/ACM transactions on audio, speech, and language processing* 29 (2021), pp. 3451–3460.

[65] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, et al. “Wavlm: Large-scale self-supervised pre-training for full stack speech processing”. In: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022), pp. 1505–1518.

[66] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, et al. “Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows”. In: *Computer Graphics Forum*. Vol. 39. 2. Wiley Online Library. 2020, pp. 487–496.

[67] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, et al. “Gesticulator: A Framework for Semantically-Aware Speech-Driven Gesture Generation”. In: *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*. 2020, pp. 242–250.

[68] Ethan Perez, Florian Strub, Harm De Vries, et al. “Film: Visual reasoning with a general conditioning layer”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[69] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. PMLR. 2021, pp. 8748–8763.

[70] Guy Tevet, Brian Gordon, Amir Hertz, et al. “Motionclip: Exposing human motion generation to clip space”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 358–374.

[71] Tenglong Ao, Zeyi Zhang, and Libin Liu. “GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents”. In: *ACM Transactions on Graphics (TOG)* 42.4 (2023), pp. 1–18. DOI: 10.1145/3592097.

[72] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, et al. “Data2vec: A general framework for self-supervised learning in speech, vision and language”. In: *International conference on machine learning*. PMLR. 2022, pp. 1298–1312.

[73] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. “Representation learning with contrastive predictive coding”. In: *arXiv preprint arXiv:1807.03748* (2018).

[74] Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, et al. “The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings”. In: *Proceedings of the 25th international conference on multimodal interaction*. 2023, pp. 792–801.

[75] Daniel Holden, Taku Komura, and Jun Saito. “Phase-functioned neural networks for character control”. In: *ACM Transactions on Graphics (TOG)* 36.4 (2017), pp. 1–13.

[76] Jungbin Cho, Junwan Kim, Jisoo Kim, et al. “DisCoRD: Discrete Tokens to Continuous Motion via Rectified Flow Decoding”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025, pp. 14602–14612.

[77] Cheng Chi, Siyuan Feng, Yilun Du, et al. “Diffusion Policy: Visuomotor Policy Learning via Action Diffusion”. In: *Proceedings of Robotics: Science and Systems*. Daegu, Republic of Korea, July 2023. DOI: 10.15607/RSS.2023.XIX.026.

[78] Kevin Black, Noah Brown, Danny Driess, et al. “π_0: A Vision-Language-Action Flow Model for General Robot Control”. In: *Proceedings of Robotics: Science and Systems (RSS)*. 2025.

[79] Delin Qu, Haoming Song, Qizhi Chen, et al. “SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Models”. In: *Robotics: Science and Systems (RSS)*. 2025.

[80] Quantao Yang, Michael C Welle, Danica Kragic, et al. “S ^{2}-Diffusion: Generalizing from Instance-level to Category-level Skills in Robot Manipulation”. In: *IEEE Robotics and Automation Letters* (2025).

[81] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, et al. “Grounding dino: Marrying dino with grounded pre-training for open-set object detection”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 38–55.

[82] Zhiliang Peng, Wenhui Wang, Li Dong, et al. “Kosmos-2: Grounding Multimodal Large Language Models to the World”. In: *ArXiv* abs/2306 (2023).

[83] Bin Xiao, Haiping Wu, Weijian Xu, et al. “Florence-2: Advancing a unified representation for a variety of vision tasks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 4818–4829.

[84] Shuai Bai, Keqin Chen, Xuejing Liu, et al. “Qwen2.5-VL Technical Report”. In: *arXiv preprint arXiv:2502.13923* (2025).

[85] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, et al. “Scanqa: 3d question answering for spatial scene understanding”. In: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 19129–19139.

[86] Yining Hong, Haoyu Zhen, Peihao Chen, et al. “3d-llm: Injecting the 3d world into large language models”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 20482–20494.

[87] Sijin Chen, Xin Chen, Chi Zhang, et al. “Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 26428–26438.

[88] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. “ScanRefer: 3D object localization in RGB-D scans using natural language”. In: *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*. Glasgow, United Kingdom: Springer-Verlag, 2020, pp. 202–221. ISBN: 978-3-030-58564-8. DOI: 10.1007/978-3-030-58565-5_13. URL: https://doi.org/10.1007/978-3-030-58565-5_13.

[89] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, et al. “ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer. 2020, pp. 422–440.

[90] Wenhui Wang, Furu Wei, Li Dong, et al. “Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers”. In: *Advances in neural information processing systems* 33 (2020), pp. 5776–5788.

[91] Ylva Ferstl, Michael Neff, and Rachel McDonnell. “ExpressGesture: Expressive gesture generation from speech through database matching”. In: *Computer Animation and Virtual Worlds* (2021), e2016.

[92] Haiyang Liu, Zihao Zhu, Giorgio Becherini, et al. “EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Expressive Masked Audio Gesture Modeling”. In: *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. Piscataway, NJ, Sept. 2024, pp. 1144–1154. DOI: 10.1109/CVPR52733.2024.00115. URL: https://pantomatrix.github.io/EMAGE/.

[93] Gilwoo Lee, Zhiwei Deng, Shugao Ma, et al. “Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 763–772.

[94] Claire McLean, Makenzie Meendering, Tristan Swartz, et al. “Embody 3d: A large-scale multimodal motion and behavior dataset”. In: *arXiv preprint arXiv:2510.16258* (2025).

[95] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, et al. “Resolving 3D Human Pose Ambiguities with 3D Scene Constraints”. In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 2282–2292.

[96] Omid Taheri, Nima Ghorbani, Michael J. Black, et al. “GRAB: A Dataset of Whole-Body Human Grasping of Objects”. In: *European Conference on Computer Vision (ECCV)*. 2020.

[97] Zan Wang, Yixin Chen, Tengyu Liu, et al. “HUMANISE: Language-conditioned Human Motion Generation in 3D Scenes”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.

[98] Yixin Chen, Qing Li, Deqian Kong, et al. “Yourefit: Embodied reference understanding with language and gesture”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1385–1395.

[99] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, et al. “TEACh: Task-Driven Embodied Agents that Chat”. In: *Proceedings of the AAAI Con-*

ference on Artificial Intelligence. Vol. 36. 2. 2022, pp. 2017–2025. DOI: 10.1609/aaai.v36i2.20097.

[100] Ziyang Lu, Yunqiang Pei, Guoqing Wang, et al. "ScanERU: Interactive 3D visual grounding based on embodied reference understanding". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 2024, pp. 3936–3944.

[101] Ylva Ferstl and Rachel McDonnell. "Investigating the use of recurrent motion modelling for speech gesture generation". In: *IVA '18 Proceedings of the 18th International Conference on Intelligent Virtual Agents*. Nov. 2018. URL: https://trinityspeechgesture.scss.tcd.ie.

[102] Joao Pedro Araujo, Jiaman Li, Karthik Vetrivel, et al. "CIRCLE: Capture In Rich Contextual Environments". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2023, pp. 21211–21221. DOI: 10.1109/CVPR52729.2023.02032. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.02032.

[103] Zicong Fan, Omid Taheri, Dimitrios Tzionas, et al. "ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 12943–12954.

[104] Xitong Yang, Devansh Kukreja, Don Pinkus, et al. "Sam 3d body: Robust full-body human mesh recovery". In: *arXiv preprint arXiv:2602.15989* (2026).

[105] Vickie Ye, Georgios Pavlakos, Jitendra Malik, et al. "Decoupling Human and Camera Motion from Videos in the Wild". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023.

[106] Zhizheng Liu, Joe Lin, Wayne Wu, et al. "Joint Optimization for 4D Human-Scene Reconstruction in the Wild". In: *The Fourteenth International Conference on Learning Representations* (2026).

[107] Lea Müller, Hongsuk Choi, Anthony Zhang, et al. "Reconstructing people, places, and cameras". In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 21948–21958.

[108] Yuhong Zhang, Jing Lin, Ailing Zeng, et al. "Motion-x++: A large-scale multimodal 3d whole-body human motion dataset". In: *arXiv preprint arXiv:2501.05098* (2025).

[109] Andy Lücking, Thies Pfeiffer, and Hannes Rieser. "Pointing and reference reconsidered". In: *Journal of Pragmatics* 77 (2015), pp. 56–79.

[110] Josh Achiam, Steven Adler, Sandhini Agarwal, et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

[111] Herbert H. Clark and Deanna Wilkes-Gibbs. "Referring as a collaborative process". In: *Cognition* 22.1 (1986), pp. 1–39. DOI: 10.1016/0010-0277(86)90010-7.

[112] Eric Kolve, Roozbeh Mottaghi, Winson Han, et al. "AI2-THOR: An Interactive 3D Environment for Visual AI". In: *arXiv* (2017).

[113] Max Bain, Jaesung Huh, Tengda Han, et al. "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio". In: *INTERSPEECH 2023* (2023).

[114] Robert Dale and Ehud Reiter. "Computational interpretations of the Gricean maxims in the generation of referring expressions". In: *Cognitive Science* 19.2 (1995), pp. 233–263.

[115] Matt Deitke, Christopher Clark, Sangho Lee, et al. “Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models”. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 91–104.

[116] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, et al. “SAM 2: Segment Anything in Images and Videos”. In: *The Thirteenth International Conference on Learning Representations*. 2025.

[117] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, et al. “From coarse to fine: Robust hierarchical localization at large scale”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 12716–12725.

[118] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, et al. “The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation”. In: *Proceedings of the 2022 international conference on multimodal interaction*. 2022, pp. 736–747.

[119] Davis Rempe, Mathis Petrovich, Ye Yuan, et al. “Kimodo: Scaling Controllable Human Motion Generation”. In: *arXiv preprint arXiv:2603.15546* (2026).

Included Publications

Paper A

Paper A

Diffusion-Based Co-Speech Gesture Generation Using Joint Text and Audio Representation

ANNA DEICHLER, SHIVAM MEHTA, SIMON ALEXANDERSON, AND JONAS BESKOW

Published in
Proceedings of the 25th ACM International Conference on Multimodal Interaction (ICMI '23)

arXiv:2309.05455v1 [eess.AS] 11 Sep 2023

Diffusion-Based Co-Speech Gesture Generation Using Joint Text and Audio Representation

Anna Deichler
deichler@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Shivam Mehta
smehta@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Simon Alexanderson
simonal@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Jonas Beskow
beskow@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

ABSTRACT

This paper describes a system developed for the GENEA (Generation and Evaluation of Non-verbal Behaviour for Embodied Agents) Challenge 2023. Our solution builds on an existing diffusion-based motion synthesis model. We propose a contrastive speech and motion pretraining (CSMP) module, which learns a joint embedding for speech and gesture with the aim to learn a semantic coupling between these modalities. The output of the CSMP module is used as a conditioning signal in the diffusion-based gesture synthesis model in order to achieve semantically-aware co-speech gesture generation. Our entry achieved highest human-likeness and highest speech appropriateness rating among the submitted entries. This indicates that our system is a promising approach to achieve human-like co-speech gestures in agents that carry semantic meaning.

KEYWORDS

gesture generation, motion synthesis, diffusion models, contrastive pre-training, semantic gestures

ACM Reference Format:
Anna Deichler, Shivam Mehta, Simon Alexanderson, and Jonas Beskow. 2023. Diffusion-Based Co-Speech Gesture Generation Using Joint Text and Audio Representation. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 9–13, 2023, Paris, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3577190.3616117

1 INTRODUCTION

Human communication is inherently multimodal involving the integration of multiple verbal and non-verbal modalities to convey the information. These modalities work in synergy, collaborating to create a joint representation of the message the speaker intends to convey [29]. In addition to complementing verbal communication, these non-verbal gestures frequently serve as substitutes for words [9, 31]. The semantic meaning contribution of gestures is multifaceted. Beat gestures primarily emphasize the verbally expressed content, serving to accentuate the spoken message. On the other hand, iconic and pointing gestures go beyond emphasizing content; they directly represent or indicate the referent being discussed. Deictic pointing gestures, often accompanying deictic words, play a crucial role in referential communication by providing vital contextual information for reference disambiguation, while iconic gestures serve to visually represent or symbolize the attributes, actions, or characteristics associated with the referent.

Co-speech gesture generation in robotics and avatars focuses on generating gestures that accompany and extend the verbal modality. However, the generation of audio-driven motion has posed a significant challenge. This difficulty arises from the fact that such motion can be accurately predicted by very strong probabilistic models, since gestures exhibit high individual variability, are inherently non-deterministic [2]. Recent advances in learning arbitrary probability distributions with diffusion models has offered a way to tackle this problem. These audio-driven gesture generation models have proven to be efficient in reproducing the high variability and expressivity of human gestures, however integrating semantic content into gesture generation by combining audio and text conditioning is another challenge.

Self-supervised pre-training methods have proven to be an efficient way to learn useful representations for downstream tasks, especially in case of limited labeled data. Multi-modal pre-training methods learn embedding spaces that encode useful relations of different data modalities. Contrastive Language-Image Pre-Training (CLIP) [32] is a contrastive multi-modal pre-training method that learns a joint representation of image and text data by contrasting positive and negative text-image pair examples in the latent space during training. This training approach encourage the model to capture the underlying relationship between the two modalities.

The problem of co-speech gesture generation involves multiple modalities, with a tight coupling between motion, text and audio. This work aims at combining the expressivity of diffusion based motion synthesis [2] with the multi-modal understanding of a CLIP-like latent embedding space that models the relations between motion, text and audio in co-speech gestures.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
ICMI '23, October 9–13, 2023, Paris, France
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0055-2/23/10.
https://doi.org/10.1145/3577190.3616117

2 RELATED WORK

2.1 Co-speech gesture generation

The primary goal of co-speech gesture generation is to synthesise natural and contextually appropriate gestures. In the early stages of gesture generation research, various rule-based approaches were employed [5, 26, 27], where the generation of gestures was triggered by predefined rules that initiated the playback of pre-recorded gestures. In recent years, this field has been dominated by the use of data-driven deep learning based modelling methodologies [31].

Early works on deep learning-based gesture synthesis treated it as a regression problem and utilised recurrent [14, 36] and convolutional [21] neural networks to model the generation process. Treating gesture synthesis as a regression problem leads to the problem of under-articulated and over-smoothened gestures because of averaging over all the possible outcomes for an input signal. To address the challenge of under-articulated and over-smoothened synthesis researchers employed various probabilistic modelling techniques such as VAEs [12], VQ-VAEs [43], Normalising Flows [1] or adversarial techniques like GANs [41, 42]. These methodologies aim to enhance the realism and expressiveness of the generated gestures by learning a distribution over the entire utterances and sampling different realisations from it or learning powerful transformations from a simple distribution, usually a Gaussian distribution, to the output motion distribution.

Diffusion models [15, 34, 35] have emerged as a notable and contemporary probabilistic generative modelling methodology. These models have shown promise in capturing complex data distributions and have gained attention in various fields, including gesture generation [2, 3, 30, 45]. Inspired by these works our system uses Denoising Diffusion Probabilistic Modelling (DDPM) [15] formulation with self-supervised representations to synthesise gestures conditioned on the input audio.

2.2 Semantic gesture generation

In order to generate contextually appropriate gestures in agents, it is crucial to take into account gesture semantics. Semantic gestures have a symbolic representational quality and contribute to the overall meaning in communication. The generation of semantic gestures is highly reliant on which input modalities are taken into account in the modeling process [31].

Audio driven generation can reproduce the coupling between gesture kinematics and the intonation, stress and rhythm present in the audio signal. These systems are good at modeling beat gestures, which can help highlight important points or add emphasis to certain words or phrases [28],[1],[2]. However, in order to generate representational gestures (e.g., iconic, deictic pointing), additional input modalities are needed. Text-based conditioning is essential to model the relation between semantic and kinematic spaces in order to generate iconic gestures [44],[22], while the generation of deictic pointing gestures needs referential target information [10]. In this work we develop a novel approach to jointly model audio and text conditioning in gesture generation through a contrastive self-supervised learning approach in order to extend the existing audio conditioned system with semantic capabilities.

2.3 Using language based pre-training approaches in motion generation

Recent works approaches have leveraged different pre-training approaches to learn the semantic coupling between text and motion spaces. [46] uses a GPT-like module to generate code indices based on text embeddings which are utilized by a VQ-VAE module in motion generation, while [17] proposes MotionGPT, which performs language modeling on both motion and text in a unified manner, treating human motion as a specific language. Previous work has also leveraged CLIP's multimodal understanding to generate meaningful motion. [37] develops an auto-encoder based motion generation model, which learns a motion embedding space aligned with CLIP's latent space, which allows for the generation of expressive and versatile text-based motions. [38] uses CLIP latents as conditioning information in diffusion based human motion generation. Similarly, [8] conditions on CLIP latents, but combines latent space based and diffusion based motion generation. Most similar to our work is [3], which learns a gesture-text joint embedding using contrastive learning and a CLIP based style encoding module in a diffusion based gesture synthesis model.

3 METHOD

3.1 Self-supervised representations of text and audio

We employ pre-trained self-supervised representations for text and audio for both the main agent and the interlocutor. Data2vec [4] which is a framework for self-supervised representation learning on data of different modalities (text, audio and images), for which pre-trained models are available [1]. Data2vec leverages transformer architecture in a self-distillation setup to achieve contextual text embedding, predicting latent representations of the full input data based on a masked view of the input.

For audio, we use the `data2vec-audio-base-960h` model, which takes one-channel 16 Khz audio as input. As output we use the last hidden layer, which gives us a sequence of 768-dimensional embedding vectors at a rate of 50 Hz. The output is then converted to 30 Hz using polyphase resampling (`scipy.signal.resample_poly`) in order to match the frame rate of the motion data.

For text, we use the `data2vec-text-base` model. Input to the model is a sequence of byte-pair encoded text tokens. Just as for the audio, we use the last hidden layer of the data2vec model to obtain a 768-dimensional vector for each input token. We use the word timed transcriptions provided in the dataset (see [23]) to maintain a start and end time for each token, then we replicate the output vector at a rate of 30 Hz for the duration of the token, The result is a text-embedding sequence that is aligned with, and of the same length as, the audio and motion data sequences.

3.2 Join representation with Contrastive Speech and Motion Pretraining (CSMP)

Contrastive pre-training can effectively capture the semantic relationships between two modalities but usually, it requires a larger batch size and larger dataset to learn efficient joint representations [7] which can be challenging especially in this case because of

[1] huggingface.co

dataset-specific properties [23] such as the presence of an interlocutor and the skeletal nodes of the characters. In such a case having representations which already capture semantic information can be used as the inputs to the CLIP module. Therefore, we devise a variation of CLIP and call it Contrastive Speech and Motion Pretraining (CSMP).

In CSMP, we propose several modifications to the original CLIP architecture within the context of multimodal understanding namely:

(1) We replace the vision transformer present in the original CLIP architecture with a regular transformer architecture, which effectively eliminates the patching process typically employed for 2-D image analysis. This modification is motivated by the nature of text and audio.
(2) The input to this modified transformer is derived from concatenated representations of the output of the pretrained data2vec module for text and audio as described in section 3.1 instead of raw tokens for original CLIP.
(3) For the text encoder in CLIP, we modify the input from discrete text tokens to continuous motion vectors thus eliminating the need for an embedding layer. This alteration is intended to mimic the semantic information contained in the text and audio representations to the motion representation in the joint space of CSMP's representations.
(4) Since the original clip takes discrete and tokenized text as an input it had a context length of 77 this, in the case of modalities like the output of data2vec and motion which is continuous in nature can be insufficient to capture longer-term dependencies. In order to overcome and increase the encoder's field of view we increased the context length to 500 timesteps.

The final architecture of the CSMP module is described in Fig. 1

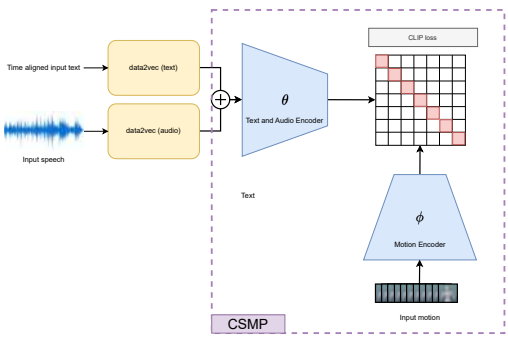


Figure 1: Architecture of Contrastive Speech Motion Pretraining (CSMP) module.

In order to train such an architecture with CLIP loss, we chunked each input $X_i = [x_1, \cdots, x_T]$ in a sliding window manner with a window length of 500 and a hop length of 250 and formed multiple splits for each utterance.

$$X_i = [[x_1, \cdots, x_{500}], [x_{250}, \cdots, x_{750}], \cdots, [x_{T-500}, \cdots, x_T]]$$

We hypothesise that this helped in the generalisation despite a fixed context size because the positional encoding could see the data at a specific timestep x_t in different relative positions while training. The source code is available on GitHub in GestCLIP branch[2].

3.3 DDPM for motion synthesis

Diffusion models are a recent class of generative models that have become popular due to their expressivity and flexible conditioning. Diffusion models are based on the idea that complex data distributions can be learned by iteratively transforming a simple known distribution, such as a Gaussian distribution, through a series of diffusion steps. Unlike VAEs, which incorporate latent variable modeling, diffusion models directly model the data distribution without explicitly introducing latent variables. Diffusion models consist of a forward process and a reverse (denoising) process. The forward process defines a Markov chain of N diffusion steps to gradually add noise to samples from the data distribution $x_0 \sim q(z)$. The noise steps are assumed to be fixed, zero-mean Gaussian distributions, without learnable parameters, $q(x_n|x_{n-1}) = \mathcal{N}(x_n; \sqrt{1-\beta_n}x_{n-1}, \beta_n\mathbf{I})$, where $\mathcal{N}$ denotes the multivariate Gaussian density function evaluated at x_n and $\{\beta_n\}_{n=1}^N$ is the noise schedule. In the reverse process the model learns to reverse the forward process so that the model is able to construct desired data samples from the noise. If β_n is small enough, the reverse step $p(x_{n-1}|x_n)$ is also Gaussian and a neural network is used to approximate the parameters of the distribution $p_\theta(x_{n-1}|x_n) = \mathcal{N}(x_{n-1}; \mu_\theta(x_n, n), \Sigma_\theta(x_n, n))$.

The Denoising Diffusion Probabilistic Model (DDPM) [15] simplifies the objective of diffusion model and establishes a connection to score matching, which is a technique used for estimating the gradients of the probability distribution of data. These gradients are then used to generate samples via Langevin dynamics, which is a stochastic process that simulates the motion of particles in a fluid. In DDPM the score-matching objective is reformulated as noise predicting objective, $\mathcal{L} = \mathbb{E}_{x_0,n,\epsilon}[\kappa_n\|\epsilon - \epsilon_\theta(x_n, n)\|_2^2]$, where ϵ_θ is a neural network intended to predict the noise ϵ that was added to x_0 and κ_n are weights.

Conditional generation in diffusion models can be achieved via classifier-guided or classifier-free models. In classifier guided diffusion models the gradients of a separately trained classifier $f_\phi(y|x_n)$ is used to guide the diffusion process $\nabla_{\mathbf{x}} f_\phi(y|x_n)$[11]. Classifier-free diffusion models combine conditional and unconditional diffusion in order to guide the diffusion. In the above formulation this means that a conditional network $\epsilon_\theta(x_n, n, c)$, with conditioning input c is trained, where the conditioning information is randomly discarded during training, so that in the reverse diffusion process conditional generation can be achieved by the combination of the input conditioned and unconditioned model $\bar{\epsilon}_\theta(x_n, n, c) = \epsilon_\theta(x_n, n, c) + \gamma(\epsilon_\theta(x_n, n, c) - \epsilon_\theta(x_n, n))$ [16]. Denoising diffusion based conditional generation has been applied in various domains. In [33], the CLIP embedding based conditioning input is randomly set to zero in order to achieve high quality image synthesis. DiffWave [20] is a denoising diffusion based model for waveform generation, which uses mel spectograms and speaker ID as conditioning information. The Listen-Denoise-Act (LDA) model

[2]https://github.com/shivammehta25/CLIP/tree/GestCLIP

[2] builds on the DiffWave model and uses mel spectogram information for human motion synthesis. Audio based conditional human motion synthesis, such as dancing and co-speech gesture generation have been a challenge in machine learning, due to the ambiguity and high versatility required for good performance in these tasks. The denoising diffusion based LDA model have proven to be a powerful model to generate versatile and expressive motion in the fields of dance and co-speech gesture generation. In our work we use the residual deonising network of LDA with a conditioning from the CSMP module for semantically-aware co-speech gesture generation.

The LDA model follows DiffWave in parameterising the denoising network ϵ_θ, but replaces the dilated convolutions in the stacked residual blocks with a stack of Transformers [39] or Conformers [13] in order to capture and integrate information over long time scales. In our experiments we use a stack of 3 translation-invariant transformers [40] in each of the 15 residual blocks. The model learns a distribution of the form $p(x_{1:T}|a_{1:T})$, where $a_{1:T}$ is the acoustic conditioning and $x_{1:T} = x_{1:T,0}$ is the output of the diffusion process and x_t is a representation of the pose at time step t in the motion sequence. In our case, the mel spectogram based acoustic conditioning of LDA is replaced with the joint audio and text based output of the CSMP module, where the outputs for interlocutor and the main agent data are concatenated into a conditioning signal of dimension $c_t \in \mathbb{R}^{1024}$. This is the conditioning input in the classifier-free diffusion guidance formulation. The outputs of the model are the same in LDA, poses of skeletal joint rotations parametrised using an exponential map representation relative to a T-pose, similarly as in [1].

4 DATA PREPARATION

The challenge dataset is a processed version of the Talking With Hands dataset[25]. The original dataset is one of the largest conversational dataset of motion and voice, incorporating 50 hours of dyadic interactions, with audio, text and motion modalities. We only used the data provided by the challenge for gesture synthesis.

4.1 Audio DC-removal and muting of cross-talk

We found that the audio data contained a number of loud transient clicking noises. On inspection, it was found that they were due to a significant DC-offset, in combination with the fact that certain sections of the audio signal had been zeroed out, as part of an anonymization process. This was easily rectified by subtracting the mean from all non-zeroed out portions.

Additionally, the data contained a non-negligible amount of cross-talk between the two speakers in the recording. We used the time stamps from the time-aligned text transcriptions to mute all audio falling outside of the intervals marked as speech in the transcription for each speaker. We used a 200 ms ramp function for the muting to avoid introducing transients.

4.2 Motion capture data cleaning

We also noticed that some of the motion capture data contained errors such as joints suddenly popping to unnatural poses. These errors were predominantly confined to the wrist joints, but also occurred at the hips. As such problems has an impact model training, and we even found our model reproducing them in synthesis, we performed some data cleanup. We transformed the data to joint positions and detected discontinuities in the wrist speeds using a Hampel filter. This was followed by a manual check of the affected files. In the end, 17 files were removed from the training set.

5 SYSTEM OVERVIEW

Schematic view of the final system can be seen in Figure 2. The system was trained on a NVIDIA GeForce RTX 3090 for 387.4k steps and achieved 0.013 loss on the training and 0.019 loss on the validation set. No post-processing was applied on the generated output motions.

6 EVALUATION

The evaluation of the generated motions was carried out by the GENEA Challenge organisers, details about the evaluation interface and experiment setups can be found in the evaluation paper [24]. The generated co-speech gesture were evaluated in three separate perceptual studies: human-likeness, appropriateness to the agent's speech and appropriateness to the interlocutor's motion and speech. The evaluation included two baseline conditions and the natural motion taken from the motion-capture recordings. The monadic baseline ('BM') was generated with [6] which uses information from the main-agent for gesture generation, while the dyadic baseline ('BD') is an adapted version of the former, which also includes information from the interlocutor in the conversation. The study participants were recruited through a crowd-sourcing platform from English-speaking countries and each study incorporated attention checks. Our system, labeled as 'SG' achieved top performance in the studies of human-likeness and speech appropriateness based on the generated motions submitted. However, it ranked among the lowest in terms of interlocutor appropriateness.

6.0.1 Human-likeness evaluation. The aim of this study was to evaluate whether the generated motion of the virtual character looks like the motion of a real human. No audio was used in order to disentangle the human-likeness evaluation from the speech appropriateness. The evaluation was based on the HEMVIP methodology [19], where multiple different motion samples are presented in parallel and the participant is asked to rate each sample. Participant could give their ratings on a scale from 0 (worst) to 100 (best). Results for the evaluation are shown on Figure 3. Our system, denoted as 'SG', achieved best performance from the entries, with mean rating of 65.6 ± 1.4. Figure 4 also shows that this results is significantly better than all of the entries, except 'SF'. Interestingly, the human-likeness score is very close to mean rating of the natural condition, which was 68.4 ± 1.4 as seen on Table 1. This indicates that our system can generate co-speech gestures which resembles the motion of real humans.

6.0.2 Appropriateness to speech. The aim of this study was to evaluate whether the motion if the virtual character is appropriate for the given speech, controlling the overall human-likeness of the motion. The participants were presented with a pair of matched and mismatched videos from the same condition in order to disentangle this study from the motion quality evaluation. Five response options were given for indicating preference over the 2 videos and

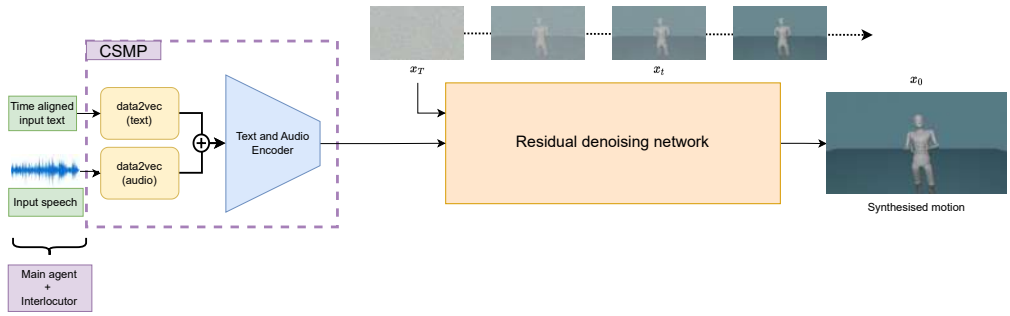


Figure 2: Architecture of the motion synthesis module

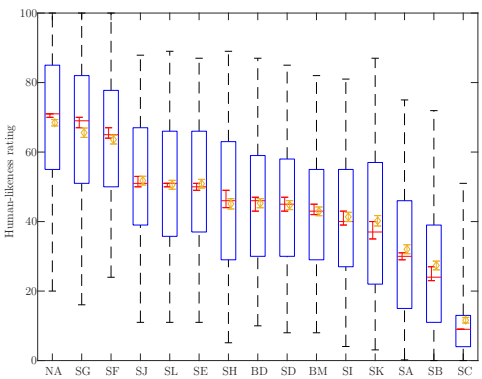


Figure 3: Box plot visualising the ratings distribution in the human-likeness study. Red bars are the median ratings (each with a 0.05 confidence interval); yellow diamonds are mean ratings (also with a 0.05 confidence interval). Box edges are at 25 and 75 percentiles, while whiskers cover 95% of all ratings for each condition. Conditions are ordered by descending sample median rating.

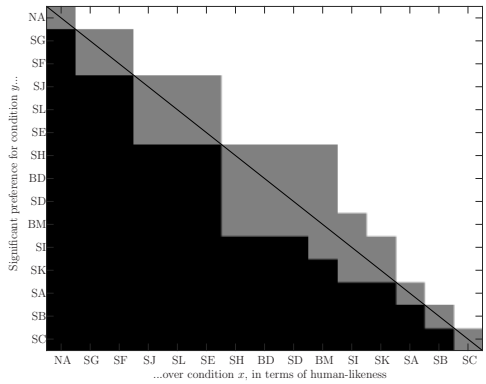


Figure 4: Significance of pairwise differences between conditions. White means the condition listed on the y-axis achieved an MAS significantly above the condition on the x-axis, black means the opposite (y scored below x), and grey means no statistically significant difference at level $\alpha = 0.05$ after correction for the false discovery rate. Conditions use the same order as the corresponding subfigure in Figure 3

the responses were converted to integer values in the range of $[-2, 2]$. Our system achieved a MAS score of 0.39 ± 0.07 at the level of $\alpha = 0.05$ and the matched motion was preferred over the mismatched in 61.8% of the evaluations. With these results it ranked highest amongst the generated motions. Figure 5 visualizes the significant differences between conditions and shows that our system, denoted by 'SG', was significantly more appropriate to speech than all of the entries of generated motions. Comparison to other entries can be found in Table 1.

6.0.3 Appropriateness to interlocutor. The aim of this study was to evaluate whether the motion of the virtual character is appropriate for the given interlocutor behavior (speech and motion). In order to evaluate the mismatched condition, synthetic interactions were created, where the main agent was the same, but the interlocutor behavior was replaced with one from another interaction. Our system achieved a MAS score of -0.09 ± 0.08 at the level of $\alpha = 0.05$ and the matched motion was preferred over the mismatched in 46.7% of the evaluations. With these results it ranked among the lowest. Figure 6 visualizes the significant differences between conditions and shows that our system, denoted by 'SG', was significantly less appropriate to interlocutor than all half of the entries of generated motions and there was no significant difference to the other half. Comparison to other entries can be found in Table 1.

The MP4-format video stimuli used in the user studies can be accessed through the following link: https://zenodo.org/record/8211449. As before, our system is denoted as 'SG'.

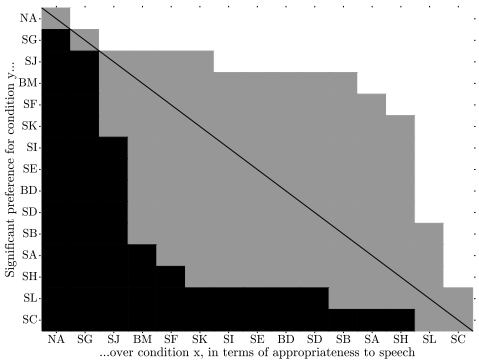


Figure 5: Appropriateness for agent speech

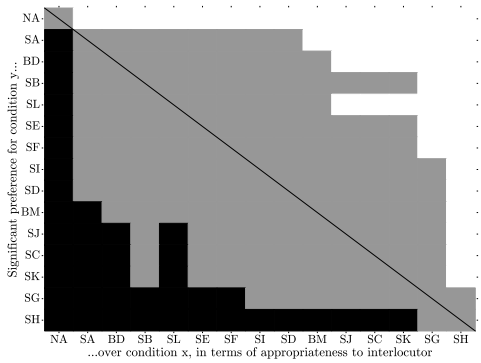


Figure 6: Appropriateness for the interlocutor

Figure 7: Significant differences between conditions in the two appropriateness studies. White means the condition listed on the y-axis achieved an MAS significantly above the condition on the x-axis, black means the opposite (y scored below x), and grey means no statistically significant difference at level $\alpha = 0.05$ after correction for the false discovery rate.

Table 1: Summary of results for subjective evaluation studies with confidence intervals for the mean appropriateness score (MAS) at the level $\alpha = 0.05$. "Pref. matched" identifies how often test-takers preferred matched motion in terms of appropriateness, after splitting ties equally.

	Human-likeness			Speech appropriateness			Interlocutor appropriateness	
Condition	Median	Mean	Condition	MAS	Pref.M.	Condition	MAS	Pref.M.
NA	71 ∈ [70, 71]	68.4±1.0	NA	0.81±0.06	73.6%	NA	0.63±0.08	67.9%
SG	69 ∈ [67, 70]	65.6±1.4	SG	0.39±0.07	61.8%	SA	0.09±0.06	53.5%
SF	65 ∈ [64, 67]	63.6±1.3	SJ	0.27±0.06	58.4%	BD	0.07±0.06	53.0%
SJ	51 ∈ [50, 53]	51.8±1.3	BM	0.20±0.05	56.6%	SB	0.07±0.08	51.8%
SL	51 ∈ [50, 51]	50.6±1.3	SF	0.20±0.06	55.8%	SL	0.07±0.06	53.4%
SE	50 ∈ [49, 51]	50.9±1.3	SK	0.18±0.06	55.6%	SE	0.05±0.07	51.8%
SH	46 ∈ [44, 49]	45.1±1.5	SI	0.16±0.06	55.5%	SF	0.04±0.06	50.9%
BD	46 ∈ [43, 47]	45.3±1.4	SE	0.16±0.05	54.9%	SI	0.04±0.08	50.9%
SD	45 ∈ [43, 47]	44.7±1.3	BD	0.14±0.06	54.8%	SD	0.02±0.07	52.2%
BM	43 ∈ [42, 45]	42.9±1.3	SD	0.14±0.06	55.0%	BM	-0.01±0.06	49.9%
SI	40 ∈ [39, 43]	41.4±1.4	SB	0.13±0.06	55.0%	SJ	-0.03±0.05	49.1%
SK	37 ∈ [35, 40]	40.2±1.5	SA	0.11±0.06	53.6%	SC	-0.03±0.05	49.1%
SA	30 ∈ [29, 31]	32.0±1.3	SH	0.09±0.07	52.9%	SK	-0.06±0.05	47.4%
SB	24 ∈ [23, 27]	27.4±1.3	SL	0.05±0.05	51.7%	SG	-0.09±0.08	46.7%
SC	9 ∈ [9, 9]	11.6±0.9	SC	-0.02±0.04	49.1%	SH	-0.21±0.05	44.0%

7 DISCUSSION

The subjective evaluation results have shown that our system is capable of generating of co-speech gestures that are human-like and speech appropriate. The high performance on the speech appropriateness shows that the current system is a promising approach to achieve semantically-aware co-speech gesture generation in virtual agents.

Our system was top-ranked in the human-likeness and appropriateness for agent speech evaluations, while receiving one of the lowest scores in the appropriateness to interlocutor evaluation. This might seem a bit counter intuitive, given that we indeed trained the system to listen to the interlocutor. We believe that there are multiple factors at play here and will outline them below. First, out system was trained to take in speech information of the interlocutor as input (in the form of CSMP embeddings), but we chose to not include interlocutor motion as one of the inputs, due to time constraints. Feeding interlocutor motion as input might have rendered a system capable of mirroring/mimicry, similar to [18] which could have resulted in a higher rating. Secondly, we would like to discuss another possible explanation, which stems from the nature of the

data and how the evaluation was carried out. In the appropriateness evaluations, each system was compared against itself, and the objective was to see to what degree raters could distinguish motion that matched the context from mis-matched motion. As mentioned in section 4.1, there was a certain amount of cross-talk present in the data, i.e. the interlocutor audio was present in the main agent's channel and vice versa. We took extra measures to eliminate such cross-talk, because not doing so would have resulted in the agent performing co-speech gestures also while listening, based on the cross-talk from the interlocutor. Inspecting the evaluation stimuli based on the output from the different systems in the challenge, it is clear that this seems to happen in certain systems. We can further speculate that such an agent might in fact score favourably in the match/mismatch paradigm, because the gestures would indeed be interlocutor aware. Future work on improving the interlocutor appropriateness could involve conditioning on interlocutor motion, as mentioned above, or training a separate model for listening behavior.

Additional evaluations on the semantic gesture generation capabilities of the model could be of interest for future work. In theory, our model is capable of capturing the semantic relations between speech and gesture spaces through the CSMP model. However, the current subjective evaluation is a bit limited in measuring the semantic gesture generation capabilities of the model, as it is difficult to disentangle from other aspects, such as speech-gesture synchrony. Objective evaluation metrics for semantic appropriateness could be helpful in quantifying and improving our system in this regard.

8 CONCLUSIONS

In this paper we described our entry system to the GENEA Challenge 2023. We presented a system, which builds on an existing diffusion based motion synthesis model and proposed a conditioning signal, which utilizes audio, text and motion data. For this we proposed a CLIP-like contrastive pre-training module, contrastive speech and motion pretraining (CSMP) in order to capture the underlying relations between speech and motion. Our system achieved top performance in human-likeness and speech appropriateness amongst the submitted entries, which proves that our system is a promising approach to generate human-like co-speech gestures in agents. Our system ranked relatively low in interlocutor approrpiateness, which is a focus in future work for improvement. Human-like, semantic and interlocutor appropriate co-speech gesture generation in virtual agents is still an open problem. Our systems high performance in the subjective evaluations is encouraging and indicates that our submitted model is a promising way to achieve these goals.

ACKNOWLEDGMENTS

This work was partially supported by the Advanced Adaptive Intelligent Agents project (Digital Futures), the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation and by grant no. 20023495 (Development of behavior-oriented HRI AI technology for long-term interaction between service robots and users) funded by the Korean Ministry of Trade, Industry and Energy (MOTIE).

REFERENCES

[1] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. *Computer Graphics Forum* 39, 2 (2020), 487–496. https://doi.org/10.1111/cgf.13946 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13946

[2] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2023. Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models. *ACM Trans. Graph.* 42, 4 (2023), 1–20. https://doi.org/10.1145/3592458

[3] Tenglong Ao, Zeyi Zhang, and Libin Liu. [n. d.]. GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents. *ACM Trans. Graph.* ([n. d.]), 18 pages. https://doi.org/10.1145/3592097

[4] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*. PMLR, 1298–1312.

[5] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. BEAT: The Behavior Expression Animation Toolkit. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*. Association for Computing Machinery, New York, NY, USA, 477–486. https://doi.org/10.1145/383259.383315

[6] Che-Jui Chang, Sen Zhang, and Mubbasir Kapadia. 2022. The IVI Lab Entry to the GENEA Challenge 2022 – A Tacotron2 Based Method for Co-Speech Gesture Generation With Locality-Constraint Attention Mechanism. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 784–789. https://doi.org/10.1145/3536221.3558060

[7] Changyou Chen, Jianyi Zhang, Yi Xu, Liqun Chen, Jiali Duan, Yiran Chen, Son Dinh Tran, Belinda Zeng, and Trishul Chilimbi. 2022. Why do We Need Large Batchsizes in Contrastive Learning? A Gradient-Bias Perspective. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=T1dhAPdS--

[8] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18000–18010.

[9] Robert O. Davis. 2018. The impact of pedagogical agent gesturing in multimedia learning environments: A meta-analysis. *Ed. Res. Rev.-Neth.* 24 (2018), 193–209.

[10] Anna Deichler, Siyang Wang, Simon Alexanderson, and Jonas Beskow. 2023. Learning to generate pointing gestures in situated embodied conversational agents. *Frontiers in Robotics and AI* 10 (2023), 1110534.

[11] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

[12] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. 2023. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. *Computer Graphics Forum* 42, 1 (2023), 206–216. https://doi.org/10.1111/cgf.14734 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14734

[13] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. arXiv:2005.08100 [eess.AS]

[14] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney, NSW, Australia) *(IVA '18)*. Association for Computing Machinery, New York, NY, USA, 79–86. https://doi.org/10.1145/3267851.3267878

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Proc. NeurIPS* (2020), 6840–6851.

[16] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[17] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. MotionGPT: Human Motion as a Foreign Language. *arXiv preprint arXiv:2306.14795* (2023).

[18] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. 2020. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[19] Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, and Gustav Eje Henter. 2021. HEMVIP: Human evaluation of multiple videos in parallel. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 707–711.

[20] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761* (2020).

[21] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing Input and Output Representations for Speech-Driven

Gesture Generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 97–104. https://doi.org/10.1145/3308532.3329472

[22] Taras Kucherenko, Patrik Jonell, Sanne Van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 242–250.

[23] Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '23)*. ACM.

[24] Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '23)*. ACM.

[25] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha Srinivasa, and Yaser Sheikh. 2019. Talking With Hands 16.2M: A Large-Scale Dataset of Synchronized Body-Finger Motion and Audio for Conversational Motion Analysis and Synthesis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 763–772. https://doi.org/10.1109/ICCV.2019.00085

[26] Jina Lee and Stacy Marsella. 2006. Nonverbal Behavior Generator for Embodied Conversational Agents. In *Intelligent Virtual Agents*, Jonathan Gratch, Michael Young, Ruth Aylett, Daniel Ballin, and Patrick Olivier (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 243–255.

[27] Margot Lhommet, Yuyu Xu, and Stacy Marsella. 2015. Cerebella: Automatic Generation of Nonverbal Behavior for Virtual Humans. *Proceedings of the AAAI Conference on Artificial Intelligence* 29, 1 (Mar. 2015). https://doi.org/10.1609/aaai.v29i1.9778

[28] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11293–11302.

[29] David McNeill. 2008. *Gesture and Thought*. University of Chicago Press.

[30] Shivam Mehta, Siyang Wang, Simon Alexanderson, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2023. Diff-TTSG: Denoising probabilistic integrated speech and gesture synthesis. arXiv:2306.09417 [eess.AS]

[31] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. *Comput. Graph. Forum* (2023).

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[34] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Proc. NeurIPS* (2019).

[35] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *Proc. ICLR*.

[36] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. 2017. Speech-to-Gesture Generation: A Challenge in Deep Learning Approach with Bi-Directional LSTM. In *Proceedings of the 5th International Conference on Human Agent Interaction* (Bielefeld, Germany) *(HAI '17)*. Association for Computing Machinery, New York, NY, USA, 365–369. https://doi.org/10.1145/3125739.3132594

[37] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 2022. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*. Springer, 358–374.

[38] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022).

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[40] Ulme Wennberg and Gustav Eje Henter. 2021. The case for translation-invariant self-attention in transformer-based language models. *arXiv preprint arXiv:2106.01950* (2021).

[41] Bowen Wu, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. 2021. Modeling the Conditional Distribution of Co-Speech Upper Body Gesture Jointly Using Conditional-GAN and Unrolled-GAN. *Electronics* 10, 3 (2021). https://doi.org/10.3390/electronics10030228

[42] Bowen Wu, Chaoran Liu, Carlos T. Ishi, and Hiroshi Ishiguro. 2021. Probabilistic Human-like Gesture Synthesis from Speech Using GRU-Based WGAN. In *Companion Publication of the 2021 International Conference on Multimodal Interaction* (Montreal, QC, Canada) *(ICMI '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 194–201. https://doi.org/10.1145/3461615.3485407

[43] Payam Jome Yazdian, Mo Chen, and Angelica Lim. 2022. Gesture2Vec: Clustering Gestures using Representation Learning Methods for Co-speech Gesture Generation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 3100–3107. https://doi.org/10.1109/IROS47612.2022.9981117

[44] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 4303–4309.

[45] Fan Zhang, Naye Ji, Fuxing Gao, and Yongping Li. 2023. DiffMotion: Speech-Driven Gesture Synthesis Using Denoising Diffusion Model. In *MultiMedia Modeling*, Duc-Tien Dang-Nguyen, Cathal Gurrin, Martha Larson, Alan F. Smeaton, Stevan Rudinac, Minh-Son Dao, Christoph Trattner, and Phoebe Chen (Eds.). Springer International Publishing, Cham, 231–242.

[46] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. *arXiv preprint arXiv:2301.06052* (2023).

Paper B

Paper B

Learning to Generate Pointing Gestures in Situated Embodied Conversational Agents

ANNA DEICHLER, SIYANG WANG, SIMON ALEXANDERSON, AND JONAS BESKOW

Published in
Frontiers in Robotics & AI, 2023

frontiers | Frontiers in Robotics and AI

TYPE Original Research
PUBLISHED 30 March 2023
DOI 10.3389/frobt.2023.1110534

Check for updates

OPEN ACCESS

EDITED BY
Nutan Chen,
Volkswagen Group, Germany

REVIEWED BY
Izidor Mlakar,
University of Maribor, Slovenia
Xingyuan Zhang,
ARGMAX.AI Volkswagen Group Machine Learning Research Lab, Germany

*CORRESPONDENCE
Anna Deichler,
deichler@kth.se

SPECIALTY SECTION
This article was submitted to Robot Learning and Evolution, a section of the journal Frontiers in Robotics and AI

RECEIVED 28 November 2022
ACCEPTED 16 March 2023
PUBLISHED 30 March 2023

CITATION
Deichler A, Wang S, Alexanderson S and Beskow J (2023), Learning to generate pointing gestures in situated embodied conversational agents.
Front. Robot. AI 10:1110534.
doi: 10.3389/frobt.2023.1110534

COPYRIGHT
© 2023 Deichler, Wang, Alexanderson and Beskow. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Learning to generate pointing gestures in situated embodied conversational agents

Anna Deichler*, Siyang Wang, Simon Alexanderson and Jonas Beskow

Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

One of the main goals of robotics and intelligent agent research is to enable them to communicate with humans in physically situated settings. Human communication consists of both verbal and non-verbal modes. Recent studies in enabling communication for intelligent agents have focused on verbal modes, i.e., language and speech. However, in a situated setting the non-verbal mode is crucial for an agent to adapt flexible communication strategies. In this work, we focus on learning to generate non-verbal communicative expressions in situated embodied interactive agents. Specifically, we show that an agent can learn pointing gestures in a physically simulated environment through a combination of imitation and reinforcement learning that achieves high motion naturalness and high referential accuracy. We compared our proposed system against several baselines in both subjective and objective evaluations. The subjective evaluation is done in a virtual reality setting where an embodied referential game is played between the user and the agent in a shared 3D space, a setup that fully assesses the communicative capabilities of the generated gestures. The evaluations show that our model achieves a higher level of referential accuracy and motion naturalness compared to a state-of-the-art supervised learning motion synthesis model, showing the promise of our proposed system that combines imitation and reinforcement learning for generating communicative gestures. Additionally, our system is robust in a physically-simulated environment thus has the potential of being applied to robots.

KEYWORDS

reinforcement learning, imitation learning, non-verbal communication, embodied interactive agents, gesture generation, physics-aware machine learning

1 Introduction

1.1 Overview

Humans rely on both verbal and non-verbal modes of communications in physically situated conversational settings. In these settings, non-verbal expression, such as face and hand gestures, often contain information which is not present in the speech. A prominent example is how people point to an object instead of describing it with words. This complementary function makes communication more efficient and robust than speech alone. In order for embodied agents to interact with humans more effectively, they need to adapt similar strategies, i.e., they need to both comprehend and generate non-verbal communicative expressions.

Regarding comprehension, one focus has been on gesture recognition in order to accomplish multimodal reference resolution and establishing common ground between human and agent Abidi et al. (2013); Häring et al. (2012); Wu et al. (2021). Regarding generation, the main focus has been on generating co-speech gestures that typically uses supervised learning techniques to map text or speech audio to motion [see Liu et al. (2021) for an overview]. While these approaches can generate natural looking gesticulation, they only model beat gestures, a redundant aspect of non-verbal communication. Moreover, these supervised methods require large labeled datasets tied to a specific embodiment and are not physics-aware, posing a major issue for transferring the results to robots.

The focus of our study is the generation of pointing gestures. According to McNeill's influential classification, pointing gestures are among the four primary types of gestures, alongside with iconical, metaphorical and beat gestures McNeill (1992). Iconic gestures depict concrete objects or actions with the hands. Metaphoric gestures express abstract concepts or ideas with the hands. Beat gestures are simple movements that emphasize or segment speech. Deictic gestures point to objects, locations, or persons in space. Pointing is the most specialized of all gestures in terms of its association with particular verbal expressions, specifically reference and referring expressions Kibrik (2011). Linguistic reference is the act of using language to refer to entities or concepts. Linguistic elements that perform a mention of a referent are called referring expressions or referential devices. People use reference in discourse to draw the listener's attention to a referent or target, which can be in either the speech-external (deictic reference) or speech-internal (anaphoric reference) environment. The anaphoric referent is an element of the current discourse, while the deictic referent is outside the discourse in the spatio-temporal surroundings Talmy (2020). Deictic expressions are used to indicate a location or point in time relative to the deictic centre, which is the center of a coordinate system that underlies the conceptualization of the speech situation. They are essential parts of human communication, since they establish a direct referential link between world and language. Demonstratives like "this," "that" or "there" are the simplest form of deictic expressions that focus the interlocutor's attention to concrete entities in the surrounding situation Peeters et al. (2021). In everyday conversations people often use non-verbal means (e.g., eyes, head, posture, hands) to indicate the location of the referent alongside with verbal expressions to describe it. Pointing and eye gaze are the most prominent non-verbal means of deictic reference and also play a key role in establishing joint attention in human interactions Diessel and Coventry (2020), a prerequisite for coordination in physical spaces. Pointing is ubiquitous in adult interaction across settings and it has been described as "a basic building block" of human communication Kita (2003). Demonstratives produced with pointing gestures are more basic than demonstratives produced without pointing Cooperrider (2020). Speakers also tend to prefer short deictic descriptions when gesture is available, in contrast to longer non-deictic descriptions when gesture is not available Bangerter (2004).

We proposed a new framework based on imitation and Reinforcement Learning (imitation-RL) for this for generating referential pointing gestures in physical situated environment. Our method, adapted from Peng et al. (2021), learns a motor control policy that imitates examples of pointing motions while pointing accurately. By providing only a few pointing gesture demonstrations, the model learns generalizable accurate pointing through combined naturalness and accuracy rewards. In contrast to supervised learning, our method achieves high pointing accuracy, requires only small amounts of data, learns physically valid motion and has high perceived motion naturalness. It is also considerably more lightweight than compared supervised-learning methods. To implement the method, we first collected a small dataset of motion-captured pointing gesture with accurate pointing positions recorded. We then trained a motor control policy on a humanoid character in a physically simulated environment using our method. The generated gestures were then compared against several baselines to probe both motion naturalness and pointing accuracy. This was done through two subjective user studies in a custom-made virtual reality (VR) environment. In the accuracy test, a full-body virtual character points at one of several objects, and the user is asked to choose the object that the character pointed at, a setup known as referential game Lazaridou et al. (2018), Steels (2001). This setup is also motivated by the notion that language use is a triadic behaviour involving the speaker, the hearer, and the entities talked about Bühler (1934).

We found that the proposed imitation-RL system indeed learns highly accurate pointing gestures while retaining high level of motion naturalness. It is also shown that imitation-RL is much better at both accuracy and naturalness than a supervised learning baseline. The results suggest that imitation-RL is a promising approach to generate communicative gestures.

1.2 Related work

Our work concerns learning communication in embodied agents through reinforcement learning, specifically learning pointing gestures in a physically simulated environment with imtation-RL.

Below we first review related work in the area of non-verbal deictic expression and gesture generation for embodied agents, followed by an overview of two reinforcement learning fields related to our work: (a) reinforcement learning for communication, (b) reinforcement learning for motion control. We believe that our work is the first that learns non-verbal communication with reinforcement learning, which can be seen as bridging the two currently disjunct reinforcement learning fields.

1.2.1 Gesture and non-verbal deictic expression generation

There has been substantial work considering the generation of speech-driven gestures in embodied agents. Early work in synthesis of gestures for virtual agents employed methods based on rules, inverse kinematics and procedural animation to generate co-speech gestures Cassell et al. (1994), Cassell et al. (2001); Kopp and Wachsmuth (2004); Ng-Thow-Hing et al. (2010); Marsella et al. (2013). Rule-based approaches, however, often require laborious manual tuning and often struggle to achieve a natural motion quality. More recently, supervised learning systems for co-speech gesture generation have achieved high naturalness in certain settings [Liu et al. (2021)]. Some of these systems take only speech

audio as input [Hasegawa et al. (2018); Kucherenko et al. (2021a); Ginosar et al. (2019); Ferstl et al. (2020); Alexanderson et al. (2020)] but are limited to generating beat gestures aligned with speech. Extending these supervised learning based gesture generation methods to produce iconic, metaphoric or deictic gestures would require more elaborate modelling and a large dataset containing these gestures Ferstl and McDonnell (2018). Adding text as input can potentially generate more semantic gestures but is still limited Kucherenko et al. (2020); Yoon et al. (2020); Ahuja et al. (2020); Korzun et al. (2020). Common for all these methods is that they do not take into account the surrounding environment and thus cannot communicate other than redundant information. On the contrary, works related to deictic expression generation for intelligent agents take into account the physical environment in which the agent is situated. Pointing and gaze generation in general have been extensively studied in virtual agents and robots. Rule-based approaches for deictic expression generation are often used in virtual agents Noma et al. (2000); Rickel and Johnson (1999); Lester et al. (1999), as well as robotics. Fang et al. (2015) and Sauppé and Mutlu (2014) both implemented rule-based pointing gestures in a humanoid robot for referential communication. Holladay et al. (2014) proposes a mathematical optimization approach to make legible pointing, i.e., disambiguation of closely situated objects. In Sugiyama et al. (2007) natural generate rule-based deictic expressions consisting of pointing and a reference term are generated in an interactive setting with a robot. Referential gaze has been studied in virtual agents Bailly et al. (2010), Andrist et al. (2017), as well as robotics Mutlu et al. (2009). Some studies have adopted learning-based approaches to develop pointing gestures generation. Huang and Mutlu (2014) proposes a learning based modeling approach using dynamic Bayesian network to model speech, gaze and gesture behavior in a narration task. Zabala et al. (2022) developed a system of automatic gesture generation for a humanoid robot that combines GAN based beat gesture generation with probabilistic semantic gesture insertion, including pointing gestures. Chao et al. (2014) takes a developmental robotics approach to generating pointing gestures using reinforcement learning for a 2D reaching task. None of the above works consider human-like and physics based pointing motion generation in 3D space.

1.2.2 RL for communication in task-oriented embodied agents

There have been a substantial amount of works in recent years that focused on extending the capabilities of embodied RL agents with language. This field has been reviewed in a recent survey Luketina et al. (2019) which separates out these works into two categories: language-conditioned RL, where language is part of the task formulation and language-assisted RL, where language provides useful information for the agent to solve the task. The simplest form of language-conditioned is instruction following, in which a verbal instruction of the task is provided to the agent. Examples are manipulation Stepputtis et al. (2019) and navigation task Qi et al. (2020). These studies focused more on comprehension of verbal expressions and usually no interaction takes place between the user and the agent. In more complex settings, the agent interacts with humans using natural language. In Lynch et al. (2022) the authors present a RL and imitation based framework, Interactive Language, that is capable of continuously adjusting its behavior to natural language based instructions in a real-time interactive setting. There have also been a recent wave of datasets and benchmarks created by utilizing 3D household simulators and crowd sourcing tools to collect large-scale task-oriented dialogue aimed at improving the interactive language capabilities of embodied task-oriented agents Padmakumar et al. (2022), Gao et al. (2022), Team et al. (2021). Most of the above mentioned works focus on the verbal mode of communication and largely on the comprehension side (e.g., instruction following). Less work has explored the non-verbal communication in situated embodied agents Wu et al. (2021), especially on the generation side, which we aim to address in our work.

1.2.3 Learning motor control through RL

RL has been extensively applied to learning motor control for both robotics and graphics applications Duan et al. (2016); Heess et al. (2017). The learned motion dynamic for humanoid characters from these methods are usually not very human-like. To improve both the motion dynamics and also to facilitate learning, it is proposed for the learning agent to imitate expert demonstration or motion-capture animation Merel et al. (2017); Ho and Ermon (2016). However, it is not until DeepMimic Peng et al. (2018) that the RL-imitation approach achieved human-like motion. The drawback in DeepMimic is that it uses several manually tuned hyper-parameters in the imitation reward, a problem addressed by a follow-up work Adversarial Motion Prior (AMP) Peng et al. (2021) which replaces the "hard" imitation in DeepMimic with a learned discriminator (similar to GAN). Through a combination of imitation reward and task reward, methods like AMP can learn human-like motion dynamics while completing a manually defined task, for example, heading in a given direction while walking or spin-kick a given target. It has been shown that even with just one motion clip as imitation target, AMP learns generalizable motion control given random task input (e.g., randomly sampled walking direction) while maintaining high motion naturalness.

2 Materials and methods

2.1 Pointing gesture data collection

Our method for synthesizing human-like pointing gestures for the humanoid character requires an appropriate full-body motion capture dataset with diverse and accurate target locations covering the 3D space surrounding the character. Multiple datasets exist of referring expression with pointing as non-verbal modality Schauerte and Fink (2010); Matuszek et al. (2014); Shukla et al. (2015); Chen et al. (2021), but since most of these focus on comprehension, they lack full body motion capture, which is essential for our method. Existing full-body motion capture datasets for gesture generation Ferstl and McDonnell (2018), Ferstl et al. (2020) focus on beat gestures with the aim to train speech-to-gesture generative models and do not contain pointing gestures. A recent dataset Islam et al. (2022) records multi-modal referring expressions, including pointing gestures, but it is restricted to a tabletop setting.

To obtain a collection of ground truth examples for training and evaluation, we recorded a pointing gesture dataset in a optical

motion capture studio equipped with 16 Optitrack Prime 41 cameras. The actor wore a suit with 50 passive markers, a pair of Manus data gloves (hand motion capture) and a head-mounted iPhone 12 (face motion and voice capture). In addition, we also recorded a pointing target, consisting of a rigid structure equipped with 4 markers. The marker data was solved to a representation of the actor's skeleton and the location of the target in 3D space using the systems software (Motive 3.0).

Three different pointing tasks were recorded: single target pointing, two targets selection and two targets moving (point-and-place). In each setting the targets were moved around in order to get a coverage of the surrounding space, while the actor was in a stationary position. Between each pointing task the actor returned to a neutral stance. The beginning and the end of the movement is defined by the pointing hand leaving and returning to the initial, downward position. We used this information to parse the continuous recording of motion capture data into pointing clips. For more details, see Deichler et al. (2022).

We focus on single target pointing in this study. Here, we provide an overview of that data. We first divided the pointing target positions to front and back. We only use the front data in our study. The total sum of single target positions in the dataset is 83, from which 52 are front target pointing movements (25 left-handed and 27 right-handed). The actor's dominant hand is right hand. The target distribution of this subset is visualized in 3D space in **Figure 1** (blue balls). The single target dataset is parsed into single arm pointing movement based on peaks in displacement on the sagittal plane (yz). We further analyse the pointing movements in terms of pointing accuracy (**Figure 2**) and velocity (**Figure 3**) dividing into dominant (right) and non-dominant (left) hand. The pointing accuracy is calculated based on inverse of alignment angle between arm and arm to pointing target (Eq. **2**), the more accurate pointing the higher this measure is. As seen in **Figure 2**, both dominant and non-dominant hands have a bell-shaped accuracy curve. This reflects the pointing motion trajectory: the hand first moves to pointing position as the accuracy rises, it stays there for some time (holding), and then retracts. Moreover, we note that neither hand achieves maximum accuracy. This shows that human pointing is not accurate in terms of alignment angle, which has been shown in previous pointing gesture studies Lücking et al. (2015). The velocity trajectory (**Figure 3**) shows clear correspondence with accuracy trajectory. No significant difference between dominant and non-dominant hands was found.

We applied left-right mirroring on the mocap data, which doubled the total amount of data available for model training. This is mainly due to the baselines MoGlow (**Section 2.4.1.1**) and GT-nn (**Section 2.4.1.2**) require large amount of data to perform well. For fair comparison, we used the same mirrored data for training all systems including our proposed method even though our proposed method does not have the same data quantity requirement as the baselines.

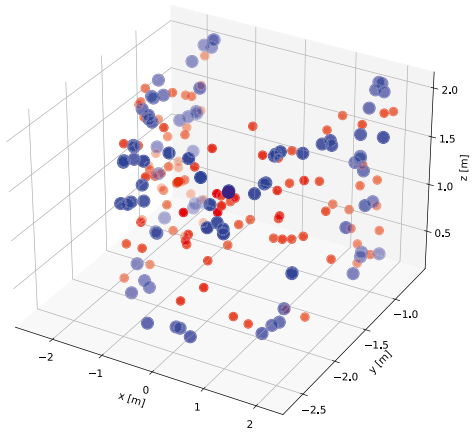


FIGURE 1
Targets from training set (blue) and perceptual test set (red) visualized in 3D space.

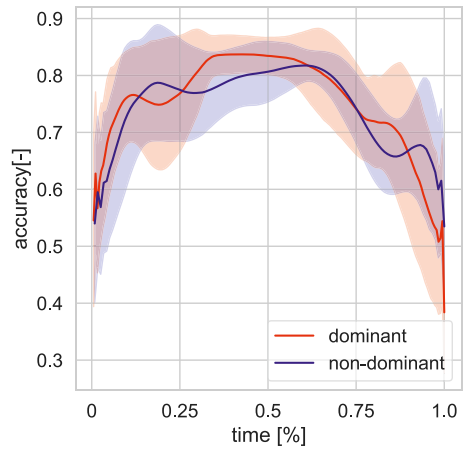


FIGURE 2
Averaged pointing accuracy profile for front targets in training set for the dominant and non-dominant hands.

2.2 Method

Our method is based on imitation-RL method AMP Peng et al. (2021), capable of learning complex naturalistic motion on humanoid skeletons and showing good transferrability of simulation-learned policy to real-world robots Escontrela et al. (2022), Vollenweider et al. (2022). AMP is a subsequent work of Generative Adversarial Imitation learning Ho and Ermon (2016), that takes one or multiple clips of reference motion and learns a motor control policy π_θ that imitates the motion dynamics of the reference (s) through a discriminator network D_ϕ. An optional task reward can be added for the specific motion task to encourage task completion in the learned policy, for example, heading in a given direction while walking. The full reward function is the combination of rewards from motion imitation r_t^I and the task reward r_t^G, weighted

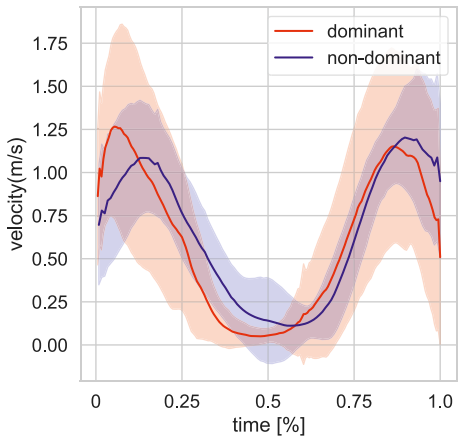


FIGURE 3
Averaged hand velocity profile for front targets in training set for the dominant and non-dominant hands.

respectively by ω^I, ω^G,

$$r_t = \omega^I r_t^I + \omega^G r_t^G. \quad (1)$$

The policies are trained using a combination of GAIL and PPO [Proximal Policy Optimization Schulman et al. (2017)]. The schematic view of the system is presented on **Figure 4**. We kept $\omega^I = \omega^G = 0.5$ and in all AMP models. We used same action and state space as humanoid motion example presented in original AMP Peng et al. (2021). The state is defined as the character's configuration in the character's local root joint based coordinate system. The state vector includes the relative positions of each link with respect to the root, their rotation and linear and angular velocities. The actions specify the target positions in the PD controllers for the controllable joints.

We designed a reward function r_t^{Pt} for pointing accuracy that takes the angle between the hand-to-target vector $\overrightarrow{V_{HT}}$ and the elbow-to-hand vector $\overrightarrow{V_{EH}}$, and gives higher reward to smaller angle, i.e., more accurate pointing. This is illustrated in **Figure 5**. The angular distance from the index finger ray has been used as a heuristics for pointing accuracy Lücking et al. (2015), here we use the lower-arm extension, due to the limitations of the simulated agent's morphology. Similarly to previous closely related work in humanoid motion generation with RL and imitation Peng et al. (2021), Peng et al. (2018), we apply exponential scaling to angular distance term to form the task reward function as follows,

$$\hat{\theta} = 1 - \frac{\angle\left(\overrightarrow{V_{HT}}, \overrightarrow{V_{EH}}\right)}{\pi} \quad , \quad r_t^{Pt} = \frac{e^{\hat{\theta}} - 1}{e} \quad (2)$$

Since our task is to point at a given position, the x-y-z coordinates of the pointing position are concatenated to policy newtork input. Additionally, it is common in methods like AMP to sample the task input randomly to achieve good generalization Peng et al. (2021); Peng et al. (2018), for example, to randomly sample walking directions when learning directional walking. We also applied this approach by randomly sampling pointing target positions in the training process. The sampling process takes in a ground-truth pointing position (corresponding to the current pointing gesture demonstration) as input and uniformly samples points within a square box of size 20 × 20 × 20 cm centered at that ground-truth position.

Through preliminary experiments, we found that the original AMP algorithm does not learn pointing gesture effectively. We subsequently implemented changes described in the following sections to improve AMP in learning pointing gesture.

2.2.1 Phase input to network: AMP base

In motion imitation it is common to provide as input normalized running time called phase ($p \in [0, 1]$, 0 is beginning of a clip, 1 is the end) to the policy network in order to synchronize the simulated character with the reference motion. This was proven to be effective in imitation of single motion clips in DeepMimic Peng et al. (2018). In AMP this is not needed, since the policies are not trained to explicitly imitate a single motion clip. However, we found that vanilla AMP without phase input struggles to learn pointing gesture motion dynamics, thus we provided the policy network with phase to help the system learn the dynamics of the pointing motion. We call this model AMP base.

2.2.2 Phase-functioned neural network with pointing target clustering: AMP-pfnn

We found that the discriminator network in AMP, which learns what "plausible" motion looks like through adversarial loss and provide imitation reward to the policy network, has difficulty learning pointing gesture dynamics even with phase as input. Specifically, it has trouble distinguishing between the raising phase of pointing gesture and the retraction phase, resulting in the learned motion stuck in the middle between idle and pointing positions. This is unsurprising since both phases of the pointing gesture have similar trajectories (in opposite directions). We tackled this issue by changing discriminator network from a simple MLP (as in original AMP) to a phase-functioned neural network (PFNN) Holden et al. (2017).

Through early experiments, we also found that having a single network to learn pointing in all directions can be challenging. Thus, we grouped the training set pointing positions into clusters in the 3D space. The clustering is done with the MeanShift algorithm, which is based on kernel density estimation (KDE). Using Scott's rule the estimated bandwidth is 0.78 for the univariate KDE, which results in 8 clusters in the target space. The estimated bandwidth is quite wide given the target space limits, but this is expected with a limited amount of data. Based on this clustering, we separated the motion clips into subgroups and trained specialized networks for the different subgroups. During test time, the subgroup which contains the closest training target to the test target in terms of Euclidean distance is chosen to generate the pointing movement. We denote this version of AMP with pfnn and cluster group training as AMP-pfnn.

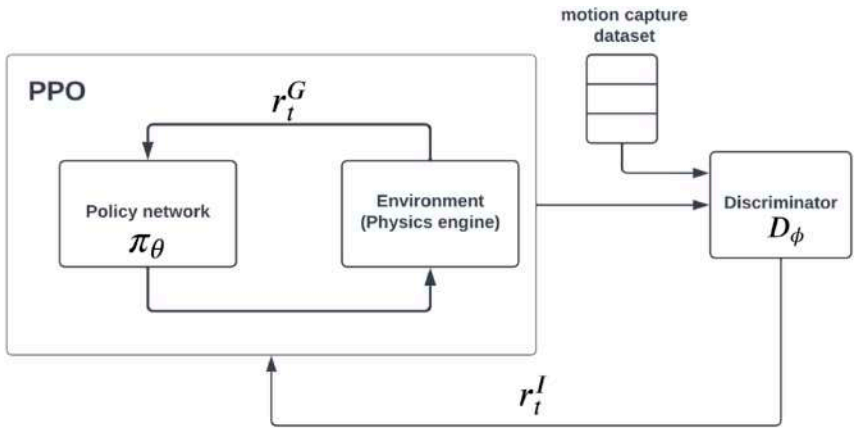


FIGURE 4
Schematic view of the AMP framework, which allows using a composite reward of task $r_t^G$and imitation r_t^I rewards by combining RL policy gradient methods (PPO) with a GAN-like disriminator.

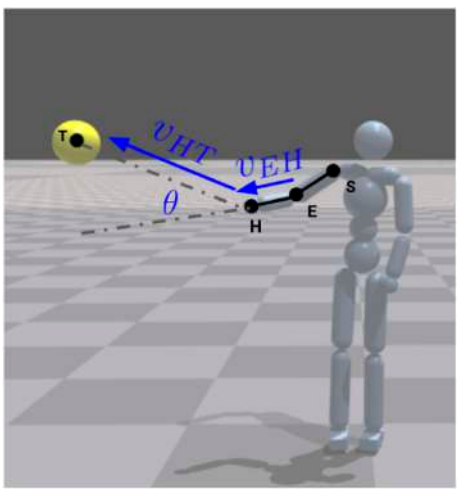


FIGURE 5
Pointing reward based on angle between vector from elbow joint (E) to hand joint (H) and vector from hand joint to target position (T). Agent rendered in the IssacGym physics engine Makoviychuk et al. (2021).

2.3 Implementation

We use a GPU-based physics simulation software Isaac gym Makoviychuk et al. (2021), a rendered example of which can be seen in **Figure 5**. For implementation of our proposed method, we modified an existing AMP implementation provided by authors of Isaac gym[1]. The modifications also include enabling learning pointing gesture in humanoid characters and adding our proposed reward function for pointing accuracy.

2.4 Experiment

The goal of our proposed system is to produce natural and accurate pointing gestures in an interactive agent. In order to evaluate both aspects in an embodied interaction, we created a novel virtual reality (VR) based perceptual evaluation using a 3D game engine. This is described with more details in **Section 2.4.2**. We further probed the specific aspects of our proposed model in an object evaluation. We benchmarked our proposed method with two baselines, a supervised learning motion synthesis model MoGlow, and a simple nearest-neighbour retrieval method.

2.4.1 Baselines

2.4.1.1 MoGlow

We used a state-of-the-art animation synthesis model MoGlow Henter et al. (2020) as supervised learning baseline. This model is based on machine learning framework normalizing flow Dinh et al. (2014) Dinh et al. (2016) which directly estimates data probability distribution with invertible neural network. MoGlow utilizes a popular normalizing flow architecture Glow Kingma and Dhariwal (2018) and is able to synthesize animation in a conditional or unconditional probabilistic manner, and has shown state-of-the-art performance in synthesizing locomotion Henter et al. (2020), dancing Valle-Pérez et al. (2021), and gesture Alexanderson et al. (2020).

1 https://github.com/NVIDIA-Omniverse/IsaacGymEnvs.

We extended the conditional generation mode of MoGlow to achieve pointing gesture synthesis. We provided the model with a 4-dimensional vector as condition, consisting of pointing target position (x-y-z coordinates) and an indicator variable of whether the current frame is pointing or idling. This indicator variable differs from the AMP phase variable, as it takes binary values, whereas the phase variable is the normalized running time. This difference stems from a fundamental difference between the two models. The MoGlow model is based on autoregressive supervised learning, and therefore has a "sense of time," AMP is RL based and does not get the history as input, therefore the phase in used as an input.

The data preparation consists of first downsampling the mocap data to 20 frames per second, then cutting it into training clips with a sliding window of window length 6 s and step size 0.5 s. Such data preparation is consistent with prior studies using MoGlow for locomotion generation Henter et al. (2020) and gesture generation Alexanderson et al. (2020).

2.4.1.2 Nearest-neighbor queried ground-truth animation (GT-NN)

We also created a simple baseline through nearest-neighbor querying of the dataset. Given an input pointing position, this baseline model queries the closest pointing position in the dataset and plays the corresponding pointing gesture animation. We call this baseline Ground-truth Nearest-Neighbor (GT-nn). This baseline is the topline in motion naturalness since it is playing back motion-captured animation. On the other hand, its performance in referential accuracy would be determined by how close is the input position to positions in the dataset. If the input is close to some pointing position already existing in the dataset, then this method is expected to perform well, otherwise poorly. Thus, comparing a more complex learned model with this simple baseline in referential accuracy should reveal how much the learned model is able to generalize beyond the training data to unseen pointing target positions.

2.4.2 VR perceptual test

The perceptual test is an embodied interaction in VR. A user is put in a shared 3D space with an embodied virtual agent. In this setting, the virtual agent makes pointing gestures and the user is asked to evaluate motion naturalness of the gesture and play a referential game, i.e., guessing which object is the virtual agent pointing at. The two aspects are evaluated in two separate stages as seen in **Figure 6**.

In the first stage, the naturalness of the pointing gesture is evaluated by presenting a pointing gesture to the user, then asking the user to rate "How natural do you find the animation?" on a 1-5 scale [in **Figure 6A**]. The avatar is in a fixed position, facing the user in this stage. We intended for this naturalness test to only evaluate motion dynamics and not pointing accuracy, so we did not show the pointing object at this stage so that the user is not distracted by potentially inaccurate pointing. It is likely that a user would rate a pointing motion as unnatural even though the motion dynamics itself is natural but it does not point accurately at the object. We avoided this by not showing the pointing object at this stage. This evaluation setup is similar to the one in the GENEA co-speech generation challenge Kucherenko et al. (2021b). In the GENEA challenge the evaluation metrics included "human-likeness" and "appropriateness" where the first one measures motion quality, like naturalness in our study, and the second one measures task performance - in their case how well motion matches speech, in our case the equivalent metric is pointing accuracy.

In the second stage, the pointing accuracy is evaluated through a simplified embodied referential game Steels (2001),Lazaridou et al. (2018). The participants were presented with 3 balls in the environment, 1 pointing target and 2 distractors (the sampling process are described in **Section 2.4.3**). In this stage, the avatar's relative position to the user is varied between "across" and "side-by-side" conditions. In case of the "side-by-side" condition, the user stands by the shoulder of the avatar, allowing for a shared perspective on the target. In case of the "across" condition, the avatar is facing the user. After observing the agent's pointing motion, participants were asked to guess which object the agent was pointing at. The target selection mechanism was implemented using a raycast selection [red ray seen in of **Figures 6B, C**]. The raycast selection becomes available to the user after the motion has ended and this is indicated to the user by the objects color changing from black to white. This setup prevents the user from selecting before seeing the entire motion.

Each participant saw 5 samples for each of the 4 models in motion naturalness test (first stage) and 10 samples for each model in referential accuracy test (second stage). Thus each participant sees 20 samples in total for motion naturalness test and 40 in total for referential accuracy test. There are more samples in the referential accuracy test because a model's pointing accuracy may vary depending on the pointing position thus requires more samples to estimate true mean, while motion naturalness is easier to judge with less samples since it varies little given different pointing positions. We separated the two stages, i.e., the participant would finish motion naturalness test (first stage) before doing referential accuracy test (second stage). We also randomized ordering of the samples within each test to mitigate unwanted factors that could bias resulting statistics, such as the participant might take some time to get used to the VR game setup thus making data in the early part of the test less reliable. Lastly, the pointing positions in both tests (5 for stage 1, 10 for stage 2) were randomly picked from the 100 sampled test positions (**Section 2.4.3**) and were shared among models for the same user to ensure fair comparison.

2.4.3 Sampling test pointing targets and distractors

We sampled 100 test pointing target positions within the range of ground-truth positions. The ground-truth positions roughly form a half-cylindrical shape as seen in **Figure 1**. We thus defined the range of ground-truth position as a parameterized half-cylinder. We used 3 parameters, the height of the cylinder, the arc, and the radius of the cylinder. Given the ranges of these 3 parameters obtained from the ground-truth data, we then sampled new positions by uniformly sampling these 3 parameters within their respective ranges. The sampled 100 test positions are visualized in **Figure 1** (red balls are sampled positions, blue balls are ground-truth).

For each sampled test position, we sampled 2 distractors in the following way. For each test position, we first sampled a distractor position the same way as the test positions (within the parameter ranges), we then checked if that distractor is within 20–40 cm away

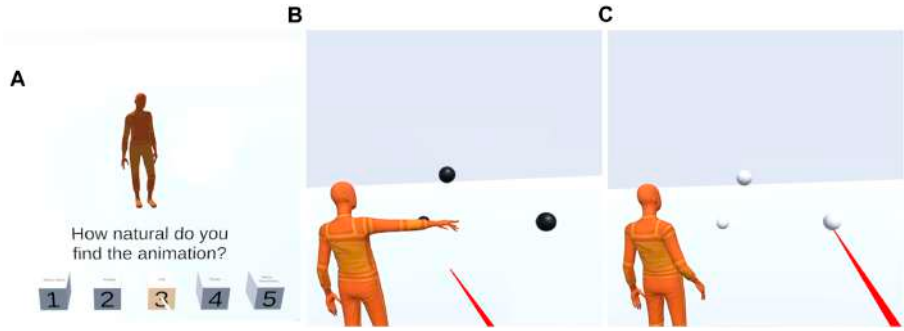


FIGURE 6
Examples of user view in the VR based perceptual test. Stage 1 shown in **(A)**: the user rates the naturalness of the pointing gesture. Stage 2 shown in **(B, C)**: the virtual agent first makes a pointing gesture in the presence of the actual pointing target and two distractor objects **(B)**, the user then guesses the pointed object (referential game) with the raycast selection mechanism **(C)**.

from the input test position. If not, we resampled. We repeated this process until we got 2 distractors that met the distance condition. This process was done for each of the 100 test pointing positions.

We note that the distance range 20–40 cm effectively determines the difficulty of the distractors. Referential game is more difficult if the distractors are closer to the actual pointing target. Through preliminary experiments, we found that 20–40 cm range is difficult enough that the model would have to point with a high level of accuracy, and at the same time not too difficult that even ground-truth pointing would not be able to distinguish the correct position from distractors. Admittedly, our process is not fully theoretically driven, and future studies could explore varying distractor distance or choose this distance in a more theoretical manner.

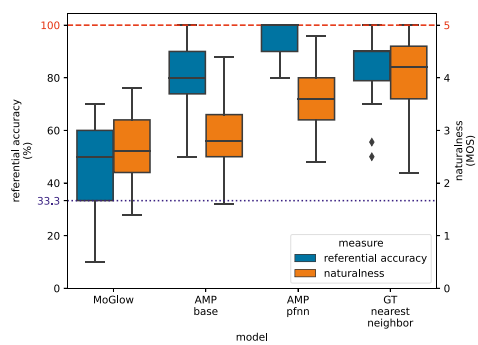


FIGURE 7
VR perceptual evaluation results. Red dashed line at the top indicates maximum possible values for referential accuracy and naturalness MOS. The blue dotted line at accuracy 33.3% corresponds to uniform guessing among 3 possible choices in the referential accuracy evaluation. All between-model differences in the two measures are significant (Wilcoxon signed-rank test with Holm-Bonferroni correction at $\alpha = 0.01$), except for referential accuracy between AMP-base and GT-nn.

3 Results

3.1 VR perceptual test results

We recruited 39 participants (age: min = 20, median = 23, max = 39; female: 16, male: 23). Before conducting statistical analysis on the results, we first averaged each participant's scores for each model in the two measures, referential accuracy and motion naturalness, thus obtaining a participant's average scores for the four models in the two measures. Subsequent analysis is done on this data.

Results from VR perceptual evaluation are shown in Figure 7. AMP-pfnn (2.2.2) obtains highest referential accuracy (mean = 95.5%) and highest naturalness MOS (mean = 3.57) among models. GT-nn (2.4.1.2) is second best overall at accuracy (mean = 85.6%) and is the naturalness MOS topline (mean = 4.01) since it is playing back motion-captured animation. AMP-base (2.2.1) obtains comparable referential accuracy (mean = 83.0%) as GT-nn, but has much worse naturalness (mean = 2.93) than either AMP-pfnn or topline GT-nn. MoGlow (2.4.1.1) performs worst in both accuracy (mean = 45.9%) and naturalness (mean = 2.61). All but one between-sample differences are significant at $\alpha = 0.01$ according to Wilcoxon signed-rank test with Holm-Bonferroni correction. The only insignificant difference is between AMP-base and GT-nn in accuracy measure. These results show that AMP-pfnn is able to point with very high referential accuracy in the presence of challenging distractors, and with high motion naturalness approaching ground-truth motion capture. Analysis on the avatar's relative position to the user in stage two shows that there is no significant between-sample difference for the "across" and "side-by-side" conditions at $\alpha = 0.01$ according to Wilcoxon signed-rank test with Holm-Bonferroni correction for any of the models. Further results on the position analysis can be found in Table 1.

10.3389/frobt.2023.1110534

TABLE 1 Referential accuracy results (mean and standard deviation) for "across" and "side-by-side" agent positions in VR perceptual evaluation.

Model	Across	Side-by-side
All	0.78 ± 0.41	0.76 ± 0.42
GT nearest neighbor	0.84 ± 0.37	0.88 ± 0.33
MoGlow	0.5 ± 0.5	0.39 ± 0.49
AMP base	0.83 ± 0.38	0.84 ± 0.37
AMP PFNN	0.96 ± 0.20	0.95 ± 0.22

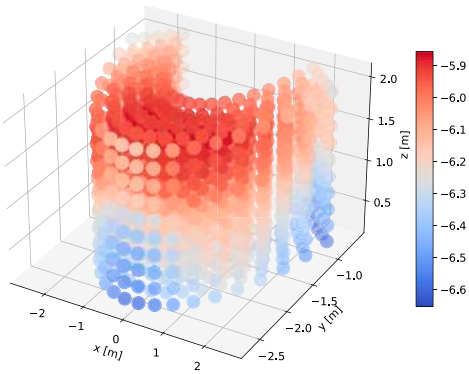


FIGURE 8
Visualization of KDE estimation values over the 3D grid for objective evaluation. The KDE was fitted on the training set targets.

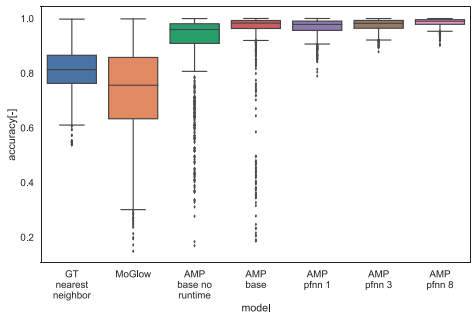


FIGURE 9
Comparison of objective evaluation results of normalized pointing accuracy for the different systems.

TABLE 2 Correlation (Spearman-r) between accuracy and KDE density measure.

Model	Acc corr. KDE	*p*-value
GT nearest neighbor	−0.118	< 0.001
MoGlow	0.718	< 0.001
AMP base no runtime	−0.020	0.532
AMP base	−0.128	< 0.001
AMP-pfnn 1	−0.043	0.176
AMP-pfnn 3	0.149	< 0.001
AMP-pfnn 8	−0.135	< 0.001

3.2 Objective evaluation results

We also conducted objective evaluations in order to gain further insights into the performance of the different models. This is done by calculating the pointing accuracy (Eq. 2) over a uniform grid in the 3D space. This is also a more principled way to measure the generalization capabilities of the learning systems. In this evaluation, the pointing accuracy was calculated from the pointing reward function (Figure 5). The grid was created by taking the x,y,z limits of the training set targets and generating a spherical grid of 1,000 test points (Figure 8). We defined the accuracy of the pointing movement as the maximum of the reward curve 1, where the velocity is below 0.5 m/s in a window of frames corresponding to 0.15 s around the time of pointing. Apart from the systems in the perceptual evaluation, we also added ablations to examine the effect of run time (phase) input to the AMP policy network (AMP base no runtime, compared to AMP base which has run time input Section 2.2.1), as well as varying the number of clusters in AMP-pfnn training (Section 2.2.2). Cluster variations include 1, 3 and 8, denoted as AMP-pfnn 1, AMP-pfnn 3 and AMP-pfnn 8, where the 8-cluster model is AMP-pfnn model evaluated in the perceptual test.

We visualized the results for the objective evaluation in Figure 9. AMP-pfnn 8 obtains the highest accuracy overall (mean = 0.62), consistent with perceptual evaluation results. The graphs also shows that using 8 clusters slightly outperforms using 1 or 3 clusters and also has less outliers. This further validates the use of clusters (and specifically 8) in AMP-pfnn for learning generalizable accurate pointing. The performance gain from concatenating the run time to the policy network input is clearly visible by comparing "AMP base no runtime" (mean = 0.56) and "AMP base" (mean = 0.59), where the latter has the runtime input. We can also observe that similarly to the perceptual evaluation, MoGlow performs worst amongst the compared systems (mean = 0.46), again showing the limitation of a supervised system for learning communicative gesture. Overall, the results from the objective evaluation show the same performance ranking as the perceptual referential accuracy with actual users, suggesting that our objective evaluation could also be used as a system development tool for pointing gestures in the future.

Since both MoGlow and AMP are systems learned from data, it is interesting to examine how well the models generalize outside training data distribution. We hypothesized that generalization (as measured by pointing accuracy) is correlated with how far away a test pointing position is from training position distribution. We first quantify training position distribution by fitting a multivariate kernel density estimation (KDE) on the training target positions. The bandwidth for KDE is estimated with Scott's rule and the resulting bandwidth is [0.80,0.36,0.35]. The estimated KDE density function

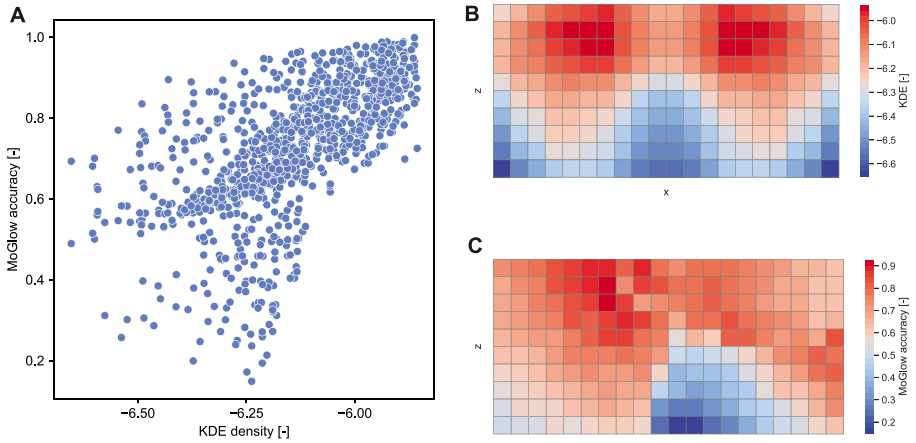


FIGURE 10
Correlation analysis between MoGlow model accuracy and training data density (measured by KDE). **(A)** Shows moderate correlation between the accuracy and the KDE that corresponds to correlation Spearman-r = 0.718 (p< 0.001) as presented in **Table 2**. **(B)** and **(C)** show the KDE grid density and MoGlow accuracy respectively condensed to the x-z axis by averaging over y coordinates.

is evaluated over 1,000 test grid points as depicted on **Figure 8**. The resulting density estimate is then correlated with the corresponding accuracy values for these points in the examined systems; **Table 2** shows the results of the correlation analysis. Only MoGlow shows strong correlation between accuracy and data density. We fully visualized the MoGlow data points in **Figure 10A**, which shows clearly that MoGlow is more accurate in test points with high training data density. We also plotted the KDE in flattened 2D heatmap (**Figure 10B**) and MoGlow's accuracy (**Figure 10C**), and it is shown that the regions with high training data density (red regions in **Figure 10B**) correspond to regions where MoGlow obtained high accuracy (red regions in **Figure 10C**). This result is not surprising since MoGlow is a supervised learning system that highly depends on training data distribution. In other systems, GT-nn, AMP base, AMP-pfnn 3, and AMP-pfnn 8 have small amount of significant correlation between training position density and model accuracy. Curiously, some of these correlations are negative, meaning that the models are more accurate further away from the training position distribution. This could be due to several factors, such as, KDE may not fully represent data density, or that the pointing position sampling mechanism in AMP models helps generalization outside of training position distribution.

4 Discussion

While perceptual test results show a trend of correlation between motion naturalness and referential accuracy, i.e., models with higher naturalness also obtained higher accuracy, it is still not clear what role does motion dynamics play in referential accuracy. We did not show referential target in motion naturalness test in order to not distract naturalness judgement with some models' inaccurate pointing. But we hypothesize that the motion dynamics itself contributes to referential accuracy. A factor of motion dynamics that we already noticed could have contributed to referential accuracy is how long was the pointing phase held and how much the hand moved during that phase, i.e., how stable was the pointing phase. One of the baselines, MoGlow, sometimes has long and unstable pointing phase; it points to one position and slowly moves to a nearby position before retraction. Different users reported both choosing the initial position or the ending position as the perceived referent.

However, one user mentioned that this type of unstable and confusing pointing gesture feels quite natural and human-like, as if the avatar is trying to convey that the pointing is not certain. This example suggests a level of communicative capacity in motion style. That is, pointing gesture style, i.e., pointing at the same position with different dynamics, can convey more information to the interlocutor than just what is the referent. It can also convey, for example, degrees of certainty of the made reference. A quick motion with short but steady pointing phase shows certainty. On the other hand, a slow motion with unsteady/constantly moving pointing phase can convey uncertainty. Previous research in cognitive science has shown that speakers tailor the kinematics of their pointing gestures to the communicative needs of the listener, by modulating the speed and duration of the different sub-components of the pointing gesture Peeters (2015) and gesture kinematics has also been exploited in human robot interaction, e.g., to signal hesitation Moon et al. (2021). Future research could explore ways to model pointing styles and uncover what communicative capacity motion style has in pointing gesture.

In our current study we focused solely on the generation of non-verbal mode for referent localization in the 3D space. However, in most cases pointing gestures appear as part of multi-modal expressions, where the non-verbal mode (gesture) and verbal (language, speech) mode carry complementary information. In

these multi-modal expressions pointing gestures are accompanied by verbal descriptors, which can range from simple demonstratives ("this," "that") Peeters and Özyürek, (2016) to more complex forms of multi-modal referring expressions describing the referent Clark and Bangerter (2004) (e.g., "that green book on the table, next to the lamp"). Furthermore, since the goal of our study was to create accurate and human-like pointing gestures, we used a simple 3D position target representation as input to the network. In order to make the pointing gesture generation more realistic and flexible, image features could be provided to the agent. This raises further questions about efficient visual representation for pointing and establishing eye-body coordination in embodied intelligent agents Yang et al. (2021).

It should also be noted that the AMP systems achieved higher pointing accuracy (**Figure 9**) than the ground-truth training data (**Figure 2**) in terms of the defined pointing reward function **Section 2.2**. On one hand, this is not surprising, since AMP is trained to maximize the pointing reward, as part of its full reward. On the other hand, it raises the question of how well-suited is the designed reward function for the pointing task. Pointing is an interactive communication process, which can be modeled as a referential game. Referential games are a type of signaling game Lewis (1969), where the sender (gesture producer) sends a signal to the listener (observer), who needs to discover the communicative intent (localize referent) from the message. In recent years language games have been used in emergent communication studies, mostly in text and image based agents Lazaridou et al. (2018), but also in embodied agents Bullard et al. (2020). In these, the sender and the listener jointly learn a communication protocol, often through an RL setup. This could also be relevant for our framework, since modeling the observer as an agent could substitute the current reward function based on simple geometric alignment. However, it is important to note that this kind of reward signal in emergent communication setup is very sparse, compared to our continuously defined dense reward function, which could pose further learning difficulties. Furthermore, current studies within emergent communication have largely focused on the theoretical aspects and the learned communication protocols are not interpretable to humans by design, therefore they have limited use in interactive applications. Extending these emergent communication frameworks with our pointing gesture generation framework is a promising direction in learning more general non-verbal communication in embodied interactive agent.

5 Conclusion

In this paper, we presented the results of learning pointing gestures in a physically simulated embodied agent using imitation and reinforcement learning. We conducted perceptual and objective evaluations of our proposed imitation-RL based method, and showed that it can produce highly natural pointing gestures with high referential accuracy. Moreover, we showed that our approach generalizes better in pointing gesture generation than a state-of-the-art supervised gesture synthesis model. We also presented a novel interactive VR-based method to evaluate pointing gestures in situated embodied interaction.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

AD: Data collection, implementations and modeling in AMP, VR experiment design. Contributed to writing. SW: Co-designed/implemented AMP models, MoGlow baseline. Co-designed VR experiment. Contributed to writing. SA Contributed with expert knowledge in data recording and processing, gave feedback to model development and helped out with the VR implementation. He also contributed to the paper writing. JB: Co-PI of the project. Planning and coordination of data collection, strategic discussions on modelling and evaluation experiments, manuscript editing.

Funding

The work is funded by Advanced Adaptive Intelligent Agents project (Digital Futures) and Swedish Research Council, grant no 2018-05409.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2022.1110534/full#supplementary-material

References

Abidi, S., Williams, M., and Johnston, B. (2013). "Human pointing as a robot directive," in 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Tokyo, Japan, 03-06 March 2013, 67–68.

Ahuja, C., Lee, D. W., Ishii, R., and Morency, L.-P. (2020). *No gestures left behind: Learning relationships between spoken language and freeform gestures*. Toronto, Canada: Association for Computational Linguistics, 1884–1895. Proc. EMNLP.

Alexanderson, S., Henter, G. E., Kucherenko, T., and Beskow, J. (2020). Style-controllable speech-driven gesture synthesis using normalising flows. *Comput. Graph. Forum* 39, 487–496. Wiley Online Library. doi:10.1111/cgf.13946

Andrist, S., Gleicher, M., and Mutlu, B. (2017). "Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters," in Proceedings of the 2017 CHI conference on human factors in computing systems, Hamburg, Germany, April 23 - 28, 2023, 2571–2582.

Bailly, G., Raidt, S., and Elisei, F. (2010). Gaze, conversational agents and face-to-face communication. *Speech Commun.* 52, 598–612. doi:10.1016/j.specom.2010.02.015

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychol. Sci.* 15, 415–419. doi:10.1111/j.0956-7976.2004.00694.x

Bühler, K. (1934). *Sprachtheorie*. vol. 2 (Jena Fischer).

Bullard, K., Meier, F., Kiela, D., Pineau, J., and Foerster, J. (2020). *Exploring zero-shot emergent communication in embodied multi-agent populations*. Ithaca, NY: Cornell University. arXiv preprint arXiv:2010.15896.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., et al. (1994). *Animated conversation: Rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents*. Philadelphia, PA: University of Pennsylvania ScholarlyCommons, 413–420. Proc. SIGGRAPH.

Cassell, J., Vilhjálmsson, H. H., and Bickmore, T. (2001). *Beat: The behavior expression animation toolkit*. Cambridge, MA: Massachusetts Institute of Technology, 477–486. Proc. SIGGRAPH.

Chao, F., Wang, Z., Shang, C., Meng, Q., Jiang, M., Zhou, C., et al. (2014). A developmental approach to robotic pointing via human–robot interaction. *Inf. Sci.* 283, 288–303. doi:10.1016/j.ins.2014.03.104

Chen, Y., Li, Q., Kong, D., Kei, Y. L., Zhu, S.-C., Gao, T., et al. (2021). "Yourefit: Embodied reference understanding with language and gesture," in Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10-17 October 2021, 1385–1395.

Clark, H. H., and Bangerter, A. (2004). "Changing ideas about reference," in *Experimental pragmatics* (Berlin, Germany: Springer), 25–49.

Cooperrider, K. (2020). *Fifteen ways of looking at a pointing gesture*. Berlin, Germany: Research Gate.

Deichler, A., Wang, S., Alexanderson, S., and Beskow, J. (2022). "Towards context-aware human-like pointing gestures with rl motion imitation," in Context-Awareness in Human-Robot Interaction: Approaches and Challenges, workshop at 2022 ACM/IEEE International Conference on Human-Robot Interaction.

Diessel, H., and Coventry, K. R. (2020). Demonstratives in spatial language and social interaction: An interdisciplinary review. *Front. Psychol.* 11, 555265. doi:10.3389/fpsyg.2020.555265

Dinh, L., Krueger, D., and Bengio, Y. (2014). *Nice: Non-linear independent components estimation*. Ithaca, NY: Cornell University. arXiv preprint arXiv:1410.8516.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2016). *Density estimation using real nvp*. Ithaca, NY: Cornell University. arXiv preprint arXiv:1605.08803.

Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. (2016). "Benchmarking deep reinforcement learning for continuous control," in International conference on machine learning (PMLR), 1329–1338.

Escontrela, A., Peng, X. B., Yu, W., Zhang, T., Iscen, A., Goldberg, K., et al. (2022). *Adversarial motion priors make good substitutes for complex reward functions*. Ithaca, NY: Cornell University. arXiv preprint arXiv:2203.15103.

Fang, R., Doering, M., and Chai, J. Y. (2015). "Embodied collaborative referring expression generation in situated human-robot interaction," in Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, Portland, OR, USA, 02-05 March 2015, 271–278.

Ferstl, Y., and McDonnell, R. (2018). "Investigating the use of recurrent motion modelling for speech gesture generation," in Proceedings of the 18th International Conference on Intelligent Virtual Agents, NSW, Sydney, Australia, November 5 - 8, 2018, 93–98.

Ferstl, Y., Neff, M., and McDonnell, R. (2020). Adversarial gesture generation with realistic gesture phasing. *Comput. Graph.* 89, 117–130. doi:10.1016/j.cag.2020.04.007

Gao, X., Gao, Q., Gong, R., Lin, K., Thattai, G., and Sukhatme, G. S. (2022). *Dialfred: Dialogue-enabled agents for embodied instruction following*. Ithaca, NY: Cornell University. arXiv preprint arXiv:2202.13330.

Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., and Malik, J. (2019). *Learning individual styles of conversational gesture*. Washington D.C: The Computer Vision Foundation. Proc. CVPR.

Häring, M., Eichberg, J., and André, E. (2012). "Studies on grounding with gaze and pointing gestures in human-robot-interaction," in *International conference on social robotics* (Berlin, Heidelberg: Springer), 378–387.

Hasegawa, D., Kaneko, N., Shirakawa, S., Sakuta, H., and Sumi, K. (2018). *Evaluation of speech-to-gesture generation using bi-directional LSTM network* in IVA '18: International Conference on Intelligent Virtual Agents, Sydney NSW Australia, November 5 - 8, 2018.

Heess, N., Tb, D., Sriram, S., Lemmon, J., Merel, J., Wayne, G., et al. (2017). *Emergence of locomotion behaviours in rich environments*. Ithaca, NY: Cornell University. arXiv preprint arXiv:1707.02286.

Henter, G. E., Alexanderson, S., and Beskow, J. (2020). Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Trans. Graph. (TOG)* 39, 1–14. doi:10.1145/3414685.3417836

Ho, J., and Ermon, S. (2016). "Generative adversarial imitation learning," in *Advances in neural information processing systems 29* (British Columbia, Canada: NeurIPS Proceedings).

Holden, D., Komura, T., and Saito, J. (2017). Phase-functioned neural networks for character control. *ACM Trans. Graph. (TOG)* 36, 1–13. doi:10.1145/3072959.3073663

Holladay, R. M., Dragan, A. D., and Srinivasa, S. S. (2014). "Legible robot pointing," in The 23rd IEEE International Symposium on robot and human interactive communication (IEEE), 217–223.

Huang, C.-M., and Mutlu, B. (2014). "Learning-based modeling of multimodal behaviors for humanlike robots," in Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction, Bielefeld, Germany, 03-06 March 2014, 57–64.

Islam, M. M., Mirzaiee, R. M., Gladstone, A., Green, H. N., and Iqbal, T. (2022). "Caesar: An embodied simulator for generating multimodal referring expression datasets," in Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

Kibrik, A. A. (2011). *Reference in discourse*. Oxford, United Kingdom: Oxford University Press. doi:10.1093/acprof:oso/9780199215805.001.0001

Kingma, D. P., and Dhariwal, P. (2018). "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in neural information processing systems 31* (British Columbia, Canada: NeurIPS Proceedings).

Kita, S. (2003). "Pointing: A foundational building block of human communication," in *Pointing* (London, England: Psychology Press), 9–16.

Kopp, S., and Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Comput. Animat. Virtual Worlds* 15, 39–52. doi:10.1002/cav.6

Korzun, V., Dimov, I., and Zharkov, A. (2020). "The finemotion entry to the genea challenge 2020," in *Proc. GENEA workshop* (Glasgow, UK: Zenodo). doi:10.5281/zenodo.4088609

Kucherenko, T., Hasegawa, D., Kaneko, N., Henter, G. E., and Kjellström, H. (2021a). Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. *Int. J. Human–Computer Interact.* 37, 1300–1316. doi:10.1080/10447318.2021.1883883

Kucherenko, T., Jonell, P., van Waveren, S., Henter, G. E., Alexanderson, S., Leite, I., et al. (2020). "Gesticulator: A framework for semantically-aware speech-driven gesture generation," in Proceedings of the 2020 International Conference on Multimodal Interaction, Virtual Event, Netherlands, October 25 - 29, 2020, 242–250.

Kucherenko, T., Jonell, P., Yoon, Y., Wolfert, P., and Henter, G. E. (2021b). "A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020," in 26th international conference on intelligent user interfaces, TX, College Station, USA, April 14 - 17, 2021. 11–21.

Lazaridou, A., Hermann, K. M., Tuyls, K., and Clark, S. (2018). "Emergence of linguistic communication from referential games with symbolic and pixel input," in *Iclr* (Ithaca, NY: Cornell University).

Lester, J. C., Voerman, J. L., Towns, S. G., and Callaway, C. B. (1999). Deictic believability: Coordinated gesture, locomotion, and speech in lifelike pedagogical agents. *Appl. Artif. Intell.* 13, 383–414. doi:10.1080/08839519 9117324

Lewis, D. K. (1969). *Convention: A philosophical study*. Cambridge, MA, USA: Wiley-Blackwell.

Liu, Y., Mohammadi, G., Song, Y., and Johal, W. (2021). "Speech-based gesture generation for robots and embodied agents: A scoping review," in Proceedings of the 9th International Conference on Human-Agent Interaction, Virtual Event, Japan, November 9 - 11, 2021, 31–38.

Lücking, A., Pfeiffer, T., and Rieser, H. (2015). Pointing and reference reconsidered. *J. Pragmat.* 77, 56–79. doi:10.1016/j.pragma.2014.12.013

Luketina, J., Nardelli, N., Farquhar, G., Foerster, J., Andreas, J., Grefenstette, E., et al. (2019). *A survey of reinforcement learning informed by natural language*. Ithaca, NY: Cornell University. arXiv preprint arXiv:1906.03926.

Lynch, C., Wahid, A., Tompson, J., Ding, T., Betker, J., Baruch, R., et al. (2022). *Interactive language: Talking to robots in real time*. Ithaca, NY: Cornell University. arXiv preprint arXiv:2210.06407.

Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., et al. (2021). *Isaac gym: High performance gpu-based physics simulation for robot learning*. Ithaca, NY: Cornell University. arXiv preprint arXiv:2108.10470.

Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., and Shapiro, A. (2013). *Virtual character performance from speech*. Eindhoven: Eurographics Association, 25–35. Proc. SCA.

Matuszek, C., Bo, L., Zettlemoyer, L., and Fox, D. (2014). Learning from unscripted deictic gesture and language for human-robot interactions. *Proc. AAAI Conf. Artif. Intell.* 28 (1), 9051. doi:10.1609/aaai.v28i1.9051

McNeill, D. (1992). *Hand and mind 1*. Chicago: University of chicago press. Advances in Visual Semiotics 351.

Merel, J., Tassa, Y., Tb, D., Srinivasan, S., Lemmon, J., Wang, Z., et al. (2017). *Learning human behaviors from motion capture by adversarial imitation*. Ithaca, NY: Cornell University. arXiv preprint arXiv:1707.02201.

Moon, A., Hashmi, M., Loos, H. M. V. D., Croft, E. A., and Billard, A. (2021). Design of hesitation gestures for nonverbal human-robot negotiation of conflicts. *ACM Trans. Human-Robot Interact. (THRI)* 10, 1–25. doi:10.1145/3418302

Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H., and Hagita, N. (2009). "Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior," in Proceedings of the 4th ACM/IEEE international conference on Human robot interaction, La Jolla, CA, USA, 11-13 March 2009, 69–76.

Ng-Thow-Hing, V., Luo, P., and Okita, S. (2010). "Synchronized gesture and speech production for humanoid robots," in 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18-22 October 2010.

Noma, T., Zhao, L., and Badler, N. I. (2000). Design of a virtual human presenter. *IEEE Comput. Graph. Appl.* 20, 79–85. doi:10.1109/38.851755

Padmakumar, A., Thomason, J., Shrivastava, A., Lange, P., Narayan-Chen, A., Gella, S., et al. (2022). Teach: Task-driven embodied agents that chat. *Proc. AAAI Conf. Artif. Intell.* 36, 2017–2025. doi:10.1609/aaai.v36i2.20097

Peeters, D. (2015). *A social and neurobiological approach to pointing in speech and gesture*. Nijmegen, Netherlands: Radboud University Nijmegen. Ph.D. thesis.

Peeters, D., Krahmer, E., and Maes, A. (2021). A conceptual framework for the study of demonstrative reference. *Psychonomic Bull. Rev.* 28, 409–433. doi:10.3758/s13423-020-01822-8

Peeters, D., and Özyürek, A. (2016). This and that revisited: A social and multimodal approach to spatial demonstratives. *Front. Psychol.* 7, 222. doi:10.3389/fpsyg.2016.00222

Peng, X. B., Abbeel, P., Levine, S., and Van de Panne, M. (2018). Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Trans. Graph. (TOG)* 37, 1–14. doi:10.1145/3197517.3201311

Peng, X. B., Ma, Z., Abbeel, P., Levine, S., and Kanazawa, A. (2021). Amp: Adversarial motion priors for stylized physics-based character control. *ACM Trans. Graph. (TOG)* 40, 1–20. doi:10.1145/3476576.3476723

Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W. Y., Shen, C., et al. (2020). "Reverie: Remote embodied visual referring expression in real indoor environments," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13-19 June 2020, 9982–9991.

Rickel, J., and Johnson, W. L. (1999). Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Appl. Artif. Intell.* 13, 343–382. doi:10.1080/08839519911 7315

Sauppé, A., and Mutlu, B. (2014). "Robot deictics: How gesture and context shape referential communication," in 2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Bielefeld, Germany, 03-06 March 2014, 342–349.

Schauerte, B., and Fink, G. A. (2010). "Focusing computational visual attention in multi-modal human-robot interaction," in International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction, Beijing, China, November 8 - 10, 2010, 1–8.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). *Proximal policy optimization algorithms*. Ithaca, NY: Cornell University. arXiv preprint arXiv:1707.06347.

Shukla, D., Erkent, O., and Piater, J. (2015). "Probabilistic detection of pointing directions for human-robot interaction," in 2015 international conference on digital image computing: techniques and applications (DICTA, Adelaide, SA, Australia, 23-25 November. 1–8.

Steels, L. (2001). Language games for autonomous robots. *IEEE Intell. Syst.* 16, 16–22. doi:10.1109/mis.2001.956077

Stepputtis, S., Campbell, J., Phielipp, M., Baral, C., and Amor, H. B. (2019). *Imitation learning of robot policies by combining language, vision and demonstration*. Ithaca, NY: Cornell University. arXiv preprint arXiv:1911.11744.

Sugiyama, O., Kanda, T., Imai, M., Ishiguro, H., and Hagita, N. (2007). "Natural deictic communication with humanoid robots," in 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, CA, USA, 29 October 2007 - 02 November 2007, 1441–1448.

Talmy, L. (2020). Targeting in language: Unifying deixis and anaphora. *Front. Psychol.* 11, 2016. doi:10.3389/fpsyg.2020.02016

Team, D. I. A., Abramson, J., Ahuja, A., Brussee, A., Carnevale, F., Cassin, M., et al. (2021). *Creating multimodal interactive agents with imitation and self-supervised learning*. Ithaca, NY: Cornell University. arXiv preprint arXiv:2112. 03763.

Valle-Pérez, G., Henter, G. E., Beskow, J., Holzapfel, A., Oudeyer, P.-Y., and Alexanderson, S. (2021). Transflower: Probabilistic autoregressive dance generation with multimodal attention. *ACM Trans. Graph. (TOG)* 40, 1–14. doi:10.1145/3478513.3480570

Vollenweider, E., Bjelonic, M., Klemm, V., Rudin, N., Lee, J., and Hutter, M. (2022). *Advanced skills through multiple adversarial motion priors in reinforcement learning*. Ithaca, NY: Cornell University. arXiv preprint arXiv:2203.14912.

Wu, Q., Wu, C.-J., Zhu, Y., and Joo, J. (2021). "Communicative learning with natural gestures for embodied navigation agents with human-in-the-scene," in 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech, 27 September 2021 - 01 October 2021, 4095–4102.

Yang, R., Zhang, M., Hansen, N., Xu, H., and Wang, X. (2021). "Learning vision-guided quadrupedal locomotion end-to-end with cross-modal transformers," in International Conference on Learning Representations.

Yoon, Y., Cha, B., Lee, J.-H., Jang, M., Lee, J., Kim, J., et al. (2020). Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Trans. Graph.* 39, 1–16. doi:10.1145/3414685.3417838

Zabala, U., Rodriguez, I., and Lazkano, E. (2022). "Towards an automatic generation of natural gestures for a storyteller robot," in In 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Napoli, Italy, 29 August 2022 - 02 September 2022, 1209–1215.

Paper C

Paper C

Incorporating Spatial Awareness in Data-Driven Gesture Generation for Virtual Agents

ANNA DEICHLER, SIMON ALEXANDERSON, AND JONAS BESKOW

Published in
Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (IVA '24)

Incorporating Spatial Awareness in Data-Driven Gesture Generation for Virtual Agents

Anna Deichler*
deichler@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Simon Alexanderson
simonal@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Jonas Beskow
beskow@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

arXiv:2408.04127v1 [cs.HC] 7 Aug 2024

ABSTRACT

This paper focuses on enhancing human-agent communication by integrating spatial context into virtual agents' non-verbal behaviors, specifically gestures. Recent advances in co-speech gesture generation have primarily utilized data-driven methods, which create natural motion but limit the scope of gestures to those performed in a void. Our work aims to extend these methods by enabling generative models to incorporate scene information into speech-driven gesture synthesis. We introduce a novel synthetic gesture dataset tailored for this purpose. This development represents a critical step toward creating embodied conversational agents that interact more naturally with their environment and users.

CCS CONCEPTS

• **Computing methodologies** → **Animation**; **Spatial and physical reasoning**.

KEYWORDS

Gesture generation, Situated virtual agents, Deictic gestures, Co-speech gesture, Synthetic data

1 INTRODUCTION

Human communication is rooted in a spatial and physical context. In situated communicative settings people engage in referential communication in identifying, describing, or giving instructions related to objects, locations, or people. They make use of multimodal expressions, consisting of spatial language and non-verbal behaviors, such as gaze and pointing gestures. For a virtual agent to function effectively in a spatially contextual environment, it must be capable of interpreting and responding to both verbal and environmental cues. This involves understanding objects, locations, and other contextual factors that influence how gestures should be formed and directed. For instance, pointing at an object while describing it, or making eye contact with an interlocutor during a conversation, requires an integration of scene understanding with gesture generation. Non-verbal behavior generation in virtual agents has a long history. Early work in gesture synthesis for virtual agents was rule-based, based on inverse kinematics and procedural animation to generate co-speech gestures [5, 15, 20]. Rule-based approaches, however, often require laborious manual tuning and often struggle to achieve a natural motion quality. In recent years data-driven solutions have become primary and supervised learning systems for co-speech gesture generation and have achieved high naturalness in certain settings. Most of these systems take audio as input [3, 10, 12], and are limited to generating beat gestures aligned with speech. Some recent work extends audio conditioning with text conditioning to obtain semantic information [6, 16]. On the one hand, this made the generation of highly natural motion possible, but on the other hand, it also narrowed the scope of these behaviors. In particular, there has been a surge in co-speech gesture generation research due to novel co-speech gesture datasets consisting of speech and motion modalities. These co-speech generation systems focus on generating gestures that lack spatial information. However, for any kind of situated interaction, these agents need exposure to the spatial context. In our work, we aim to develop generative models that extend existing speech-driven gesture synthesis frameworks by incorporating scene information. In this paper, we propose a novel synthetic gesture dataset that augments an existing co-speech dataset with multimodal referring expressions, as a first step towards extending speech-driven gesture synthesis with scene information conditioning.

2 RELATED WORK

2.1 Situated Gesture generation

Generation of situated gestures, such as deictic expressions, has been considered in virtual agents for a long time [14, 18, 22, 23]. Similar approaches have been employed in referential human-robot communication [9, 24, 25]. All of these systems are based on rules and/or procedural generation, which limits the degree of naturalness, fluidity, and variability that can be achieved in the resulting animations, and makes it difficult to match the gesture quality produced by contemporary generative models for co-speech gestures as described above. Recent work started considering the data-driven generation of pointing gestures [8].

2.2 Scene conditioning in human motion generation

In recent years there has been more focus on scene-conditioned human motion generation [13, 26, 27]. These works in human-scene interaction focus mainly on non-conversational human locomotion generation, ranging from generating static human poses in 3D environments [29] to temporally extended human motion synthesis, conditioning on the 3D layout of the environment and human-object interaction [28]. Conditional motion generation works include diffusion-based [13] and reinforcement learning-based methods.

Currently gesture generation in virtual agents and scene-conditioned human motion generation are separate fields, but in order to have situated gesturing in virtual agents there should be more interaction between these fields.

3 DATASET GENERATION

In order to train conversational embodied agents that are situated in a spatial context, datasets are needed that contain the relevant environment information. Currently, there is a lack of such datasets. One way to acquire such a dataset is to extend and combine existing datasets that contain the necessary annotations in themselves. In recent years, generating synthetic data has offered a solution to machine learning problems, where data is scarce and difficult to access. It has become common practice in the domain of text generation[19], and there are also works in motion generation [4],[21]. Similar strategies could be applied for generating synthetic datasets that could help extend the current capabilities of co-speech gesture generation agents.

In the realm of deictic expressions, which involve pointing and other indicating gestures, there is a critical need for systems that not only recognize speech and text but also incorporate the physical environment. This integration allows virtual agents to perform gestures that are not only relevant to the spoken content but are also spatially aware, enhancing their utility in real-world interactions. Such advancements are vital for developing more sophisticated and interactive virtual agents that can operate effectively within dynamic environments.

As a step towards exposing current co-speech gesture agents to a spatial context, we combined an existing co-speech gesture dataset, which mainly contains beat gestures, with an existing pointing gesture dataset. The co-speech gesture dataset we used[11] is one of the largest natural conversation datasets of synchronized motion capture and speech recordings with a duration of approximately 6 hours. The dataset contains recordings of a single English speaker. Synthesizing human-like pointing gestures in virtual avatars requires an appropriate full-body motion capture dataset with diverse and accurate target locations covering the 3D space surrounding the character. We utilized the solely available dataset that includes both motion capture data of pointing gestures and the corresponding 3D target locations [7]. This dataset also includes recorded speech as the actor is pointing. Example utterances in single target pointing are simple demonstrative expressions, such as "This one!", "I want that one!", "Give me that one over there!". The speech data is synchronized with mocap.

We replaced these simple demonstratives with synthesized ones to better match the other co-speech gesture dataset's acoustic and semantic qualities.

Extending the pointing gesture dataset. The original pointing dataset contained approximately 140 isolated single target pointing gestures, ranging from 0.5 to 9 seconds in duration. To account for the small size of this dataset, all motion files were mirrored and each of the individual pointing gesture clips was stretched in time in the range of [0.7, 1.6] three times. This resulted in a pointing gesture dataset of 1160 clips. The distribution of pointing gesture lengths can be seen on Figure 1. The mean duration of the pointing gesture was 4.85 seconds.

Text generation. We used GPT-4 [2] to generate simple demonstrative references for the pointing gesture dataset. LLMs using Generative Pre-trained transformers (GPT) can be utilized to generate large amounts of task-specific text with prompts appropriate for the given task. To have semantic similarity between the beat gesture and pointing gesture datasets, we provided the transcripts of the beat gesture dataset in prompting the LLM as contextual information. Furthermore, each utterance should be an exophoric referring expression that could be coupled with a pointing gesture. The main instruction prompt can be found in Appendix A.

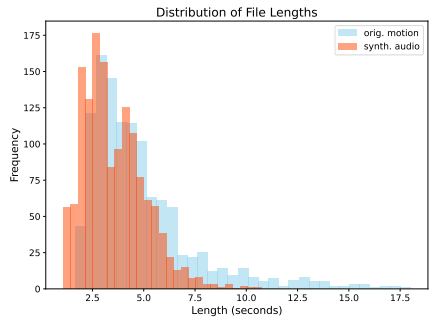


Figure 1: Distribution of length of pointing motion clips and synthesized audio clips.

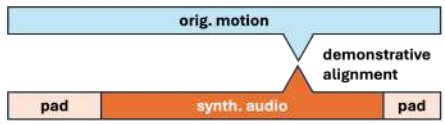


Figure 2: Alignment of demonstratives in motion and speech, with pre-and post-padding of audio clips.

Speech synthesis. Based on the generated text, speech segments were synthesized using a commercially available TTS engine [1]. Emphasis on demonstrative words in the synthesized utterances text was added using the Speech Synthesis Markup Language (SSML) phoneme tags. In total, 1400 sentences were synthesized.

```
<speak><emphasis level=strong>THIS</emphasis>
one, towering above the rest, is the oldest
tree in the forest.</speak>
```

```
<speak>On this shelf, can you spot <emphasis
level='strong'> THAT</emphasis> vase there?
It's handmade.</speak>
```

Matching the generated speech segments with pointing gestures. The generated speech segments were matched with the pointing gestures using the Hungarian algorithm.

Figure 1 compares the length of the synthesized audio clips against the length of the pointing gestures.

The cost function considered the difference in length between synthetic and pointing segments, as well as the different in alignment between demonstratives, i.e. penalizing matches where the synthetic event does not temporally align with the original demonstrative event (e.g., starting before or ending after the demonstrative). The aim was to maintain consistency with the original data's

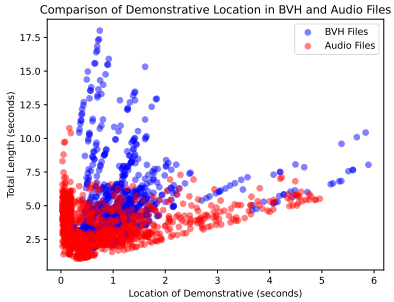


Figure 3: Demonstrative location against total length (in seconds) for motion (BVH) and audio files.

temporal distribution of demonstrative within the utterance as seen on Figure 3.

As can be seen from Figures 1 and 3, there is a discrepancy between animation and audio lengths. This is intentional since we wanted the speech segments to be fully contained within the animation clips after the demonstratives are aligned in the two streams. The speech segments were subsequently padded with leading and trailing silence, to the same length as the animations, see Figure 2.

Figure 4 shows a visualization of a pointing gesture and a beat gesture from the dataset. Further examples can be found on the demo website. [1].

Figure 4: Visualization of dataset examples containing synchronized clips of audio and gesture for (a) pointing and (b) beat gestures.

4 CONCLUSIONS AND FUTURE WORK

In this paper, we highlighted one of the major shortcomings of current data-driven gesture generation systems that the gestures are generated completely in isolation, without any spatial context that would make the gestures situated. We provided a synthetic dataset, which could serve as a benchmark to test computational models that are capable of joint co-speech and referential gesture generation. These simple, synthetic datasets are beneficial to extend the capabilities of current co-speech gesture generation models, but in the future, more datasets are needed that have contextual information, for example, spatial context. These datasets could include annotations about the scene, for example, details about the objects and spatial configurations present in the environment during each recorded interaction. Processing this enriched dataset will require sophisticated data handling strategies to ensure that the model can generalize from training data to real-world applications.

Another important consideration is how to properly evaluate situated gestures in a virtual agent. We propose to build such evaluation methods on previous gesture generation evaluations. For example, referential games can be utilized for pointing accuracy, as in [8]. For gesture naturalness and appropriateness, we can refer to the methodologies used in the GENEA challenge [17].

Going forward, we believe that generating synthetic datasets or relying on multiple separate datasets in training will be necessary to achieve agents capable naturalistic of situated gesture and this dataset provides a first benchmark to develop methods that integrate spatial awareness into the natural gesture generation capabilities of virtual agents.

[1]https://huggingface.co/spaces/annadeichler/spatial-gesture

REFERENCES

[1] [n. d.]. ElevenLabs ElevenLabs. https://elevenlabs.io/. Accessed: 2024-04-01.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 487–496.

[4] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. 2023. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8726–8737.

[5] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 477–486.

[6] Anna Deichler, Shivam Mehta, Simon Alexanderson, and Jonas Beskow. 2023. Diffusion-based co-speech gesture generation using joint text and audio representation. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 755–762.

[7] Anna Deichler, Siyang Wang, Simon Alexanderson, and Jonas Beskow. 2022. Towards context-aware human-like pointing gestures with rl motion imitation. In *Context-Awareness in Human-Robot Interaction: Approaches and Challenges, workshop at 2022 ACM/IEEE International Conference on Human-Robot Interaction*.

[8] Anna Deichler, Siyang Wang, Simon Alexanderson, and Jonas Beskow. 2023. Learning to generate pointing gestures in situated embodied conversational agents. *Frontiers in Robotics and AI* 10 (2023), 1110534.

[9] Rui Fang, Malcolm Doering, and Joyce Y Chai. 2015. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. 271–278.

[10] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2020. Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics* 89 (2020), 117–130.

[11] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2021. ExpressGesture: Expressive gesture generation from speech through database matching. *Computer Animation and Virtual Worlds* (2021), e2016.

[12] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 79–86.

[13] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. 2023. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16750–16761.

[14] Sri Rama Kartheek Kappagantula, Nicoletta Adamo-Villani, Meng-Lin Wu, and Voicu Popescu. 2019. Automatic deictic gestures for animated pedagogical agents. *IEEE Transactions on Learning Technologies* 13, 1 (2019), 1–13.

[15] Stefan Kopp and Ipke Wachsmuth. 2004. Synthesizing multimodal utterances for conversational agents. *Computer animation and virtual worlds* 15, 1 (2004), 39–52.

[16] Taras Kucherenko, Patrik Jonell, Sanne Van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 international conference on multimodal interaction*. 242–250.

[17] Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 792–801.

[18] James C Lester, Jennifer L Voerman, Stuart G Towns, and Charles B Callaway. 1999. Deictic believability: Coordinated gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence* 13, 4-5 (1999), 383–414.

[19] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024. Best Practices and Lessons Learned on Synthetic Data for Language Models. *arXiv preprint arXiv:2404.07503* (2024).

[20] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics symposium on computer animation*. 25–35.

[21] Shivam Mehta, Anna Deichler, Jim O'regan, Birger Moëll, Jonas Beskow, Gustav Eje Henter, and Simon Alexanderson. 2024. Fake it to make it: Using synthetic data to remedy the data shortage in joint multimodal speech-and-gesture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1952–1964.

[22] Tsukasa Noma, Liwei Zhao, and Norman I Badler. 2000. Design of a virtual human presenter. *IEEE Computer Graphics and Applications* 20, 4 (2000), 79–85.

[23] Jeff Rickel and W Lewis Johnson. 1999. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied artificial intelligence* 13, 4-5 (1999), 343–382.

[24] Elena Rubies, Ricard Bitriá, Eduard Clotet, and Jordi Palacín. 2023. Remote Control Device to Drive the Arm Gestures of an Assistant Humanoid Robot. *Applied Sciences* 13, 19 (2023), 11115.

[25] Allison Sauppé and Bilge Mutlu. 2014. Robot deictics: How gesture and context shape referential communication. In *Proceedings of 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 342–349.

[26] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. 2024. Move as You Say Interact as You Can: Language-guided Human Motion Generation with Scene Affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 433–444.

[27] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. 2022. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems* 35 (2022), 14959–14971.

[28] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez-Pellitero, and Gerard Pons-Moll. 2024. FORCE: Dataset and Method for Intuitive Physics Guided Human-object Interaction. *arXiv preprint arXiv:2403.11237* (2024).

[29] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. 2020. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6194–6204.

A GPT-4 PROMPTS

The demonstrative references were generated using this prompt:

> Demonstratives in exophoric reference are used to indicate concrete physical entities in space. For example, I was thinking of sitting down. See that bench there, we could sit there. Here 'that' would be the demonstrative. I want you to generate examples like the following, which connect to a story. The rule is that each generated sentence should contain one exophoric reference. Also, put the demonstrative in the middle of the sentence. Examples: " See that bench over THERE, we could sit there", "Look in the middle of the room, THAT small dog is my new best friend.", "We could get there faster. See THAT bike over there.", "Look at the top shelf. THAT is the vase I wanted to buy".

After this, summaries of the [11] dataset transcripts were provided as additional context. Furthermore, the location of the demonstrations was also directly prompted to cover better the original pointing dataset's demonstrative location in the accompanying pointing gesture motion sequence.

Paper D

Paper D

Grounded Gesture Generation: Language, Motion, and Space

ANNA DEICHLER, JIM O'REGAN, TÉO GUICHOUX, DAVID JOHANSSON, AND JONAS BESKOW

Published in
Humanoid Agents Workshop, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025

arXiv:2507.04522v1 [cs.CV] 6 Jul 2025

Grounded Gesture Generation: Language, Motion, and Space

Anna Deichler[1] Jim O'Regan[1] Téo Guichoux[2] David Johansson[1] Jonas Beskow[1]
[1]KTH Royal Institute of Technology
[2]Sorbonne University

Abstract

Human motion generation has advanced rapidly in recent years, yet the critical problem of creating spatially grounded, context-aware gestures has been largely overlooked. Existing models typically specialize either in descriptive motion generation, such as locomotion and object interaction, or in isolated co-speech gesture synthesis aligned with utterance semantics. However, both lines of work often treat motion and environmental grounding separately, limiting advances toward embodied, communicative agents. To address this gap, our work introduces a multimodal dataset and framework for grounded gesture generation, combining two key resources: (1) a synthetic dataset of spatially grounded referential gestures, and (2) MM-Conv, a VR-based dataset capturing two-party dialogues. Together, they provide over 7.7 hours of synchronized motion, speech, and 3D scene information, standardized in the HumanML3D format. Our framework further connects to a physics-based simulator, enabling synthetic data generation and situated evaluation. By bridging gesture modeling and spatial grounding, our contribution establishes a foundation for advancing research in situated gesture generation and grounded multimodal interaction. Project page: https://groundedgestures.github.io/.

1. Introduction

Situated referential communication, where speech, gesture, and gaze are coordinated to ground meaning in the surrounding environment, is a core aspect of human interaction. It is essential not only for resolving ambiguities and clearly expressing communicative intent in everyday conversation, but also for pedagogical instruction, language development, and overcoming language barriers. In recent years, there has been a surge in technologies related to the generation of humanoid agents with believable and communicative behaviors. However, the generation of *spatially grounded, context-aware gestures* for these agents remains a significant challenge.

To date, research on motion generation using deep generative models has largely focused on two separate problem domains. One is co-speech gesture generation, which aims to generate believable and semantically plausible gestures that match a verbal message in speech and/or text. These models are trained on large corpora of motion capture or video, conditioned with speech and/or text with speech, typically with no spatial information other than that of the interlocutor (if present). The other class of models are so-called "text-to-motion" models, where natural language prompts are used to drive human behavior. These models are trained on large and diverse datasets of human motion, and can also in some cases incorporate semantic and geometric informa-

tion about the scene, allowing instructions like "walk over to the sofa and sit down", but these models have no concept of speech and gesture. In order to build humanoid agents that can function in real life scenarios, we need models that can integrate spatial information and communicative behavior, including speech and gesture. Progress is constrained by the lack of standardized datasets that combine motion, language, gaze, and 3D scene information in a way that supports both training and evaluation of situated behaviors. Existing datasets often treat gesture generation and environmental grounding separately, limiting advances in embodied, interactive AI systems.

One issue that hinders efforts trying to bridge these domains is discrepancy in terms of motion representation. For text-to-motion models, HumanML3D or SMPL-X are the dominant formats. In co-speech gesture, the tradition is to represent motion as skeletal joint angles (e.g. BVH-format). Recently [21] introduced SMPL-X representation for one co-speech gesture dataset, but most of the available co-speech data is still in BVH format which hinders integration tasks.

In this work, we address this gap by introducing a multimodal dataset and framework for situated referential gesture research. We combine two complementary resources, a synthetic dataset of spatially grounded referential gestures and MM-Conv, a VR-based dataset of spontaneous two-party dialogues, converted into HumanML3D-compatible format. Beyond dataset integration, we connect the data to a physics-based simulator, enabling synthetic data generation, situated evaluation, and future extensions toward embodied agents. As a proof-of-concept, we also fine-tune a motion generation model (OmniControl [31]) on the data to produce spatially controllable pointing gestures.

Our main contributions are:

- **Standardization:** Integration of synthetic and real referential gesture datasets into the HumanML3D format, aligning with emerging standards in human-scene modeling.
- **Framework:** A proposed modular architecture that integrates spatial cues into gesture generation, supporting both synthetic data generation and situated evaluation in dynamic 3D environments.
- **Resources:** Release of over 7.7 hours of synchronized motion, speech, and scene data, supporting the study of grounded communication in immersive settings.

2. Related Work

Research at the intersection of gesture, language, and 3D scene understanding has advanced significantly in recent years, but grounded referential gesture generation remains underexplored. Most progress has occurred in two relatively separate domains: (1) language-conditioned motion generation, which focuses on producing motion from textual instructions, and (2) co-speech gesture generation, which aims to synchronize gestures with speech for naturalistic communication. In this section, we review both domains and highlight their limitations for building communicative, scene-aware humanoid agents.

A structured comparison of co-speech and motion-language datasets, including modality coverage, representation format, and communicative focus, is provided in Table 1.

2.1. Language-conditioned human motion generation

The AMASS (Archive of Motion Capture as Surface Shapes) [24] dataset is a comprehensive collection of human motion data, integrating 15 different optical marker-based motion capture datasets into a unified framework. It provides realistic 3D human meshes represented by a rigged body model, making it valuable for various applications in animation, visualization, and deep learning. It contains over 11k motion sequences, 40 hours of data, more than 300 individuals.

HumanML3D [13] is derived from AMASS and HumanAct12 dataset and is designed to bridge natural language and 3D human motions, facilitating tasks like text-to-motion generation and motion captioning. It contains over 14k motion sequences, approximately 28.6 hours of data. The average motion length is 7.1 seconds, while average description length is 12 words. The minimum and maximum duration are 2s and 10s respectively. In terms of the textual descriptions, their average and median lengths are 12 and 10, respectively.

HUMANISE [30] introduces a large-scale synthetic dataset for language-conditioned human motion generation in 3D scenes. It aligns 19,600 AMASS motion sequences with richly annotated ScanNet 3D scans and auto-generated language descriptions that specify both actions and interacting objects (e.g., "sit on the armchair near the desk"). Spanning 643 indoor scenes, the dataset enables research on spatial reasoning between motion, object geometry, and language for embodied AI.

Recent advances in human motion generation increasingly rely on diffusion models, which have demonstrated strong capabilities in producing diverse, high-quality motion sequences. Language-conditioned models such as those used in HUMANISE, as well as scene-aware and language-guided frameworks like TeSMo [34], now commonly adopt diffusion-based architectures. This reflects the growing dominance of diffusion models for generating realistic and contextually grounded human motions.

Table 1. Comparison of human motion and gesture datasets highlighting modalities, text types, settings, and scales.

Dataset	Modalities	Text Type	Motion Representation	Duration	Focus
Grounded Gestures (Ours)	Motion, Speech, Scene Info	Referential utterances	SMPL-X	7.7 h	Referential/ conversational gestures
MM-Conv [7]	Motion, Speech, Gaze, Scene Graphs	Spontaneous dialogue	BVH	6.5 h	Spontaneous referential communication
HumanML3D [12]	Text, 3D motion	Descriptive captions	SMPL+H/DMPL	28.59h	Text-conditioned general motion
HUMANISE [30]	Text, 3D motion	Descriptive captions	SMPL+H/DMPL	28.59h	Text-conditioned general motion
BEAT [22]	Text, Speech, Motion	Transcripts	BVH	76h	Co-speech beat and iconic gestures
BEAT2 [21]	Text, Speech, Motion	Transcripts	SMPL-X	60h	Co-speech beat and iconic gestures
TalkSHOW [33]	Text, Speech, Motion	Transcripts	SMPL-X	27h	Video extracted Co-speech gestures
Trinity Speech Gesture (I [10] and II [11])	Text, Speech, Motion	Transcripts	BVH	6h + 4h	Co-speech gestures
KIT ML [27]	Text, MoCap	Action labels	SMPL [13]	11.23h	Action-mapped motions

2.2. Co-Speech Gesture Generation

Co-speech gesture generation focuses on synthesizing naturalistic hand and body movements that accompany spoken language. Early approaches often relied on rules or heuristics in combination with procedural animation, and could support different types of co-speech gestures, including pointing to specific spatial locations [14,17,25,28]. The motion produced by these systems often had an unmistakable synthetic quality. With the availability of large-scale datasets of synchronized speech and motion data, data-driven methods have become the dominant paradigm. These datasets fostered the development of models that achieve high naturalness and improved speech-gesture synchronization, however, at the cost of lacking control and spatial awareness.

The Trinity Speech-Gesture Dataset [9] contains 23 recordings totaling 244 minutes of synchronized full-body motion capture and audio of a male native English speaker delivering spontaneous monologues on various topics. The Talking with Hands16.2M dataset [16] provides approximately 50 hours of synchronized full-body and detailed hand motion, recorded during natural dyadic conversations in a motion capture laboratory setting. The BEAT dataset [22] offers large-scale multimodal data combining speech, text, and full-body motion. The BEAT2 [21] dataset combines SMPL-X body estimated using MoSh [23] with head parameters estimated using FLAME [19].

Diffusion models have also shown strong performance in gesture generation. For example, GestureDiffuClip [1] related approaches use diffusion processes to model complex motion distributions conditioned on speech (audio and text) embeddings. Other works have incorporated contrastive learning strategies to better align motion and speech representations [6].

While these methods have led to substantial improvements in the realism and timing of gestures relative to speech, most focus on isolated gesture generation without grounding gestures in a shared spatial or interactional context. As a result, they often neglect the spatial communicative functions of gestures essential for situated communication.

3. Dataset

We aim to standardize situated gesture generation datasets. To this end, we build on two separate motion-capture-based datasets, which contain conversational and pointing gestures. Together, these datasets enable the development of models that combine co-speech and deictic gestures in situated settings.

- The **synthetic pointing gesture dataset** consists of isolated pointing gestures, with accurate 3D target locations captured via marker-based motion capture.
- The **VR-based referential dataset (MM-Conv)** contains natural conversations recorded in virtual reality environments, using motion capture, where participants interact with shared virtual spaces.

3.1. Motion data preparation

We use [32] to transform the global joint position data from the motion capture system into SMPL-X [26] format. Then we follow the steps described in [13] to transform the SMPL-X data format into the HumanML3D format.

In HumanML3D, each body pose is parameterized by a vector $\mathbf{x} \in \mathbb{R}^{263}$, whose entries are

$$\underbrace{\omega_r}_{1}, \underbrace{\mathbf{v}_r}_{2}, \underbrace{y_r}_{1}, \underbrace{p}_{21\times3}, \underbrace{R}_{21\times6}, \underbrace{\dot{R}}_{22\times3}, \underbrace{c}_{4}$$

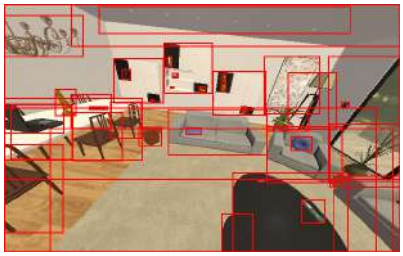


Figure 1. Ground truth bounding boxes of objects in current view, extracted from simulator [15].

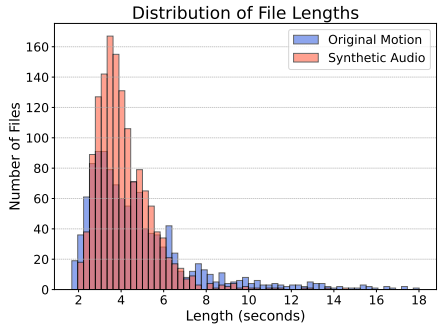


Figure 2. Distribution of file lengths for synthetic audio and original motion segments.

corresponding respectively to root rotational velocity, root linear velocity, root vertical position, rotation-invariant joint positions, joint rotation parameters, joint rotational velocities, and binary foot-contact indicators, with total dimensionality $\dim(\mathbf{x}) = 1{+}2{+}1{+}21{\cdot}3{+}21{\cdot}6{+}22{\cdot}3{+}4 = 263$. Pelvic translation is encoded as the displacement from the preceding frame, while all other joints are expressed relative to the pelvis; this representation facilitates learning and leads to more realistic synthesized motions [31]. The motion sequences are sampled at 20 frames per second.

3.2. Synthetic pointing gesture data generation

The paper [5] proposes integrating spatial context into virtual agents' gesture generation by combining co-speech and pointing gestures. A synthetic dataset was created to extend current speech-driven models with spatially grounded behaviors, enabling agents to perform deictic gestures aligned with speech. We follow the main generation steps outlined in the paper to create semantically aligned data with the VR-based referential dataset [7]:

1. Take the augmented pointing dataset from [8].
2. Generate new demonstrative utterances using a VLM.
3. Synthesize utterances with TTS.
4. Temporally align the synthesized utterances with the pointing motions using the Hungarian algorithm, including padding to match durations.

We transform the pointing dataset BVH files using the process described in section 3.1. This results in 1135 motion files with average duration of 4.85 seconds.

We use GroundingGPT [20] to adjust the prompt generation for the MM-Conv dataset. Specifically, we rendered an RGB image from the character's viewpoint and extracted the ground truth segmentation mask with object labels from the simulator (Figure 1). We randomly selected one object and passed it to the VLM using the prompt in Appendix 6.1. We generate exophoric referring expressions involving demonstratives ('this' or 'that') in a conversational style. Examples of generated outputs:

- *See that chair over there, it's the perfect spot for a relaxing read.*
- *Look at that painting on the wall over there, it's really beautiful.*

We used [4] to generate natural-sounding synthetic speech from the generated sentences, using voice cloning of the original actors. The resulting speech samples are then processed using WhisperX [3] to obtain word-level timestamps. The output of WhisperX is additionally compared to the original text using a Word Error Rate threshold of 0.3, to filter for major divergences in generation, discarding any samples that deviate too much from the original sentence. 94 samples were removed by this step. The remaining 1406 files were matched with the 1135 pointing gesture motions using the Hungarian algorithm based on the demonstrative location. Figure 2 shows the distribution of durations for the synthesized audio clips and the original pointing gesture segments.

Examples of aligned, synthesized pointing gestures are provided in the supplementary material. Additional illustrative examples can be found in Appendix 6.2.

3.3. VR-based referential dataset

The MM-Conv dataset [7] is a multimodal conversation dataset designed for training speech-driven gesture generation models. It contains motion capture recordings of human conversations conducted in virtual reality (VR) environments using the Ai2-Thor simulator [15]. Participants interact naturally, referring to shared virtual environments that include objects and spatial landmarks. The dataset includes synchronized speech audio, text transcripts, and full-body motion capture data encompassing hands, head, and

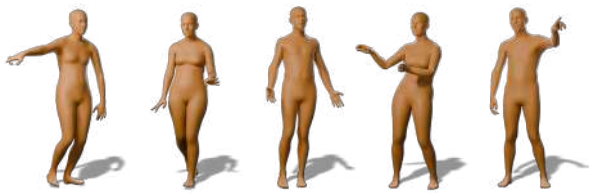


Figure 3. Rendered SMPL-X frames from the VR-based referential dataset (MM-Conv).

Listing 1. Example of a structured JSON annotation for a referential utterance.

```
"9": {
  "snippet": "Oh yeah, yeah, I stole this
      from the Louvre.",
  "references": [
    {
      "ref_id": 0,
      "phrase": "this",
      "resolved_ref": "Painting_eda6d5e1",
      "plurality": "single",
      "original_label": "pronominal",
      "object_id": "Painting_eda6d5e1",
      "phrase_start": 38.86,
      "phrase_end": 39.04
    }
  ],
  "start": 37.880602861595214,
  "end": 39.702,
  "high_level": {
    "current_topic": "Painting_eda6d5e1",
    "topic_duration_id": 1
  }
}
```

body movements. MM-Conv is specifically designed to link language, motion and space, providing a rich foundation for research in co-speech gesture generation and situated communication tasks.

Reference Annotation Output. The output of the reference annotation process from [2] is stored in a structured JSON format. Each utterance is associated with a snippet of transcribed speech and a list of extracted references. For each reference, we store the resolved object ID, the original phrase, plurality information (e.g., "single" or "multiple"), the reference type (e.g., "exact" noun phrase), and the phrase's start and end timestamps (in seconds) within the utterance. An example is shown on Listing 1.

Each utterance also includes high-level topic annotations, which track the object of interest across multiple utterances. This structure enables fine-grained multimodal grounding, supporting tasks such as reference resolution, object localization, and dialog analysis.

Motion Segmentation Output. The motion data associated with each recording is segmented into *referential* and *non-referential* motion segments. Referential segments correspond to utterances that contain at least one resolved reference to an object in the environment. Each referential segment is extracted based on the start and end timestamps of the utterance, covering the motion that co-occurs with the referential expression. Non-referential segments capture all remaining motion, including both speech without object references and silent periods between utterances. To ensure complete coverage, non-referential motion is computed as the complement of the referential intervals over the full recording timeline. This partitioning enables training and evaluation of models for both grounded, reference-driven gesture generation and background or spontaneous motion modeling.

Following the filtering practices in prior motion-language datasets such as HumanML3D [13], we apply duration-based filtering to ensure data consistency. Referential motion segments shorter than 0.5 seconds or longer than 20 seconds are discarded. Non-referential motion is segmented into clips strictly between 0.5 and 20 seconds by splitting longer intervals and discarding very short ones. This process improves dataset uniformity and facilitates stable training for generative models.

Table 2 summarizes the motion segment statistics after filtering, with 2,394 referential and 2,721 non-referential motion clips. The overall dataset spans 6.14 hours, with a mean clip duration of 4.32 seconds across 64 joints in 3D space.

Table 3 summarizes the distribution of references across referential utterances. Out of 2,394 referential utterances, the majority involve a single reference, while a substantial proportion involve multiple references, highlighting the presence of both simple and complex referential structures. On average, each utterance includes 1.65 annotated refer-

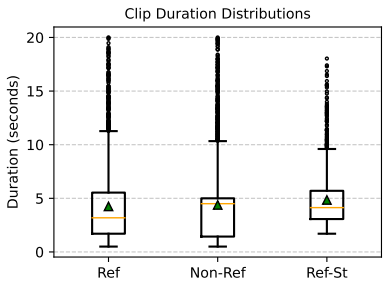


Figure 4. Distribution of motion clip durations for referential and non-referential segments after filtering. The boxplots show median (orange line), mean (green triangle), and outliers (black dots).

ences, enabling modeling of both simple and complex referential structures. At the reference level, most annotations point to single objects, with a smaller but notable fraction referring to multiple objects. These statistics provide insight into the diversity and richness of referential behavior in the dataset.

Statistic	REF	NON-REF
Number of clips	2394	2721
Total duration (min)	169.41	198.75
Mean duration (sec)	4.25	4.38
Std duration (sec)	3.56	3.65
Median duration (sec)	3.18	4.50
Motion dimensionality	64 joints × 3D	
Total dataset duration (min)	368.16 (6.14 hours)	
Overall mean duration (sec)	4.32	

Table 2. Statistics of the dataset after filtering, broken down into referential and non-referential motion segments.

Statistic	Value
Total referential utterances	2394
Utterances with 1 reference	1377 (57.5%)
Utterances with >1 reference	1017 (42.5%)
Total references	3955
References (single object)	2991 (75.6%)
References (multiple object)	964 (24.4%)
Average references per referential utterance	1.65

Table 3. Statistics of references across referential utterances after filtering.

3.4. Dataset Summary

Overall, the combined corpus spans over 7.7 hours of 3D motion data, with consistent clip durations and detailed reference labels. These resources enable the development and evaluation of generative models for co-speech and deictic gesture generation in situated contexts. Figure 4 illustrates the distribution of motion clip durations across referential and non-referential segments, summarizing the temporal structure of the dataset after filtering.

(a) Rendered RGB 3rd person view.

(b) Rendered RGB 1st person view.

Figure 5. Examples of rendered data sample of SMPL-X character in simulator [15].

3.5. Training Splits

To facilitate model training under different scenarios, we define three training splits with increasing data coverage:

- **ST only**: Only synthetic pointing motion sequences.
- **REF + ST**: Referential motion segments combined with synthetic pointing data.
- **ALL (REF + ST + NONREF)**: Full dataset including referential, synthetic pointing, and non-referential motion clips.

Table 4 summarizes the statistics of the three splits, highlighting the number of motion clips, total duration, and duration distribution characteristics.

Statistic	ST	REF+ST	ALL
# clips	1135	3529	6250
Total (min)	91.70	261.11	459.86
Mean (sec)	4.85 ± 2.65	4.45 ± 3.10	4.35 ± 3.29
Motion dim	64 joints × 3D		
Total (min)	459.86 (7.66 hours)		
Mean (sec)	4.45		

Table 4. Statistics of the three defined training splits: ST only (synthetic pointing), REF + ST (referential and synthetic pointing), and ALL (referential, synthetic pointing, and non-referential).

4. Experiments

In situated conversational agents, the generation task shifts from goal-driven locomotion to understanding spatial references in dialog using a vision-language model and generating referential language combined with gestures. An

open question is the appropriate spatial semantic representation for gesture generation: does the gesture module require the same spatial granularity as humanoid locomotion, or a more semantically nuanced representation? This affects how gestures align with communicative intent in 3D scenes.

Recent advances in controlled motion generation suggest that modular systems, where perception and generation are decoupled, offer better scalability and interpretability [18, 29]. Inspired by these approaches, we propose a modular framework where the system first derives referenced objects from the scene (e.g., via bounding boxes or graphs) and predicts a spatial cue, such as a referent's location, using a VLM or 3D scene understanding module (Figure 6). Conditioned on the detected spatial cue, a generative motion model then produces a gesture that grounds the intent toward the detected location. This decomposition enables flexible training, better interpretability, and supports extensions toward fully embodied agents in complex 3D environments.

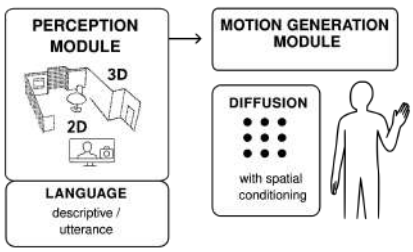


Figure 6. Proposed framework.

4.1. Motion generation with spatial conditioning

OmniControl [31] is a diffusion-based motion generation model trained on the HumanML3D dataset, which enables precise control over any joint in time, guided by text prompts and spatial constraints. It uses joint-conditioned guidance during training to align generated motions with the specified constraints while maintaining naturalness in the generated motion. This approach allows for flexible and high-fidelity motion synthesis, outperforming previous methods limited to pelvis control.

To incorporate spatial control, OmniControl introduces a spatial guidance mechanism that enforces joint-level constraints during the denoising process. At each diffusion timestep t, given a partially noised motion sequence $\mathbf{x}_t$, the model predicts the clean motion $\hat{\mathbf{x}}_0$ by minimizing a composite loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{denoise}} + \lambda_{\text{spatial}}\mathcal{L}_{\text{spatial}} + \lambda_{\text{realism}}\mathcal{L}_{\text{realism}}, \quad (1)$$

where $\mathcal{L}_{\text{denoise}}$ is the standard denoising loss over the motion sequence, $\mathcal{L}_{\text{spatial}} = \|\mathbf{x}_t^j - \mathbf{g}_t^j\|_2$ is the spatial loss ensuring that the generated joint positions $\mathbf{x}_t^j$ match the target positions $\mathbf{g}_t^j$ for controlled joints j, and $\mathcal{L}_{\text{realism}}$ encourages plausible and coherent motion trajectories. The spatial guidance is implemented by converting generated motions to global coordinates and aligning them with the spatial targets, allowing control over arbitrary joints and frames without retraining. This is especially relevant for grounded gesture generation, since the target representation is naturally defined in the global coordinate frame, allowing the model to produce gestures that are spatially aligned with the referent's position in the shared environment.

4.2. Training Configurations

We fine-tuned OmniControl [31] on three subsets of our dataset, defined in section 3.5, and evaluated models using three control joints —pelvis, left wrist, right wrist— chosen for their importance in modeling pointing gestures. Each configuration was evaluated in two modes: the base model trained exclusively on HumanML3D, and its fine-tuned counterpart adapted to each of the three dataset splits.

4.3. Results

Table 5 summarizes the Fréchet Inception Distance (FID) and Control L_2 error for each experimental configuration, using the same evaluation protocol as Guo et al. [12]. The FID measures the naturalness of the generated motion by quantifying the discrepancy between the distributions of generated versus ground-truth motions in the Inception latent space, while the Control L_2 error measures the average Euclidean distance between the model's predicted joint positions and the provided control hints. Lower values indicate better motion quality and control accuracy. Fine-tuning consistently improves both metrics across all training subsets and joints, demonstrating the benefit of adapting the pretrained OmniControl model to task-specific motion distributions.

Notably, the ST only fine-tuned model achieves the lowest overall errors, particularly for wrist joints (FID: 0.65–0.82, Control L_2: 0.058–0.060), highlighting the benefit of training on focused, synthetic pointing data. Incorporating referential data in the REF + ST split leads to improved pelvis control (FID: 1.03 vs. 1.72), suggesting that referential clips help better model full-body coordination. However, this comes with a modest increase in wrist error.

The ALL split, which includes the most diverse set of motions (referential, synthetic, and non-referential), results in slightly higher FID and Control L_2 values for most joints. This suggests a trade-off: while broader training data improves generalization, it may introduce variance that slightly reduces precision for fine-grained control joints. Nevertheless, fine-tuned models remain consistently supe-

Table 5. FID and Control L_2 error (↓) for base model (OmniControl) vs. fine-tuned models across the training sets.

Eval. Set	Controlled Joint	Base		Fine-tuned	
		FID ↓	L_2 ↓	FID ↓	L_2 ↓
ST only	Pelvis	6.37	0.036	1.72	0.020
	Left Wrist	3.82	0.114	0.82	0.060
	Right Wrist	4.60	0.116	0.65	0.058
REF + ST	Pelvis	5.75	0.112	1.03	0.034
	Left Wrist	3.99	0.099	1.56	0.064
	Right Wrist	4.26	0.129	1.41	0.109
ALL	Pelvis	7.96	0.125	2.00	0.055
	Left Wrist	4.08	0.113	1.25	0.099
	Right Wrist	4.49	0.132	0.84	0.123

rior to pretrained ones across all splits, underscoring the importance of task-adaptive training.

We also conduct initial evaluations on pointing gesture generation. Specifically, we select pointing motions from the test set, extract the joint positions of the wrist and elbow at the peak pointing frame, and provide these as fixed targets to the OmniControl algorithm. The resulting motion sequences, which are included in the supplementary material, demonstrate that fine-tuning on pointing gesture data contributes to generating more natural and conversational full-body motion. As a clear example, in `sample24.mp4` in the supplementary material, the base model performs a *punching motion* to reach the target pose, while the fine-tuned model performs a pointing gesture. This is promising and highlights the model's capacity to adapt to spatial constraints. However, this evaluation does not fully reflect the desired behavior in pointing tasks, where the goal is not to match absolute joint positions but to produce accurate pointing directions. Achieving this requires modifying the training procedure and redefining the spatial loss in OmniControl to enforce directional alignment, e.g., by aligning the vector from the elbow to the wrist with the vector toward the referent, rather than relying solely on positional targets.

5. Discussion

We standardized and extended two complementary motion capture datasets to advance research in *grounded gesture generation*. The first is a synthetic dataset consisting of spatially grounded pointing gestures paired with demonstrative utterances generated via vision-language prompting and aligned speech synthesis. The second is a VR-based referential dataset capturing natural, multimodal dialogue, annotated with object references and conversational structure. Both datasets are converted into a unified SMPL-X and HumanML3D-compatible format and segmented into referential and non-referential clips. The synthetic data supports controlled experimentation on isolated deictic gestures, while the VR dataset enables the study of gesture generation in dynamic, conversational settings where language, motion, and spatial context interact.

Our current evaluations primarily focus on isolated pointing gestures. While this serves as a useful baseline for spatial grounding, future work should extend to more temporally complex behaviors that better reflect the expressive range of referential communication. Nonetheless, our results demonstrate that fine-tuning on task-specific data significantly improves motion quality and control, laying a strong foundation for further advances.

To build on this foundation, future directions include:

Motion modeling. While models like OmniControl can be effectively fine-tuned for referential gesture generation, there remains room for improving how spatial intent is represented. The spatial constraints and position-based losses in OmniControl need to be re-evaluated to better capture the communicative function of gestures. Additionally, pretraining on general motion corpora like AMASS may be suboptimal for communicative behaviors. Subsampling AMASS to retain only socially relevant sequences, or integrating our data with multimodal datasets like BEAT2, could support more effective representation learning for grounded interaction.

Multimodal conditioning. Audio and speech conditioning are essential for co-speech gesture generation. To move toward full referential communication, future models should leverage synchronized speech input alongside spatial information. Our dataset is well-suited for such multimodal training regimes, combining natural dialogue, annotated references, and full-body motion in shared 3D environments.

Perception A critical component in this pipeline is the perception module, which provides spatial conditioning signals to the motion generator. These signals may be derived from 2D object detectors, depth maps, or 3D scene graphs, depending on the available modality. In our setup, spatial targets are extracted using a vision-language model or 3D simulation environment and projected into world coordinates. Future work could investigate the effect of different spatial representations, such as bounding boxes, keypoints, or object centroids, on gesture quality and grounding accuracy. Better perception modules may also support dynamic referent tracking and enable learning of spatial saliency over time.

References

[1] Tenglong Ao, Zeyi Zhang, and Libin Liu. GestureDiffuCLIP: Gesture diffusion model with CLIP latents. *ACM Transactions on Graphics (TOG)*, 42(4):1–18, 2023. 3

[2] Anonymous Authors. Multimodal referencing: Grounding generative models in situated communication, 2025. Under review. 5

[3] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023. 4

[4] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024. 4

[5] Anna Deichler, Simon Alexanderson, and Jonas Beskow. Incorporating spatial awareness in data-driven gesture generation for virtual agents. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents (IVA '24)*, pages 1–4, Glasgow, United Kingdom, 2024. ACM. 4

[6] Anna Deichler, Shivam Mehta, Simon Alexanderson, and Jonas Beskow. Diffusion-based co-speech gesture generation using joint text and audio representation. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 755–762, 2023. 3

[7] Anna Deichler, Jim O'Regan, and Jonas Beskow. MM-Conv: A multi-modal conversational dataset for virtual humans. *arXiv preprint arXiv:2410.00253*, 2024. 3, 4

[8] Anna Deichler, Siyang Wang, Simon Alexanderson, and Jonas Beskow. Towards context-aware human-like pointing gestures with rl motion imitation. In *Context-Awareness in Human-Robot Interaction: Approaches and Challenges, workshop at 2022 ACM/IEEE International Conference on Human-Robot Interaction*, 2022. 4

[9] Ylva Ferstl and Rachel McDonnell. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 93–98, 2018. 3

[10] Ylva Ferstl and Rachel McDonnell. IVA: Investigating the use of recurrent motion modelling for speech gesture generation. In *IVA '18 Proceedings of the 18th International Conference on Intelligent Virtual Agents*, Nov 2018. 3

[11] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. *CoRR*, abs/1906.04160, 2019. 3

[12] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR 2022*, pages 5152–5161, June 2022. 3, 7

[13] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022. 2, 3, 5

[14] Sri Rama Kartheek Kappagantula, Nicoletta Adamo-Villani, Meng-Lin Wu, and Voicu Popescu. Automatic deictic gestures for animated pedagogical agents. *IEEE Transactions on Learning Technologies*, 13(1):1–13, 2019. 3

[15] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. AI2-THOR: An interactive 3D environment for visual AI. *arXiv preprint arXiv:1712.05474*, 2017. 4, 6

[16] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. Talking With Hands 16.2M: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 763–772, 2019. 3

[17] James C Lester, Jennifer L Voerman, Stuart G Towns, and Charles B Callaway. Deictic believability: Coordinated gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence*, 13(4-5):383–414, 1999. 3

[18] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pages 54–72. Springer, 2024. 7

[19] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6), Nov. 2017. 3

[20] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, et al. GroundingGPT: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*, 2024. 4

[21] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. EMAGE: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1154, 2024. 2, 3

[22] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. *arXiv preprint arXiv:2203.05297*, 2022. 3

[23] Matthew Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and Shape capture from sparse markers. *ACM Trans. Graph.*, 33(6), Nov. 2014. 3

[24] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 2

[25] Tsukasa Noma, Liwei Zhao, and Norman I Badler. Design of a virtual human presenter. *IEEE Computer Graphics and Applications*, 20(4):79–85, 2002. 3

[26] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 3

[27] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT Motion-Language Dataset. *Big Data*, 4(4):236–252, Dec. 2016. 3

[28] Jeff Rickel and W Lewis Johnson. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied artificial intelligence*, 13(4-5):343–382, 1999. 3

[29] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–444, 2024. 7

[30] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned human motion generation in 3D scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3

[31] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. OmniControl: Control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 4, 7

[32] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-X: Towards versatile human-human interaction analysis. In *CVPR*, pages 22260–22271, 2024. 3

[33] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J. Black. Generating holistic 3D human motion from speech. In *CVPR 2023*, pages 469–480, Los Alamitos, CA, USA, June 2023. IEEE Computer Society. 3

[34] Hongwei Yi, Justus Thies, Michael J Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision*, pages 246–263. Springer, 2024. 2

6. Appendix

6.1. Exophoric Demonstrative Generation Prompt

The following prompt was used to generate exophoric referring expressions involving demonstratives (‘this’ or ‘that’) in a conversational style, using GPT-4o:

Listing 2. Prompt used for generating exophoric demonstrative references using GPT-4o.

```
\centering
Demonstratives in exophoric reference are
    used to indicate concrete physical
    entities in space.
For example, I was thinking sitting down.
    See that bench there, we could sit there
    .
Here 'that' would be the demonstrative.

I want you to generate examples like the
    following, which connect to a story.
The rules are that each generated sentence
    should contain one exophoric reference.

Examples:
- "See that bench over THERE, we could sit
    there."
- "Look in the middle of the room, THAT
    small dog is my new best friend."
- "We could get there faster. See THAT bike
     over there, we could take that one."
- "Look at the top shelf. THAT is the vase
    I wanted to buy."

Here's a bounding box: {bbox}, which
    highlights the object '{chosen_obj['name
    ']}'.
Can you write a casual, natural-sounding
    sentence that mentions '{chosen_obj['
    name']}'
using either 'this' or 'that'?
For example, something like: "And I haven't
     told you, but that table over there? I
    inherited it from my mother."
```

6.2. Examples of synthetic pointing gestures.

(a) Example one of rendered third person view of an aligned synthetic pointing example.

(b) Example one of rendered first person view of an aligned synthetic pointing example.

(c) Example two of rendered third person view of an aligned synthetic pointing example.

(d) Example two of rendered first person view of an aligned synthetic pointing example.

Figure 7. Rendered aligned synthetic pointing gesture examples in third and first person views.

Paper E

Paper E

A Benchmark for Scene-Aware Referential Gesture Generation

ANNA DEICHLER, RISHABH DABRAL, FETHIYE IRMAK DOĞAN, ANINDITA GHOSH, AND JONAS BESKOW

Published in
1st Workshop on Human–Scene Interaction at ECCV 2026, Malmö, Sweden

A Benchmark for Scene-Aware Referential Gesture Generation

Anna Deichler
KTH Royal Institute of Technology
Sweden

Rishabh Dabral
Max Planck Institute for Informatics
Germany

Fethiye Irmak Dogan
University of Cambridge
United Kingdom

Anindita Ghosh
Max Planck Institute for Informatics
Germany

Jonas Beskow
KTH Royal Institute of Technology
Sweden

Abstract

Referential gestures, pointing, indicating, and orienting the body toward objects in shared space, are fundamental to embodied communication. For virtual agents and physical robots operating in human environments, the ability to generate spatially grounded gestures is essential for disambiguation, instruction, and collaborative interaction. Yet, research on communicative gesture generation has largely focused on co-speech beat and iconic gestures, trained on corpora in which spatial grounding is absent or incidental. This lack of active research on referential gestures can be attributed to three key factors: datasets that pair gestures with 3D scene context are scarce, referential gesture generation lacks task formulation, and metrics for evaluating spatial grounding do not exist. In this work, we address all three gaps by introducing the MM-Conv Referential Gesture Generation Challenge. Specifically, the benchmark consists of three components: (i) a paired data release of 3,000 pointing-annotated clips from MM-Conv and SGS-HSI, with pointing-quality annotations and scene-disjoint splits; (ii) a task formulation that requires systems to produce spatially grounded reference gestures aligned with speech, without oracle apex timing or motion templates; (iii) a spatio-temporal evaluation protocol decomposing referential gesture quality into temporal alignment, spatial accuracy, and referent recall. We present a modular baseline based on OmniControl and position the benchmark as the foundation for the scene-aware gesture generation challenge at the 1st Workshop on Human–Scene Interaction at ECCV 2026. We envision this challenge as a testbed for the next generation of referential gesture synthesis works.

Current gesture evaluation relies on distributional metrics such as the Fréchet Gesture Distance (FGD), diversity, smoothness, that measure whether generated motion *looks realistic*, not whether it *successfully communicates spatial intent*. This mirrors a broader challenge in motion generation: in reinforcement learning approaches like Adversarial Motion Priors, a discriminator enforces distributional realism while task-specific rewards impose grounding constraints; gesture generation lacks the latter. Two generators with identical FGD can produce gestures that refer to entirely different objects, and no existing benchmark distinguishes them.

CCS Concepts

• **Computing methodologies** → **Animation**; *Computer vision.*

Keywords

gesture generation, human–scene interaction, referential grounding, co-speech motion, benchmark

ACM Reference Format:
Anna Deichler, Rishabh Dabral, Fethiye Irmak Dogan, Anindita Ghosh, and Jonas Beskow. 2026. A Benchmark for Scene-Aware Referential Gesture Generation. In *Proceedings of 1st Workshop on Human–Scene Interaction at ECCV 2026 (HSI '26).* ACM, New York, NY, USA, 7 pages. https://doi.org/toappear

1 Introduction

When a person says "put the cup on *that* one" while pointing at a chair, the meaning of the utterance lives jointly in the speech and in the speaker's body. A gesture generation model that reproduces realistic motion but points nowhere in particular has failed at the task that makes gestures useful in the first place: communicating something about the scene. Yet, this is exactly the failure mode that standard evaluation permits. Fréchet Gesture Distance [20] measures distributional realism; diversity measures coverage; neither tracks whether a generated gesture refers to the right object. Two models with identical FGD can differ arbitrarily in referential grounding, and no current benchmark distinguishes them.

This paper introduces a benchmark that closes that gap. The core claim is methodological: gesture generation evaluation needs per-sample, referent-grounded complements to FGD, and the data to construct them now exists. We build on MM-Conv [4], whose referring expressions are each annotated with a ground-truth referent in a 3D scene graph, thereby turning "does the motion look plausible" into "would a listener find the right object."

The benchmark comprises three core components: The first is objective grounding metrics, comprising:

(1) A **paired data release**: approximately 2K pointing-annotated clips from MM-Conv's naturalistic dialogue, and 1K clean single-target pointing clips from SGS-HSI with shared actors (§3).
(2) A **task formulation** treating referential gesture as an end-to-end problem: systems receive speech, text, and scene context, and must determine when to gesture, where to point, and how to move—without oracle timing or spatial targets (§4.1).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
HSI '26, Malmö, Sweden
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN to appear
https://doi.org/toappear

(3) A **spatio-temporal evaluation protocol** decomposing referential gesture quality into three axes (§4.3):
- *Axis 1a, temporal alignment.* Did the gesture peak at the communicatively appropriate moment?
- *Axis 1b, spatial accuracy.* At the system's chosen moment, did the pose point at the right place?
- *Axis 2, referent recall.* Did the pose uniquely indicate the intended object among scene-graph distractors?

The product of Axes 1a and 1b yields a combined spatio-temporal grounding (STG) score that prevents accuracy inflation: a gesture that accidentally passes through the correct pointing vector at the wrong time scores near zero. Furthermore, submissions are evaluated through both objective metrics (eg. FGD) and perceptual studies rating motion quality and semantic appropriateness.

We present a reference baseline (§5): a modular OmniControl-based pipeline that uses a learned IK module and phrase-anchored timing. The benchmark is designed to extend to more challenging scene-grounding conditions in future iterations (§7).

2 Related Work

Co-speech gesture generation. Recent work has driven rapid progress on speech-conditioned motion synthesis, with diffusion and flow-matching models trained on corpora such as BEAT2 [11] producing increasingly natural co-speech gestures. Evaluation is dominated by FGD [20] and diversity, which measure distributional properties of generated motion without reference to the environment or to what the utterance is about.

Scene-aware human motion. Human–scene interaction (HSI) and human–object interaction (HOI) have received significant attention [15, 16, 19, 21], with a focus on contact, affordance, and physically plausible motion. Wang et al. [17] use a dense skeleton-to-scene distance field to guide language-conditioned motion generation. Ghosh et al. [8] synthesize scene-aware motion through geometry-grounded token planning, demonstrating effective spatial reasoning for 3D motion generation in indoor environments. This literature largely treats the scene as a physical substrate rather than a referential one: motion is grounded in geometry but not in what the speaker is talking about.

Referential grounding. Referring expression benchmarks in 3D [1, 2] and dialogue [3, 14] have established the task of linking language to scene objects. MM-Conv [4] extended this to spontaneous multimodal dialogue with synchronized motion and gaze. Doğan et al. [7] address referential ambiguity in human-robot interaction through semantic disambiguation strategies. Ges3ViG [12] introduced the 3D Embodied Reference Understanding (3D-ERU) task, combining pointing gestures with language for object identification in ScanNet scenes, and showed that gesture substantially improves grounding accuracy. We repurpose the referential annotation paradigm in the opposite direction: from *understanding* reference to *generating* and *evaluating* it.

Challenge precedent. The GENEA Challenge [10] established community-wide benchmarking for speech-driven gesture generation, with FGD as the dominant objective metric. We follow the same convention for the in-context recall scorer (§4.3).

3 Corpus and Data Sources

The benchmark release comprises two paired data sources that share actors and capture conditions but differ in scene structure.

MM-Conv [4] is the primary evaluation data: approximately 2,000 six-second clips drawn from 6.7 hours of dyadic VR interaction across five AI2-THOR [9] indoor rooms, each centered on a referring expression accompanied by a detectable pointing gesture. Each clip is annotated with word-level speech transcripts, full-body SMPL-X motion capture at 30 fps, binocular gaze, the ground-truth referent, the dominant pointing hand, and the full scene graph (object identities, 3D bounding boxes, and positions; typically ∼50 candidate objects per scene). Expressions span exact noun phrases, partitive/attribute noun phrases, and pronominals ("that one," "this one"), reflecting the natural distribution in spontaneous dialogue.

SGS-HSI is a paired synthetic-scene resource: 1,138 clips recorded with the same five actors in simplified single-target setups. Each clip has a single annotated target coordinate and apex frame, but no surrounding scene graph. SGS-HSI provides clean supervision for the geometric primitive of pointing at an arbitrary 3D coordinate, uncluttered by distractors or referential ambiguity.

All sources are available in HumanML3D-compatible format. Participants may use any subset, though the declaration of training data is mandatory.

Spatial coverage. MM-Conv contributes room-scale geometry: targets reach up to 5.86 m from the speaker (mean 3.02 m, 95th percentile 5.19 m). SGS-HSI concentrates in the close-to-mid-range volume (mean distance 2.20 m, max 2.85 m).

Phrase metadata. We release referring-expression phrase intervals from the original MM-Conv corpus alongside each clip. These intervals reference the original session timeline and do not always correspond cleanly to the clip's annotated gesture: a 6-second clip may capture a gesture whose referring expression ends after the clip cut, or two distinct gestures from a dense pointing dialogue. The majority of MM-Conv test clips contain a single uniquely overlapping phrase; remaining clips have multiple, partial, or non-overlapping phrases. The benchmark scoring (§4.3) does not consume phrase timing; participants whose systems use phrase metadata as input should validate phrase-gesture alignment per-clip. SGS-HSI clips lack this property due to their single-target structure.

4 Benchmark

4.1 Task Formulation

We define two tracks.

Track A, target-aimed pointing. Given (i) a spoken utterance with word-level timing, (ii) a 3D target coordinate, and (iii) the referring expression timestamps, generate SMPL-X motion containing a pointing gesture directed at the target. No apex frame is provided; the system must determine when the gesture peak occurs. Evaluable on both MM-Conv and SGS-HSI.

Track B, scene-aware referential gesture. Given (i) a spoken utterance, (ii) preceding dialogue context, and (iii) the full 3D scene graph, generate SMPL-X motion that is communicatively appropriate and uniquely indicates the referenced object. Neither the target

Table 1: Scene-disjoint split statistics on MM-Conv and SGS-HSI.

Split	Rooms	MM-Conv	SGS-HSI	Total
Train	4 rooms	1,503	1,022	2,525
Test	1 room	339	116	455

coordinate nor any timing information is provided. Evaluable only on MM-Conv.

Track A is a strict sub-problem of Track B with the grounding step short-circuited.

4.2 Scene-Disjoint Splits

Evaluation uses a scene-disjoint split to test generalization to unseen environments. One room is held out as the test set; the remaining four rooms are released as training data. Participants may further subdivide the training rooms into train/validation splits as needed for their methods.

Table 1 summarizes the split statistics. The test room contains 339 MM-Conv referring expressions, with spatial and linguistic diversity comparable to the training rooms. By holding out an entire environment rather than random samples, the benchmark isolates the challenge of spatial generalization: models must ground references in novel room layouts with unseen object arrangements.

4.3 Evaluation Protocol: Spatio-Temporal Grounding

The protocol decomposes referential gesture quality into three independently measurable axes.

4.3.1 Axis 1a: Temporal alignment. Axis 1a measures whether the system's gesture occurs during the same time window as the ground-truth (GT) gesture. The GT *hold region* $[h_{\text{start}}, h_{\text{end}}]$ is defined kinematically as the contiguous frames where the speaker's forearm extension exceeds 99% of the clip-maximum after Gaussian smoothing ($\sigma = 2$ frames). On the test set, the hold has a median duration of 35 frames (≈1.2 s at 30 fps), reflecting that the gesture apex is a regional event rather than a single-frame instant in motion-capture data.

The submission ships motion only; an apex frame is not declared. The scorer applies a small kinematic detector (§4.3.1) to identify f_{peak} from the submitted motion. The temporal score is

$$\text{Score}_{1a} = \begin{cases} 1 & \text{if } h_{\text{start}} \le f_{\text{peak}} \le h_{\text{end}}, \\ \exp\left(-\frac{d^2}{2\sigma_t^2}\right) & \text{otherwise,} \end{cases} \tag{1}$$

where $d = \min(|f_{\text{peak}} - h_{\text{start}}|,\ |f_{\text{peak}} - h_{\text{end}}|)$ and $\sigma_t = 20$ frames. A system whose detected apex falls within the GT hold receives full credit; a system offset by 20 frames (≈667 ms) outside the hold scores ≈0.61 (one-σ); a system offset by a full second (≈30 frames) scores ≈0.32. The score is bounded in $[0, 1]$, enabling the multiplicative combination with Axis 1b (Eq. 3).

The 99% extension threshold and $\sigma_t = 20$ are calibrated empirically: the threshold isolates the apex event from the surrounding stroke and retraction, and σ_t is approximately half the median hold duration, so the soft-decay scale matches the natural temporal scale of the event.

Table 2: Pointing-quality statistics on ground-truth motion at the annotated gesture apex (test split, 455 clips). SGS-HSI clips use a 30 cm uniform cube around the target; MM-Conv clips use the per-object axis-aligned bounding box from the scene graph.

Source	N	Med. θ_{eff}	Score_{1b}	Hit rate
MM-Conv	339	9.6°	0.804	19.7%
SGS-HSI	116	10.9°	0.827	0.0%
All test	455	10.3°	0.809	14.8%

Apex Detector. The kinematic apex of the submitted motion is identified by a logistic-regression detector. The detector identifies *when* the gesture peaks (temporal information) by analyzing *how* the body moves (spatial kinematics. Per-frame features are pooled over a ±10-frame window and consist of forearm extension, wrist speed, wrist acceleration, forearm-direction stability, wrist height, and elbow angle (six raw features, max- and mean-pooled, yielding 12 dimensions). The detector is trained on the benchmark training split with frames within ±3 of the annotated apex labeled positive.

On the test set, the detector lands within the GT hold on 85.5% of clips, yielding a mean Score_{1a} of 0.92 on GT motion. This is the metric ceiling: human-annotated apex frames within the regional hold are not fully recoverable from kinematic features alone, so even GT motion does not score 1.0. Submission scores are interpreted against this ceiling.

4.3.2 Axis 1b: Spatial Accuracy. At the detected apex frame f_{peak} from Axis 1a, we measure spatial accuracy using an extent-aware angle that corrects for object size. Let $\mathbf{r}$ be the pointing ray (origin $\mathbf{o}$ at the wrist, direction along the forearm) and B the target's bounding box with centroid $\mathbf{p}$. The effective pointing angle is

$$\theta_{\text{eff}} = \arctan\left(\frac{d(\mathbf{r}, B)}{\|\mathbf{p} - \mathbf{o}\|}\right), \tag{2}$$

where $d(\mathbf{r}, B)$ is the minimum distance from the ray to the bounding-box surface (0 if the ray intersects B). The angular score at a single frame is $\exp(-\theta_{\text{eff}}^2 / 2\sigma_{\text{angle}}^2)$ with $\sigma_{\text{angle}} = 20°$. This width is calibrated to the scene ambiguity scale: across the benchmark scene graphs the mean target distance is 2.16 m and the mean object angular diameter at that distance is 30.7° (≈ 22° at 3 m). σ_{angle} thus sits in the regime where a one-standard-deviation pointing error corresponds to missing the intended target among typical nearby distractors, anchoring the soft falloff to a task-meaningful scale rather than an arbitrary kernel width. Table 2 summarizes pointing quality on ground-truth motion.

SGS-HSI's effectively-zero hit rate reflects the small 30 cm target volume; humans rarely pierce a sub-half-meter cube even when aiming correctly. The MM-Conv hit rate (20%) reflects the larger object-level bounding boxes. Both subsets achieve $\text{Score}_{1b} \approx 0.81$, indicating that the angular Gaussian dominates over hit/miss in the spatial axis: human pointing is consistently in the ∼10° neighborhood of the target regardless of object size.

4.3.3 Axis 2: Referent recall. Axis 2 measures whether the generated pose uniquely indicates the intended referent among scene-graph distractors. Using the kinematic-detector apex f_{peak} from Axis 1a, we extract pose-derived angular features pooled over a temporal window and rank all objects in the scene graph. We report top-K recall: whether the ground-truth referent appears in the first K candidates. We evaluate two rankers. *Heuristic* is a geometric rule-based scorer over ray angle, head direction, and distance, with no learned parameters. *Learned* is a 73K-parameter MLP trained on MM-Conv using leave-one-room-out cross-validation [6]. Both rankers consume the same pose-derived angular features and operate at the kinematic-detector apex used in Table 4. Challenge submissions use the Heuristic ranker to ensure full transparency and reproducibility; the Learned ranker is reported as an analysis metric.

Axis 2 is meaningful on MM-Conv (with scene-graph distractors); SGS-HSI lacks this distractor structure, so top-K referent recall is not comparably informative there.

Combined Spatio-Temporal Grounding (STG).. The product of temporal and spatial scores yields the combined STG metric:

$$\text{STG} = \text{Score}_{1a} \times \text{Score}_{1b}. \tag{3}$$

This multiplicative combination prevents accuracy inflation: a gesture that accidentally passes through the correct pointing vector at the wrong time scores near zero, as does a gesture with perfect timing but poor spatial accuracy.

Diagnostic utility. The three-axis decomposition enables interpretable failure-mode diagnosis. A system with high Axis 1a but low Axis 1b fails on spatial precision while timing correctly; the converse indicates temporal misalignment despite accurate pointing. Axis 2 complements Axis 1b by testing whether the pose disambiguates the referent among scene distractors: a system can achieve moderate Axis 1b (pointing near the target within angular tolerance) but low Axis 2 if the learned ranker finds the body configuration ambiguous among nearby candidates. Together, the axes expose *where* modeling effort should be directed rather than collapsing multi-dimensional failure into a single scalar.

5 Reference Baseline

We present a reference baseline that produces SMPL-X motion and score it with the proposed grounding evaluation metrics.

5.1 OmniControl-PT - Learned IK + Phrase Heuristic

This baseline is fully modular: timing and spatial placement are handled by separate components with no joint learning between them.

Temporal: phrase-anchored heuristic. We predict the apex frame at a fixed offset before the end of the referring expression: $f_{\text{peak}} = f_{\text{phrase_end}} - \delta$, where δ is calibrated on training data as the median offset of GT apex from phrase end. On our train split this gives $\delta = 11$ frames ($\approx$367 ms at 30 fps), consistent with the gesture-stroke-precedes-lexical-affiliate pattern reported in gesture studies [13?]. The heuristic uses only the released phrase timestamps; no learned parameters.

Table 3: Conditioning gap: predicted vs. GT control signals on the test set, before any motion generation. Temporal offset $\Delta f = f_{\text{pred}} - f_{\text{GT}}$, in frames; positive values indicate the predicted apex falls later than GT. Spatial errors compare learned-IK joints (at the predicted apex) to GT joints (at the GT apex).

Subset	n	Mean Δf	In-hold (%)	Wrist (m)	Dir. err. (°)
All test	455	+22	62.9	0.18	15.2
SGS-HSI	116	+1	68.1	0.10	8.0
MM-Conv	339	+29	61.1	0.23	21.2

The predictor's accuracy depends on phrase-gesture alignment in the source data (§3). On clips whose phrase recording-time overlaps the clip window (all of SGS-HSI; ∼60% of MM-Conv test), the heuristic lands in the GT hold reliably; on remaining MM-Conv clips, the predicted apex misses the gesture by ∼1 s on average. The resulting gap serves as the baseline floor against which alternative timing predictors—audio-conditioned or learned from acoustic features—can be evaluated.

Spatial: learned IK module. At the predicted apex frame, OmniControl receives wrist and elbow positions as spatial control hints. These are produced by a learned IK module: a small MLP (∼108K parameters) that maps (target_{xyz}, hand, $\boldsymbol{\beta}$) to predicted wrist and elbow positions at the apex frame, where $\boldsymbol{\beta}$ are SMPL-X body shape parameters. The MLP is trained on the benchmark training split with mean-squared loss with respect to GT joint positions extracted via SMPL-X forward kinematics at the GT apex frame, plus a cosine similarity term on forearm direction. Five hidden layers ($256 \to 256 \to 128$, LayerNorm + ReLU) with a 16-dimensional hand embedding feed a 6-dimensional output (wrist + elbow).

The body shape parameters $\boldsymbol{\beta}$ implicitly encode actor identity given the five-actor structure of the corpus; per-actor validation error is uniform (16–20° forearm direction error across actors), indicating that residual error arises from target-to-pose ambiguity rather than per-speaker variation. As a comparison point, a rest-pose IK heuristic (yaw-rotated rest skeleton, wrist placed at 99% extension along the shoulder-to-target ray, small perpendicular elbow offset) achieves wrist Euclidean RMSE 0.46 m and forearm direction error 34.1° on validation; the learned IK reduces these to 0.24 m and 18.0°. Closing the remaining gap would require conditioning on speaker pose at apex (pelvis position, torso orientation), which is not available prior to motion generation; we leave this for future work.

The conditioning gap is markedly narrower on SGS-HSI (mean apex offset +1 f, wrist 10 cm, direction 8°) than on MM-Conv (+29 f, 23 cm, 21°). The learned IK module also performs better on SGS-HSI's simpler single-target geometry; MM-Conv's room-scale variation introduces more target-to-pose ambiguity. SGS-HSI is therefore the cleaner subset for isolating model-side effects from data-side noise; MM-Conv reflects the harder real-world condition that the challenge will probe.

Motion generation: OmniControl. OmniControl [18] is a diffusion-based motion generation model that enables precise control over any joint through spatial guidance during denoising. Following [5],

Table 4: Spatio-temporal grounding on the test split (455 clips). Scores reflect kinematically detected apex frames for all methods. STG = $\text{Score}_{1a} \times \text{Score}_{1b}$. FGD is computed against EMAGE VAE.

Condition	temporal ↑	spatial ↑	STG ↑	Med. θ_{eff} ↓	FGD ↓
GT MoCap	0.909	0.766	0.718	11.8°	0.961
OmniControl-PT	0.823	0.515	0.431	23.4°	5.612

Table 5: Axis 2: referent recall on the test split. Top-K candidate recall over scene-graph distractors. Heuristic is a geometric ranker; Learned is a 73K-parameter MLP. All methods are evaluated at the kinematic-detector apex.

Condition	Scorer	Top-5 ↑	Top-10 ↑
GT MoCap	Heuristic	44.1%	63.6%
	Learned	51.0%	68.7%
OmniControl-PT	Heuristic	21.6%	41.1%
	Learned	25.4%	45.2%

we fine-tune OmniControl on the benchmark training set: MM-Conv referential clips from train set rooms, combined with SGS-HSI synthetic pointing data. Fine-tuning reduces control L2 error on the wrist joints (base vs. fine-tuned). At inference, the model receives wrist and elbow positions at the predicted apex frame and generates full SMPL-X motion without consuming ground-truth data.

6 Results

We report results on the Test Split for GT motion capture and the modular baseline.

6.1 Spatio-Temporal Grounding

Table 4 reports spatio-temporal grounding results on the test split (455 clips). GT MoCap establishes the metric ceiling: it represents human pointing in natural interaction, scored end-to-end by the same kinematic apex detector applied to all submissions. The OmniControl-PT baseline is fully automatic; no ground-truth motion or timing is consumed at inference, and it is scored identically. A second baseline will be reported at the workshop alongside challenge submissions.

Referent recall ceiling. Human pointing in natural conversation recovers the intended referent in roughly two-thirds of clips at top-10 (68.7%, Learned) against ~50 candidates per scene—a 3.4× improvement over uniform random ranking. Top-5 and top-10 are the informative axes here: many references in MM-Conv are intentionally underspecified at the gesture apex, with speakers gesturing toward a region containing several plausible candidates and relying on dialogue to disambiguate. The metric is designed to capture whether the gesture brings the referent into the speaker's attentional shortlist, not whether it uniquely identifies a single object from kinematics alone.

Learned vs. heuristic scorer. The Learned scorer outperforms the geometric Heuristic on both motion sources (top-5 on GT: 51.0% vs. 44.1%; on OmniControl-PT: 25.4% vs. 21.6%). The gap is consistent across motion conditions, indicating that the learned scorer captures pointing geometry that a fixed-rule ranker misses—for example, the relative weighting of head direction and ray angle, which is target-distance dependent. This validates Axis 2 as a learned metric rather than a hand-tuned threshold, and supports its use as the referent-recall component of the challenge protocol.

Generation gap. OmniControl-PT preserves about half of the GT referent signal (top-5 Learned: 25.4%/51.0% ≈ 50%; top-10: 45.2%/68.7% ≈ 66%). The drop is consistent with the Axis 1b spatial accuracy gap (Table 4: 0.515/0.766 ≈ 67%): a generator that points in roughly the right direction but with ~2× the angular error of human pointing recovers the correct referent in roughly half as many cases at strict K, and roughly two-thirds at lax K. Axis 2, therefore, does not introduce a new failure mode beyond what Axis 1b already exposes; it converts the same geometric error into a directly interpretable communicative outcome ("did the gesture identify the right object"), which is the signal practitioners and downstream HRI applications care about.

Metric ceiling. GT MoCap scores $\text{Score}_{1a} = 0.909$ rather than 1.0 because the kinematic apex detector, applied uniformly to all submissions, does not perfectly recover annotated apex frames from kinematics alone (85.5% within-hold rate; §4.3.1). The spatial score $\text{Score}_{1b} = 0.766$ reflects detector-chosen frames rather than the annotated apex; evaluated at the annotated apex directly, human pointing achieves $\text{Score}_{1b} = 0.809$ (Table 2). Both numbers are upper bounds for the generated motion under this protocol.

OmniControl-PT baseline. The modular pipeline achieves STG = 0.431, against a GT ceiling of 0.718. The gap decomposes cleanly across the two axes. Temporal alignment ($\text{Score}_{1a} = 0.823$ vs. 0.909) reflects the phrase-anchored heuristic: the heuristic places the apex in the GT hold on 62.9% of clips (Table 3), with larger miss rates on MM-Conv's naturalistic dialogue than on SGS-HSI's controlled setups. Spatial accuracy ($\text{Score}_{1b} = 0.515$ vs. 0.766) reflects the combined cost of learned-IK imprecision (18.0° forearm direction error at the predicted apex) and the diffusion model's soft rather than hard enforcement of spatial hints: OmniControl's guidance mechanism steers joint positions toward the hint but does not guarantee exact placement, so spatial accuracy is evaluated at whatever frame the kinematic detector finds, which may not coincide with the intended apex. The median angular error ($\theta_{\text{eff}} = 23.4°$ vs. 11.8° for GT) confirms that the baseline points in roughly the right direction but lacks the precision required for object-level grounding.

Together, these results characterize the modular baseline as a functional lower bound: it produces directionally plausible gestures with reasonable timing on well-aligned clips, but falls substantially short of human performance on both axes. The decomposition in Table 3 shows that the spatial gap is larger on MM-Conv than SGS-HSI, consistent with room-scale target variation being harder for the fixed IK module. Both gaps are targets for challenge participants.

Spatial accuracy ceiling. GT motion achieves $\text{Score}_{1b} = 0.809$ at annotated apex frames, with median angular error ≈10°. This is the natural ceiling: humans pointing at real targets in natural conversation. Systems that receive explicit target coordinates as input (Track A) can, in principle, approach this number; systems

that must infer the target from language and scene context (Track B) face an additional grounding step that introduces further error. The gap between Track A and Track B performance on the leaderboard will quantify the cost of scene-aware grounding.

7 Challenge

The benchmark serves as the basis for the scene-aware gesture generation challenge at the 1st Workshop on Human–Scene Interaction at ECCV 2026.

Phase 1: AI2-THOR scenes (this release). Participants receive MM-Conv training data from four rooms with full annotations, plus SGS-HSI as auxiliary data. One room is held out as the test set. Submissions consist of generated SMPL-X motion, scored under FGD, Axis 1a, 1b, STG, and—for Track B—Axis 2. Axis 2 scoring uses the geometric Heuristic ranker (Table 5) to ensure full transparency and reproducibility. Track B baselines integrating automatic language-to-target grounding are left for future releases.

Future extensions. The current benchmark uses controlled AI2-THOR environments with canonical object arrangements. Future iterations may extend to more complex real-world scene layouts and integrate with referring expression grounding baselines for Track B evaluation, testing generalization to cluttered environments, and automatic target inference.

No single-scalar ranking. The results will be reported for all axes separately. Pareto-dominant submissions are highlighted. Perceptual evaluation will be conducted during the challenge to contextualize objective scores and identify cases where motion scores well on metrics but lacks communicative appropriateness.

8 Conclusion

Gesture generation evaluation has been dominated by distributional realism metrics because, until recently, the data required for other metrics has been unavailable. MM-Conv's referential annotations close that gap. The spatio-temporal decomposition proposed here—separating *when* the system gestures (Axis 1a), *where* it points (Axis 1b), and *what* it indicates (Axis 2)—turns those annotations into complementary scoring functions whose divergence is itself diagnostic of generator failure modes. By withholding oracle timing from submissions, the benchmark evaluates referential gesture generation rather than conditioned motion synthesis.

The OmniControl-PT baseline establishes a functional lower bound: STG = 0.431 (at $\sigma_{\text{angle}} = 20°$) against a GT ceiling of 0.718, with the gap split between timing (phrase-heuristic misses on naturalistic MM-Conv dialogue) and spatial accuracy (soft constraint enforcement and learned-IK imprecision). This decomposition is the benchmark's primary contribution to participants: knowing *where* the gap is tells them where to direct their modeling effort.

The benchmark establishes a foundation for scene-aware gesture evaluation on controlled environments with ground-truth scene graphs. Future extensions to real-world scans will test transfer to unstructured geometry, connecting the gesture generation community with the broader 3D referential grounding literature.

Limitations. The extent-aware angle uses axis-aligned bounding boxes, which overestimate extent for elongated objects at oblique angles. Phrase-gesture association in MM-Conv is structurally noisier than in single-target setups; systems consuming phrase metadata as input may need to validate alignment per-clip. The current evaluation focuses on controlled environments; extension to more complex real-world layouts will test robustness to cluttered scenes and challenging spatial configurations.

Code and model release. The benchmark scorer (Axes 1a, 1b, 2), the apex detector, and the OmniControl-PT baseline weights will be released at https://huggingface.co/mm-conv-scene-gesture alongside the data release to ensure reproducibility and lower barriers to participation.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. 2020. ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes. In *European Conference on Computer Vision (ECCV)*. 422–440. doi:10.1007/978-3-030-58452-8_25

[2] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. 2020. ScanRefer: 3D Object Localization in RGB-D Scans Using Natural Language. In *European Conference on Computer Vision (ECCV)*. 202–221. doi:10.1007/978-3-030-58565-5_13

[3] Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. 2021. YouRefIt: Embodied Reference Understanding with Language and Gesture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 1385–1395.

[4] Anna Deichler, Jim O'Regan, Fethiye Irmak Dogan, Lubos Marcinek, Anna Klezovich, Iolanda Leite, and Jonas Beskow. 2026. MM-Conv: A Multimodal Dataset and Benchmark for Context-Aware Grounding in 3D Dialogue. In *Proceedings of the Language Resources and Evaluation Conference (LREC-COLING)*. LREC 2026.

[5] Anna Deichler, Jim O'Regan, Téo Guichoux, David Johansson, and Jonas Beskow. 2025. Grounded Gesture Generation: Language, Motion, and Space. *arXiv preprint arXiv:2507.04522* (2025). Presented at the Humanoid Agents Workshop, CVPR 2025.

[6] Anna Deichler, Siyang Wang, Simon Alexanderson, and Jonas Beskow. 2023. Learning to Generate Pointing Gestures in Situated Embodied Conversational Agents. *Frontiers in Robotics and AI* 10 (2023), 1110534. doi:10.3389/frobt.2023.1110534 Author version also available as arXiv:2509.12507.

[7] Fethiye Irmak Dogan, Maithili Patel, Weiyu Liu, Iolanda Leite, and Sonia Chernova. 2025. A Model-Agnostic Approach for Semantically Driven Disambiguation in Human-Robot Interaction. In *IEEE RO-MAN*. 649–656.

[8] Anindita Ghosh, Vladislav Golyanik, Taku Komura, Philipp Slusallek, Christian Theobalt, and Rishabh Dabral. 2026. SceMoS: Scene-Aware 3D Human Motion Synthesis by Planning with Geometry-Grounded Tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. 2017. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv preprint arXiv:1712.05474* (2017).

[10] Taras Kucherenko, Pieter Jonell, Youngwoo Yoon, Patrik Wolfert, and Gustav Eje Henter. 2023. The GENEA Challenge 2023: A Large-Scale Evaluation of Gesture Generation Models in Monadic and Dyadic Settings. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI)*. 792–801.

[11] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. 2024. EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Expressive Masked Audio Gesture Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1144–1154.

[12] Atharv Mahesh Mane, Dulanga Weerakoon, Vigneshwaran Subbaraju, Sougata Sen, Sanjay E Sarma, and Archan Misra. 2025. Ges3ViG: Incorporating Pointing Gestures into Language-Based 3D Visual Grounding for Embodied Reference Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9017–9026.

[13] David McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

[14] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tür. 2022. TEACh: Task-Driven Embodied Agents That Chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 2017–2025.

[15] Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. 2025. TokenHSI: Unified Synthesis of Physical Human-Scene Interactions through Task Tokenization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5379–5391.

[16] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. 2025. TriDi: Trilateral Diffusion of 3D Humans, Objects, and Interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5523–5535.

[17] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. 2024. Move as You Say, Interact as You Can: Language-Guided Human Motion Generation with Scene Affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 433–444.

[18] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. 2024. OmniControl: Control Any Joint at Any Time for Human Motion Generation. In *The Twelfth International Conference on Learning Representations*.

[19] Pradyumna Yalandur Muralidhar, Yuxuan Xue, Xianghui Xie, Margaret Kostyrko, and Gerard Pons-Moll. 2025. PhySIC: Physically Plausible 3D Human-Scene Interaction and Contact from a Single Image. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 1–12.

[20] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Transactions on Graphics* (2020). doi:10.1145/3414685.3417838

[21] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya A Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez Pellitero, and Gerard Pons-Moll. 2025. FORCE: Dataset and Method for Intuitive Physics Guided Human-Object Interaction. In *12th International Conference on 3D Vision*. IEEE.

Paper F

Paper F

MM-Conv: A Multimodal Dataset and Benchmark for Context-Aware Grounding in 3D Dialogue

ANNA DEICHLER, JIM O'REGAN, FETHIYE IRMAK DOĞAN, LUBOŠ MARCINEK, ANNA KLEZOVICH, IOLANDA LEITE, AND JONAS BESKOW

Published in
Proceedings of the 15th Language Resources and Evaluation Conference (LREC 2026), Palma de Mallorca, Spain

MM-Conv: A Multimodal Dataset and Benchmark for Context-Aware Grounding in 3D Dialogue

Anna Deichler, Jim O'Regan, Fethiye Irmak Dogan, Lubos Marcinek, Anna Klezovich, Iolanda Leite, and Jonas Beskow
KTH Royal Institute of Technology, Stockholm, Sweden
{deichler, joregan, fidogan, lubosm, annkle, iolanda, beskow}@kth.se

Abstract

Grounding language in the physical world requires AI systems to interpret references that emerge dynamically during conversation. While current vision-language models (VLMs) excel at static image tasks, they struggle to resolve ambiguous expressions in spontaneous, multi-turn dialogue. We address this gap by introducing (1) a benchmark for referential communication in dynamic 3D environments, built from 6.7 hours of egocentric VR interaction with synchronized speech, motion, gaze, and 3D scene geometry, and (2) a two-stage grounding pipeline that explicitly resolves conversational ambiguity before visual localization. The benchmark includes over 4,200 manually verified referring expressions spanning full, partitive, and pronominal types. Our contextual rewriting approach improves grounding performance by 11–22 percentage points on average, with a pure detector (GroundingDINO) reaching 56.7% on pronominals after rewriting—nearly double the best end-to-end baseline. Results demonstrate that decoupling linguistic reasoning from visual perception is more effective than end-to-end approaches for conversational grounding.

Keywords: vision-language models, referential communication, multimodal grounding, egocentric dialogue, embodied AI

1. Introduction

Understanding and resolving referring expressions in situated, real-world contexts remains a core challenge for multimodal AI. While recent progress in vision-language models (VLMs) has brought significant advances in grounding natural language to images and videos, these models often fall short when it comes to reference resolution in dynamic, embodied environments. Here, language is rarely disambiguated by words alone; instead, meaning is distributed across multiple modalities, including speech, gesture, gaze, and spatial context, unfolding over time and grounded in interaction (Gross et al., 2017). These capabilities are critical for applications such as home-assistive robots or AR-based guides, where agents must interpret ambiguous phrases like 'that one' in context.

Existing benchmarks for referring expression grounding, such as ScanRefer (Chen et al., 2020) or ReferIt3D (Achlioptas et al., 2020), have driven progress but are limited by their reliance on scripted, single-turn textual descriptions and static 3D scenes. This leaves a critical gap in evaluating models on the spontaneous, messy, and multimodal nature of real human dialogue. To address this, we introduce a new benchmark that shifts the focus to the challenge of resolving spontaneous, conversational references as they emerge naturally in rich, situated dialogues. These expressions, such as "this one here", "the yellowish thing", or simply "it", are inherently ambiguous unless grounded in a shared perceptual and interactional context.

In realistic, situated dialogue, language is often insufficient on its own. The problem of referential understanding has been approached using descriptive features in language (Shridhar et al., 2020; Doğan et al., 2022; Dogan et al., 2025), but in spontaneous interaction, non-verbal cues are crucial for resolving ambiguity, with studies showing that referential gaze is tightly synchronized with speech (Staudte and Crocker, 2011) and that gestures provide effective guidance for disambiguating objects (Zhang et al., 2024; Sauppé and Mutlu, 2014). Recent work has explored learning to generate such pointing behaviors for embodied agents (Deichler et al., 2023).

Early benchmarks miss core conversational phenomena—ellipsis, anaphora, and multimodal deixis, common in situated dialogue. Current VLMs, largely trained on text–image pairs, struggle to resolve such ambiguities without nonverbal cues (e.g., gaze, gesture) and often lack explicit, structured reasoning for highly underspecified references. Our contribution is the first benchmark that *synchronizes* these signals with conversational language, setting the stage for truly multimodal grounding.

To support this, we built a 6.7-hour dataset of immersive, multimodal interactions with synchronized speech, motion, gaze, and 3D scene geometry. We annotated 4,211 naturally occurring referring expressions and designed an evaluation framework spanning explicit nouns to highly con-

textual pronouns. Results reveal critical limitations of state-of-the-art VLMs for temporally grounded references and demonstrate the effectiveness of a modular, ambiguity-first pipeline for context-aware grounding.

Our main contributions are:

- **A new benchmark for spontaneous, multimodal referential grounding**, with 4,211 annotated expressions from 6.7 hours of conversational data in dynamic 3D environments.
- **A two-stage grounding pipeline** that separates conversational disambiguation from visual localization, improving performance by 11–22 percentage points across reference types.
- **Systematic evaluation** revealing that (1) current VLMs struggle with conversational references despite native context support, and (2) modular architectures outperform end-to-end approaches for situated grounding.
- **A rich, reusable dataset** with synchronized full-body motion, gaze, facial expression, and egocentric video to support downstream research in embodied interaction.

The rest of the paper is structured as follows: Section 2 describes related work in referential understanding and grounding in vision-language models. Section 3 describes the dataset and the annotations, and Section 4 presents the human and VLM experiments.

2. Related Work

2.1. Referential Understanding Benchmarks

Foundational datasets like ScanRefer (Chen et al., 2020) and ReferIt3D (Achlioptas et al., 2020) established 3D referring expression grounding but rely on single-turn, text-only descriptions. While YouRefIt (Chen et al., 2021) added multi-turn dialogue with gesture, it uses third-person video. TEACh (Padmakumar et al., 2022) includes embodied dialogue but lacks continuous multimodal streams (gaze, motion).

Our work addresses this gap by integrating temporally synchronized speech, full-body motion, gaze, and egocentric vision from immersive VR, enabling the investigation of reference resolution with all modalities naturally co-occurring in interaction. Table 1 summarizes key differences.

2.2. Vision–Language Models for Grounding

Vision-language models (VLMs) for grounding have evolved significantly, moving from dual-encoder architectures to more flexible generative models. This progression is critical for handling the spontaneous, conversational language, which stands in contrast to the clean, template-like queries found in traditional referring expression benchmarks. Models like GroundingDINO (Liu et al., 2024), built on CLIP (Radford et al., 2021) embeddings, perform well on short to medium-length prompts; however, their ability to generalize to longer, more conversational inputs is limited. In contrast, Florence-2 (Xiao et al., 2024) integrates a co-trained causal language model that supports prompt-driven decoding over both visual and textual inputs. This makes Florence-2 particularly well-suited for grounding tasks in ecologically valid settings where referring expressions are embedded in natural dialogue. Recent advances in GPT-style vision-language models, such as GroundingGPT (Li et al., 2024), further push this boundary by leveraging large-scale, decoder-only architectures that unify visual and textual modalities in a generative framework. Table 2 provides a comprehensive comparison of these vision-language grounding models, highlighting their architectural differences and capabilities.

3. Dataset

To create a benchmark for spontaneous, multimodal referential grounding, we collected a new dataset consisting of 6.7 hours of interaction data recorded during a referential communication task in a virtual environment. Our primary goal was to capture the richness of embodied dialogue, including synchronized speech, full-body motion, gaze, and 3D scene geometry, providing a foundation for studying grounding in a controlled yet naturalistic setting. Further details on data collection can be found in the Appendix.

3.1. Dataset Collection and Experimental Setup

The dataset was collected in dyads following a typical instruction-giver/follower paradigm, common in cognitive studies of referential communication. The main actor (instruction-giver), equipped with a full-body motion capture suit, finger-tracking gloves, and a VR headset with gaze and facial tracking, described objects and spatial arrangements within simulated apartment environments (AI2-THOR (Kolve et al., 2017)). The interlocutor (instruction-follower) responded naturally but was instructed not to introduce new referents, to elicit a high density of spontaneous referring expressions

Table 1: Comparison with prior multimodal referential benchmarks.

Dataset	Modalities	Scale	Embodied	Dialogue
ScanRefer (2020)	Text, 3D	51.6k, 800 scenes	No	One-shot
YouRefIt (2021)	Text, RGB, Gesture	4k, 432 scenes	Yes	Multi-turn
TEACh (2022)	Text, Sim3D, Actions	3k dialogues	Nav/Manip	Multi-turn
Ours	**Speech, RGB, Depth, Motion, Gaze**	**4.2k, 5 scenes**	**Full VR**	**Spontaneous**

Table 2: Comparison of vision-language grounding models, including text/visual encoders, training status, and notable capabilities.

Model	Text Enc.	Visual Enc.	Visual In.	Token	Notable Features
Grounding DINO (Liu et al., 2024)	BERT	DETR (Swin)	RGB (2D)	256	Referring expression grounding via cross-attention; zero-shot localization with box output.
Florence-2 (Xiao et al., 2024)	Transformer (BART)	DaViT	RGB (2D)	1024	Foundation model with prompt-driven generation; supports grounding using polygon or box output.
KOSMOS-2 (Peng et al., 2024)	Transformer (causal LM)	ViT (patch14-224)	RGB (2D)	2048	Grounds text spans to boxes via "location tokens"; outputs grounded captions/VQA with box coordinates (open-vocab grounding).
Ferret (You et al., 2024)	Vicuna-style LLM	CLIP ViT-L/14 (region features)	RGB (2D)	4096	Referring and region-level understanding via explicit region features; supports referring/grounded dialogue with box-based region specification.
GroundingGPT (Li et al., 2024)	Vicuna-v1.5	CLIP ViT-L/14, Q-former (video)	RGB (2D), Video	4096	Multi-modal model achieving SOTA performance in image, video, and audio grounding.
Qwen2.5-VL (Bai, 2025)	Qwen2.5 (causal LM)	ViT (window attn, SwiGLU)	RGB (2D), Video	4–16k (dyn.)	Dynamic resolution with mRoPE; native bounding-box grounding; supports multi-turn dialogue and visual agent tasks.

and associated nonverbal behaviors from the main actor. We selected five apartment environments from AI2-THOR, each containing a diverse set of interactable objects, with 2–3 short scenarios per room to vary discourse goals (e.g., showing a new apartment, landlord inspection, interior-designer suggestions) while maintaining ecological validity. This setup yields synchronized speech, full-body motion, gaze, facial expressions, and structured 3D scene graphs with object-level metadata for every frame. Further details on the data collection and annotation processes are provided in the Appendix.

Main–Interlocutor protocol.

- **Roles**: Main actor introduces and describes objects; interlocutor reacts but avoids adding novel referents, preserving a single discourse locus per segment.
- **Scenarios**: Per room, 2–3 role-play scenarios (e.g., "bragging about a new apartment," "landlord inspection," "interior designer tips") encourage varied yet natural reference types (full, partitive/attribute, pronominal).

Hardware and synchronization. Collection used OptiTrack mocap (50 markers), MANUS MetaGloves (finger tracking), and Meta Quest Pro (binocular gaze, 52 facial blendshapes). All streams were synchronized via SMPTE timecode injected into Unity, with headset pose calibrated to mocap for egocentric alignment.

Environments and objects. Five apartment rooms from AI2-THOR (Unity) were pre-selected to span common household categories and spatial layouts; scene graphs (per frame) were exported from the simulator with canonical object identifiers to support instance-level grounding. (See Appendix for object distributions.

3.2. Annotation

All modalities were aligned using the shared SMPTE timecode recorded during capture (audio-

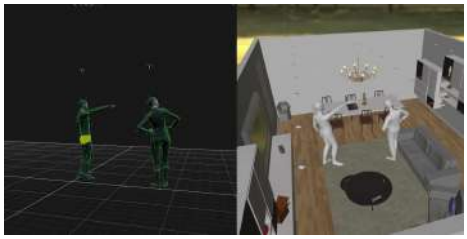


Figure 1: The multimodal data collection environment. A participant in a full-body motion capture suit and VR headset (left) interacts within a virtual scene. Their actions are rendered in real-time in the AI2-THOR simulator (right), enabling the synchronized capture of egocentric vision, speech, motion, and 3D scene geometry.

mocap-gaze/face-simulation). The timecode stream was also recorded inside the simulator to ensure frame-accurate association between referring expressions and rendered egocentric frames. This synchronization enables precise mapping from word-level timestamps to simulator frames for RGB, depth, and per-pixel instance masks.

To transform the raw data into a curated benchmark, we developed a multi-stage annotation pipeline.

3.2.1. Speech Transcription

The audio data was transcribed using WhisperX (Bain et al., 2023), which was prompted with a manually transcribed portion of the audio to encourage the inclusion of filled pauses, repeated words, and other aspects of spontaneous speech. The resulting transcripts were then edited by human annotators to ensure precision. The corrected transcripts were then aligned using the CTC-forced-aligner to obtain word-level timings. Further information on speech transcription can be found in the Appendix [1]

3.2.2. Reference Annotation

We developed a two-stage pipeline using GPT-4o to generate visually grounded reference annotations. First, GPT-4o produced topic annotations for each utterance—based on WhisperX VAD segmentation and the visible objects in the Unity scene—highlighting discourse focus and shifts in attention. Next, GPT-4o parsed the utterances to classify referring expressions into three categories: full noun phrase, partitive/attribute noun phrase, and pronominal. All GPT-4o classifications were manually verified and corrected where necessary by the research team, with most corrections involving boundary cases between partitive and pronominal types.

Categories.

Full noun phrases (full NP) are explicit, uniquely identifying descriptions (e.g., “the black sofa in the corner”).

Partitive/attribute noun phrases (partitive NP), used here as a cover term for underspecified references, refer to parts, features, subsets, salient attributes, or spatial/deictic indicators of an object (e.g., "the cloud (in the painting)", "the rubber thing (part of the lamp)", "the yellowish one", "there") that are insufficiently specific to uniquely identify a referent without additional context. We subsume spatial indicators under this category rather than introducing a separate type, as they share this defining property of underspecification.

Pronominal references (pronouns) are expressions like “it”, “that”, or “those”, which rely heavily on discourse context and shared perceptual attention.

These categories align with established distinctions in referring expression research: full NPs as definite descriptions (Dale and Reiter, 1995; Kazemzadeh et al., 2014), partitive/attribute NPs for part-whole and property-based reference (van der Sluis and Krahmer, 2001; Viethen and Dale, 2008), and pronominal references for deictic or anaphoric pronouns (Hobbs, 1978; Staudte and Crocker, 2011). Together, they span a continuum from lexically explicit to contextually dependent reference.

Grounding and validation. We grounded each expression to specific scene objects using raycasting and Unity-derived object masks. All links between referring phrases and referents were manually verified for correctness, yielding rich, context-sensitive annotations that combine linguistic and perceptual cues.

3.3. Final Dataset Format

To build a usable multimodal dataset for grounded language modeling, we aligned the referring expressions with the visual stream via word-level timestamps, yielding the exact frame where each reference occurs. For that frame, we extract synchronized egocentric assets: (1) an RGB image, (2) a per-pixel metric depth map, and (3) a segmentation mask with per-pixel object IDs. Frames are rendered in Blender for high-quality egocentric imagery, and the pipeline’s flexible camera

[1] For word-level alignment, we used a CTC forced aligner: https://github.com/MahmoudAshraf97/ctc-forced-aligner.

setup supports alternative viewports (e.g., third-person views) for future tasks. This format grounds language in both semantics and spatial structure, supporting reference resolution, object localization, and multimodal grounding in 3D settings.

(a) Visual egocentric RGB scene rendered from the main speaker viewpoint for reference 'box'.

(b) Segmentation mask.

(c) Target object.

Figure 2: A grounded data sample from the benchmark. For a referring expression like "box", our dataset provides synchronized data streams: (a) The egocentric RGB view with the ground-truth referent (box) highlighted by a green segmentation mask. (b) The raw RGB image. (c) The corresponding depth map. This structure facilitates evaluation on precise, pixel-level grounding.

3.4. Dataset statistics

Our dataset comprises a total of 4,211 referring expressions collected over 6.7 hours of recorded interaction. Each expression was annotated for linguistic form, visibility, and grounding success. Overall, virtually all referring expressions were successfully matched to objects in the environment, demonstrating the high referential clarity of the collected data. After filtering for expressions directed at objects not visible at the moment of reference, a final set of 4,001 references was established for analysis. The data also reveals natural variations in referential strategies across different participants and recording environments (Tables 3 and 4), reflecting diverse interaction styles and the trade-offs between scene complexity and referential clarity.

A key finding is the prevalence of context-dependent language, which is often underrepresented in existing multimodal datasets. As shown in Table 5, pronominal references (e.g., "it," "that one") are especially common. In the main (egocentric) view, pronouns account for 2,078 expressions (49.3% of 4,211 total; 1,591 single + 487 multiple). In the inter (partner) view, pronouns are even more dominant, totaling 1,070 expressions (69.2% of 1,547; 901 single + 169 multiple). This high frequency underscores the central role of pronouns in natural dialogue and highlights a core challenge for current models, which often struggle with such context-dependent forms. By explicitly including and annotating these expressions, our dataset fills a critical gap and enables the development of grounded models that better capture the fluid, context-dependent nature of situated language. The categorization of expression types was obtained using GPT-4o (OpenAI, 2024), guided by a custom prompt (described in the Appendix) and manually verified for correctness.

Beyond overall rates, the two perspectives reflect distinct discourse roles. Main speakers more often introduced objects into the conversation, yielding relatively more full mentions, whereas the interlocutor tended to react to already-salient items, relying heavily on pronominal and deictic forms to maintain common ground (Table 5).

Table 3: Aggregate and participant-wise reference statistics for main speaker (m) and interlocutor (i).

pid	# files	Total Refs (m)	Total Refs (i)
All	54	4211	1,547
3	8	432	220
4	13	974	296
5	12	925	391
6	10	1008	322
7	11	872	318

Table 4: Aggregate and room-wise reference statistics

Room	Total Refs	Invisible	Unique Objs (avg)
All	4211	210	23.74
209	807	18	20.42
210	942	51	22.75
211	726	33	23.22
222	844	37	30.38
227	892	71	24.00

Table 5: Reference distribution by type and interlocutor (single/multiple).

	Main			Inter		
	full	part	pron	full	part	pron
Single	1215	435	1591	248	48	901
Multiple	348	135	487	137	44	169

4. Experiments

To gain a comprehensive understanding of referential grounding, we adopt a dual evaluation strategy combining crowd-sourced human judgments and vision-language model (VLM) evaluation. Human studies serve not only to benchmark model performance but also to validate the interpretability of our dataset. However, it is important to note that crowdsourcing takes place outside the original interaction context. As such, crucial 3D spatial relationships, scene dynamics, and nonverbal cues available to the original speaker and listener may be absent. Complementing this, we evaluate state-of-the-art VLMs (GroundingGPT, Ferret, Kosmos-2, Florence-2, and Qwen2.5-VL) to understand how well models generalize to open-ended, conversational referring expressions. Together, these evaluations shed light on the challenges of reference resolution in both human and computational agents.

4.1. Experimental Setup

We split our study into subsets based on the referential expression categories in Table 5. We focus on single-object references, which we further divide into 3 subsets, based on the referring expression categories.

1. **Exact noun phrases:** Explicit object names resembling classic referring expression benchmarks, where grounding depends on direct lexical match.
2. **Filtered partitives:** Partitive NPs with abstract or spatial terms (e.g., "the area," "there") removed, retaining only those grounded in identifiable scene elements.
3. **Subsampled pronominals:** A representative subset of pronouns requiring discourse or visual context for correct resolution.

Evaluation Metrics. We evaluate visual grounding using Intersection over Union (IoU) between predicted and ground-truth bounding boxes, with ground-truth boxes derived from instance segmentation masks. We report accuracy at two standard thresholds, Acc@0.3 and Acc@0.5, as well as mean IoU across samples.

4.2. Human Evaluation

Human evaluations were conducted using the crowd-sourcing platforms Prolific (Prolific, 2024) and Cognition.run (Run, 2024). Participants were presented with the first-person-view images and corresponding utterance with highlighted referring expressions, and were asked to click on the referred object in the picture (see Figure 3). We tracked the use of exact noun phrases, partitive noun phrases, and pronoun-based references under the single object condition, while also comparing isolated utterances with context-based utterances. In order to add context, we sampled the previous five utterances, selecting only those that matched the topic; in the event that this did not yield any text, we selected text from the word-level transcriptions for the previous 20 seconds. This contextual history was then added to the utterance text. Attention checks were included to ensure data quality.

Figure 3: The interface for our human evaluation study. Crowd-workers were presented with an egocentric image and a corresponding utterance from the dataset. They were tasked with clicking on the object being referred to, providing a human baseline for reference resolution.

Each of the 1940 stimuli was evaluated by three participants. The total number of participants included in the analysis, after discarding those 4.9% participants who failed attention checks, is 78.

The percentage of clicks that are within the bounding box around the ground-truth image mask was recorded as accurate. A three-way majority agreement was also calculated for each of the data subsets.

Table 6: Crowd-sourcing results for different types of referring expressions.

Subset	**Acc**	Majority Agreement	Median time (m)
full np, w ctx	62.45%	60.92%	18:09
full np, no ctx	73.18%	74.43%	16:19
part np, w ctx	60.99%	61.96%	28:09
part np, no ctx	47.93%	47.01%	15:56
pron, w ctx	55.42%	55.16%	25:12
pron, no ctx	37.43%	33.33%	18:28

The crowd-sourced evaluation (Table 6) reveals clear trends across expression types and context conditions. Full NPs without context achieved the highest accuracy (73.18%) and agreement

(74.43%), indicating that explicit references are easiest to resolve. Adding context slightly reduced both (62.45%, 60.92%), suggesting that additional information can introduce ambiguity. The smaller difference in median completion time between with- and without-context conditions further implies participants relied less on context when full NPs were explicit. Context substantially benefits underspecified forms. Partitive NPs improved from 48% to 61% accuracy, while pronominals rose from 37% to 55%. Humans achieved non-trivial accuracy on pronominals even without linguistic context (37%), likely by leveraging visual cues such as the speaker's pointing gesture visible in the egocentric view. In contrast, VLMs often grounded to the hand itself or nearby salient regions rather than following the deictic gesture to the intended referent.

Inter-Annotator Reliability. We assessed annotation consistency using object-level agreement with a 30px clustering tolerance. Annotators achieved 75–83% unanimous agreement across conditions (Krippendorff's α = 0.42–0.66), with tight localization (24–44px) when identifying the ground-truth referent. Unanimous misses—where all annotators clicked outside the target mask—fell into three patterns: near-misses due to conservative mask boundaries (15–20%), coherent agreement on alternative referents reflecting genuine ambiguity ($\sim$30%), and scattered responses indicating true confusion (24–51%, highest for pronouns without context). The median pairwise distance among misses (11–64px) remains far below chance ($\sim$270px), indicating consistent interpretations even when diverging from ground truth.

4.3. VLM evaluation

To understand how vision-language models handle natural language grounding in situated interaction, we evaluated several state-of-the-art approaches on our benchmark. We compared end-to-end VLMs against a novel two-stage pipeline designed to explicitly handle conversational context through linguistic disambiguation prior to visual grounding.

4.3.1. Baseline VLM Performance

Table 7 presents baseline results for five VLMs: GroundingGPT (Li et al., 2024), Ferret (You et al., 2024), Kosmos-2 (Peng et al., 2024), Florence-2 (Xiao et al., 2024), and Qwen2.5-VL (Bai, 2025). We evaluated each model in two conditions: without context (only the referring utterance) and with context (previous 5 topic-matched utterances or 20 seconds of dialogue).

Full NPs achieve moderate success, with Qwen2.5-VL reaching 53.2% with context. Context provides minimal benefit (+2 points), as these expressions are already self-contained. Partitive NPs prove challenging for all models (14–30% range). Qwen2.5-VL achieves 29.6% with context, but this remains far below human performance (61% in our crowd-study). Models struggle to infer part-whole relationships from conversational cues alone.

Pronominal references reveal the starkest limitation. Without context, all models fail catastrophically (5–9%), as expected for expressions like "it" or "that." With context, performance improves to 9–30%, but remains poor. Even Qwen2.5-VL, which can explicitly process dialogue history, achieves only 30.4% on pronominals—indicating that simply providing context to VLMs is insufficient.

4.3.2. Two-Stage Grounding Pipeline

The baseline results expose a critical gap: VLMs struggle to leverage conversational context for ambiguity resolution. We hypothesize this stems from conflating two distinct challenges—linguistic disambiguation and visual localization—in a single forward pass. To address this, we propose a modular pipeline.

In Stage 1 (Contextual Rewriting), Qwen2.5-VL rewrites ambiguous references into explicit descriptions (e.g., "it" $\rightarrow$ "the wall painting", "the figures" $\rightarrow$ "the figures on the rug"). The rewriter resolves pronouns using dialogue history, adds part-whole relationships for underspecified expressions, and preserves disambiguating qualifiers. Importantly, the rewriter sees only text—no images—making this pure linguistic reasoning.

In Stage 2 (Visual Grounding), the disambiguated expression is passed to a visual grounding model (Florence-2, GroundingDINO, or Qwen2.5-VL) without fine-tuning. Critically, GroundingDINO—a pure detector with no conversational capabilities—can now successfully ground references it could not handle in the baseline condition.

Rewrite Quality Validation. Validation of rewrites shows 83.4% correctly preserve referent identity, with grounding success strongly correlating with rewrite quality: good rewrites achieve 55.0% (IoU$\geq$0.5), borderline 32.5%, bad 23.4%— confirming genuine disambiguation rather than fortuitous errors. Error analysis reveals systematic patterns (wrong object: 7.5%; missing part relations: 20.3%; details in Appendix C).

4.3.3. Rewriting Results

Table 8 presents results for our two-stage pipeline. The improvements are substantial and consistent across all models and reference types.

Pronominals show the most dramatic improvements. GroundingDINO jumps from approximately

Table 7: VLM Grounding Performance. Match Rate at IoU 0.5/0.3 (M@.5/M@.3) and mean IoU for matched samples (mIoU@.5). Best M@.5 per row in **bold**.

Expr. Type	Context	GroundingGPT	Ferret	Kosmos-2	Florence-2	Qwen2.5-VL
Full NP	w/o ctx	37.9/47.3	39.1/45.2	42.6/48.1	46.3/52.6	**51.8**/60.9
	w/ ctx	28.2/41.1	40.9/47.9	42.0/48.2	28.0/33.1	**53.2**/62.0
Partitive NP	w/o ctx	14.2/16.3	14.7/16.7	16.5/17.9	16.3/18.8	**22.5**/26.4
	w/ ctx	14.7/20.9	18.3/21.6	19.7/22.2	21.6/23.6	**29.6**/34.2
Pronominal	w/o ctx	4.7/6.9	5.1/7.4	6.3/9.1	**9.2**/10.1	5.9/8.3
	w/ ctx	10.7/16.0	12.3/15.4	9.3/11.4	25.8/28.5	**30.4**/38.0

Table 8: Rewrite-based Grounding Results. Referring expressions are rewritten by Qwen2.5-VL then grounded. Match Rate at IoU 0.5 (M@.5) shown with improvement over baseline in parentheses.

	Full NP		Partitive NP		Pronominal	
Model	Baseline	Rewrite	Baseline	Rewrite	Baseline	Rewrite
Qwen2.5-VL	53.2	54.4 (+1.2)	29.6	40.8 (+11.2)	30.4	50.3 (+19.9)
Florence-2	28.0	49.1 (+21.1)	21.6	39.7 (+18.1)	25.8	48.9 (+23.1)
GroundingDINO	—*	**61.1**	—*	**49.5**	—*	**56.7**
Avg. Gain†	+11.2		+14.7		+21.5	

*No conversational capability; baseline not applicable.
†Average gain for Qwen2.5-VL and Florence-2 only.

5% to 56.7%, Florence-2 from 25.8% to 48.9%, and Qwen2.5-VL from 30.4% to 50.3%. The massive gain for GroundingDINO is particularly revealing: this model has no ability to process conversational context, yet after rewriting it outperforms Qwen2.5-VL's baseline with context (56.7% vs 30.4%). This demonstrates that the bottleneck is not visual grounding capability, but rather the conflation of linguistic and visual reasoning.

Partitive NPs also benefit substantially. GroundingDINO improves from approximately 15% to 49.5%, Florence-2 from 21.6% to 39.7%, and Qwen2.5-VL from 29.6% to 40.8%. The rewriter adds explicit part-whole relationships that models struggle to infer from conversation alone. GroundingDINO again achieves the highest absolute performance (49.5%), nearly doubling Qwen's baseline.

Full NPs show varied patterns. Qwen2.5-VL shows minimal gain (+1.2%) because it already handles explicit references well (53.2% baseline). However, Florence-2 and GroundingDINO benefit substantially from rewriting, which removes object IDs, preserves qualifiers, and adds spatial context.

Remarkably, GroundingDINO with rewriting achieves the highest performance across all categories (61.1% / 49.5% / 56.7%), outperforming Qwen2.5-VL's best baseline with context (53.2% / 29.6% / 30.4%). This is despite GroundingDINO having no native ability to process conversational history. Mean IoU remains high (77–89%) across all models, indicating that when models successfully identify the correct object, they localize it precisely. The challenge is identification, not localization—which our rewriting approach directly addresses.

5. Discussion and conclusions

In this work, we present a multimodal benchmark for situated referential communication, combining spontaneous VR dialogue with synchronized 3D scene data. This enables analysis of grounding in realistic, multimodal contexts beyond prior datasets.

A human text-only evaluation established a lower bound for grounding performance. Participants easily resolved explicit noun phrases but relied on dialogue history for ambiguous partitive and pronominal references. By removing non-verbal cues, this setup provides a fair baseline for comparison with current vision–language models (VLMs).

Compared to this baseline, state-of-the-art VLMs such as Florence-2 and GroundingGPT exhibit a clear domain gap: trained on self-contained captions, they struggle with the spontaneous, ambiguous language dominant in our data. With nearly 50% of references being pronominal, this failure affects the majority of situated expressions. The gap between human text-only and full multimodal performance reveals a larger *multimodal gap*, the benefit humans gain from gaze and gesture cues, highlighting our dataset's unique potential to close it.

Our two-stage pipeline, which separates linguistic reasoning from visual localization, outperforms monolithic VLMs and shows that resolving ambiguity prior to grounding yields more robust results. Analysis further shows that:

- Conversational history benefits both humans and models, especially for ambiguous references.
- Added context can slightly lower match rates but increase localization precision (IoU), indicating a trade-off between recall and spatial accuracy.
- Even when semantically off, VLM predictions remain spatially close to ground truth.

While effective, our model introduces latency. A practical next step is teacher–student distillation, where GPT-4o converts ambiguous expressions into explicit, grounded forms to train faster, deployable models.

Our current 2D text–image evaluation forms only the first layer of analysis. Future work will exploit synchronized temporal and non-verbal streams to enable: (1) video-based grounding over unfolding interactions, (2) gaze-conditioned attention mechanisms, and (3) full 3D scene reasoning for embodied references.

Ultimately, progress in multimodal grounding requires moving beyond static image–text alignment toward models capable of temporal reasoning, dialogue-aware interpretation, and multimodal integration. By providing a rigorous baseline and identifying key gaps, our benchmark lays the groundwork for the next generation of situated AI systems. The full dataset, with synchronized motion, speech, and gaze, will be publicly released to support advances in both grounding and embodied behavior modeling.

6. Bibliographical References

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. 2020. ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 422–440. Springer.

S. Bai. 2025. Qwen2.5-vl technical report. *arXiv*.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-accurate speech transcription of long-form audio. In *Interspeech 2023*, pages 4489–4493, Dublin, Ireland. ISCA.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. 2020. ScanRefer: 3D object localization in RGB-D scans using natural language. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*, page 202–221, Berlin, Heidelberg. Springer-Verlag.

Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. 2021. Yourefit: Embodied reference understanding with language and gesture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1385–1395.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Anna Deichler, Siyang Wang, Simon Alexanderson, and Jonas Beskow. 2023. Learning to generate pointing gestures in situated embodied conversational agents. *Frontiers in Robotics and AI*, 10:1110534.

Fethiye Irmak Dogan, Maithili Patel, Weiyu Liu, Iolanda Leite, and Sonia Chernova. 2025. A model-agnostic approach for semantically driven disambiguation in human-robot interaction. In *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 649–656.

Fethiye Irmak Doğan, Ilaria Torre, and Iolanda Leite. 2022. Asking follow-up clarifications to resolve ambiguities in human-robot conversation. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 461–469.

Stephanie Gross, Brigitte Krenn, and Matthias Scheutz. 2017. The reliability of non-verbal cues for situated reference resolution and their interplay with language: implications for human robot interaction. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI '17, page 189–196, New York, NY, USA. Association for Computing Machinery.

Jerry R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), pages 787–798. Association for Computational Linguistics.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. 2017. AI2-THOR: An interactive 3D environment for visual AI. *arXiv preprint arXiv:1712.05474*.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. 2024. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6657–6678.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer.

OpenAI. 2024. Gpt-4o technical report. https://openai.com/research/gpt-4o. Accessed: 2025-04-10.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. TEACh: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. Grounding multimodal large language models to the world. In *International Conference on Learning Representations*, volume 2024, pages 51575–51598.

Prolific. 2024. Prolific. https://www.prolific.com. Version used: April 2025. First released in 2014. Prolific, London, UK.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Cognition Run. 2024. Cognition.run. https://www.cognition.run. Platform for running online behavioral experiments. Accessed: April 12, 2025.

Allison Sauppé and Bilge Mutlu. 2014. Robot deictics: how gesture and context shape referential communication. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '14, page 342–349, New York, NY, USA. Association for Computing Machinery.

Mohit Shridhar, Dixant Mittal, and David Hsu. 2020. Ingress: Interactive visual grounding of referring expressions. *The International Journal of Robotics Research*, page 0278364919897133.

Maria Staudte and Matthew W. Crocker. 2011. Investigating joint attention mechanisms through spoken human–robot interaction. *Cognition*, 120(2):268–291.

Ielka van der Sluis and Emiel Krahmer. 2001. Generating referring expressions in a multimodal context: An empirically oriented approach. In *Computational Linguistics in the Netherlands 2000: Selected Papers from the Eleventh CLIN Meeting*, pages 158–176. Rodopi.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the 5th International Natural Language Generation Conference*, pages 59–67. Association for Computational Linguistics.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2024. Ferret: Refer and ground anything anywhere at any granularity. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zhuoyang Zhang, Kun Qian, Bo Zhou, Fang Fang, and Xudong Ma. 2024. Gaze-assisted visual grounding via knowledge distillation for referred object grasping with under-specified object referring. *Engineering Applications of Artificial Intelligence*, 133:108493.

7. Language Resource References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate

speech transcription of long-form audio. https://github.com/m-bain/whisperX.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. 2017. AI2-THOR: An interactive 3D environment for visual AI. https://ai2thor.allenai.org/.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, Zhida Huang, and Tao Wang. 2024. Groundinggpt: Language-enhanced multimodal grounding model (code/models). https://github.com/lzw-lzw/GroundingGPT.

MANUS. 2023. MANUS quantum metagloves. https://www.manus-meta.com/products/quantum-mocap-metagloves.

OptiTrack. 2023. Optitrack motion capture system. https://optitrack.com/.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Unified representation for vision tasks (code/models). https://github.com/microsoft/Florence-2.

7.1. Resource to be Shared

We release **MM-Conv**, a multimodal benchmark for context-aware grounding in situated 3D dialogue. The resource comprises synchronized *speech*, *text*, *egocentric RGB*, *metric depth (normalized for the starter pack)*, *segmentation masks*, *full-body motion*, *gaze*, *facial blendshapes*, and *3D scene geometry*.

Scope and Scale.

- Duration: ~6.7 hours of dyadic interaction (VR).
- Instances: 4,211 referring expressions (after filtering: 4,001 visible).
- Linguistic forms: full NP, partitive/attribute NP, pronominal.
- Speech: ~250k transcribed words with word-level timecodes.
- Visual: Five AI2-THOR apartment rooms, egocentric frames with per-pixel masks.

Modality Summary.

Modality	Details
Speech/Text	WhisperX transcripts+word timings
RGB/Depth	Egocentric images+normalized depth
Segmentation	Per-pixel object IDs and GT masks
Motion	Full-body (main+interlocutor)
Hands/Face	Finger tracking, 52 facial blendshapes
Gaze	Binocular gaze from HMD
Scene Graph	Object instances and 3D metadata

Table 9: Modalities included in the resource.

Format. Each RE is aligned to its egocentric frame; we provide JSON annotations (utterance, token indices, category, referent ID), RGB, mask, depth (normalized in the sample pack), and metadata.

Availability and License. The resource will be made available for research use upon publication under **CC BY-NC 4.0**. A *starter pack* ($\leq$20 MB) with schemas, 10–20 samples, and an evaluation script is provided for review. The full release (with documentation and datasheet) will follow camera-ready.

Ethics and Privacy. Data were collected under informed consent; audio is transcribed and anonymized; egocentric renders are from simulation (no real faces). We remove metadata that could reveal identity or location. Redistribution is limited to non-commercial research; re-identification is prohibited.

7.2. Resources Used in This Work

We used the following LRs/tools; each is (or will be) separately entered in the LRE Map:

- **AI2-THOR** simulator for interactive 3D environments Kolve et al. (2017).
- **WhisperX** for ASR with word-level alignment (Bain et al., 2023).
- **Florence-2** (tool) (Xiao et al., 2024); paper (Xiao et al., 2024).
- **GroundingGPT** (tool) (Li et al., 2024); paper (Li et al., 2024).
- **OptiTrack** motion capture system (OptiTrack, 2023).
- **MANUS Quantum MetaGloves** for finger tracking (MANUS, 2023).

7.3. Reproducibility and Evaluation Artifacts

We provide a minimal evaluation script (IoU, Match@{0.3,0.5}) and JSON schemas for predictions/ground truth to standardize reporting. Benchmarks include single-object full/partitive/pronominal subsets with and without dialogue context.

A. Data Collection and Experimental Setup

This section details the methodology for collecting our multimodal dataset. We combine motion capture with virtual reality, recording conversations within the AI2-THOR physics simulator[2]. This setup provides experimental control for replicating conditions precisely and simplifies the annotation of objects and scenes.

A.1. Experimental Scenarios

To elicit natural, spontaneous referential language, participants engaged in one of three conversational scenarios within the virtual environment. The setup involved a main actor (the "speaker," wearing the VR headset) and a secondary actor (the "interlocutor"). The scenarios were designed to encourage descriptive language and interaction centered on objects in the scene.

- **Scenario 1: Bragging/Introducing New Apartment**
  - *Roles:* Main Actor: Apartment Owner, Secondary Actor: Friend.
  - *Description:* The apartment owner shows off their new apartment to their friend, highlighting various features and objects in the room.
  - *Dialogue Focus:* Description of the objects, personal anecdotes about their acquisition, and the benefits of each item.
- **Scenario 2: Landlord Asking About Objects**
  - *Roles:* Main Actor: Landlord, Secondary Actor: Tenant.
  - *Description:* The landlord inquires about various objects in the apartment, possibly checking for maintenance needs or understanding the tenant's living conditions.
  - *Dialogue Focus:* Questions about the objects, their usage, condition, and any issues.
- **Scenario 3: Interior Designer Giving Tips**
  - *Roles:* Main Actor: Interior Designer, Secondary Actor: Client.
  - *Description:* The interior designer provides suggestions and advice on improving the apartment's aesthetics and functionality.
  - *Dialogue Focus:* Suggestions for rearranging furniture, adding new decor items, and making the space more efficient.

[2] https://github.com/allenai/ai2thor

A.2. Scene Details and Object Distribution

The experiments were conducted across five different virtual rooms from the AI2-THOR simulator. The average number of interactable objects in these rooms was 38 ± 3.16. This variety ensures a diverse distribution of objects and spatial layouts. Figure 4 shows the frequency of different object categories across all environments used in the study.

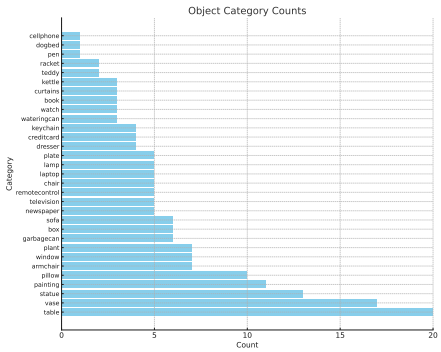


Figure 4: Object category distribution across simulated environments used in the experiments.

A.3. Hardware and Software Setup

Our data capture system integrated several components to record synchronized multimodal data streams. An illustration of the hardware setup is provided in Figure 5.

- **Motion Capture:** Skeletal information was recorded using an OptiTrack system[3] with 16 Prime x 41 cameras, tracking 50-marker skeletons for both participants.
- **Finger Tracking:** Accurate finger movements were captured using Quantum Mocap Metagloves[4].
- **VR and Gaze Tracking:** A META Quest Pro headset[5] was used to immerse the main speaker in the virtual environment and to record face and gaze tracking data.
- **Simulation Environment:** The virtual scenes were rendered in the AI2-THOR physics simulator[6].
- **Synchronization:** Tentacle Sync E devices[7] were used to generate SMPTE time codes, ensuring fine-grained synchronization between the audio recordings, OptiTrack motion data, and AI2-THOR simulation data. The position of the Quest headset was aligned with the motion-captured head position to ensure visual and physical synchrony.

Figure 5: Illustration of hardware setup in motion capture lab.

B. Annotation Pipeline Details

This section details the methodology for annotating our multimodal dataset.

B.1. Speech Transcription and Alignment

To support speech annotation, we used Label Studio[8] to verify and correct the generated transcriptions. Each annotation consisted of a pair of linked elements: a timed audio chunk and its corresponding transcription text box. This interface allowed annotators to listen to speech segments in context and make corrections directly on the aligned text.

We initialized Label Studio using audio chunk boundaries provided by WhisperX's voice activity detection (VAD) module, yielding more accurate and robust segmentation compared to sentence-level heuristics. Each chunk was pre-filled with WhisperX's ASR output, providing a fast starting point for correction. Annotators focused on preserving spontaneous speech characteristics (e.g.,

[3] https://optitrack.com/
[4] https://www.manus-meta.com/products/quantum-mocap-metagloves
[5] https://www.meta.com/quest/quest-pro/
[6] https://github.com/allenai/ai2thor
[7] https://tentaclesync.com/sync-e
[8] https://labelstud.io

filled pauses, repetitions) while ensuring accuracy and clarity.

To assist with triage and quality control, we implemented color-coding in the annotation interface: segments containing non-speech events—such as laughter or spoken noise—were visually highlighted in distinct colors. This allowed annotators to quickly identify and skip or deprioritize these segments during transcription.

As only one of the speakers is a native speaker, annotators also took care to check for effects of accented speech, such as the Spanish native's pronunciation of "vase" being recognized as "base".

Figure 6: Label Studio interface for speech annotation. Annotators reviewed WhisperX transcriptions aligned to VAD-based audio segments and edited text within the linked text box.

B.2. GPT-4o prompts

B.2.1. Reference Classification Prompt

The following prompt was used to classify referring expressions into semantic categories using GPT-4o:

Listing 1: Prompt used for classifying referring expressions into semantic categories using GPT-4o.

```
You are analyzing a sentence and a
    highlighted phrase within it.
The object of interest is called:
    '{object_name}'.

Utterance: '{utterance}'
Referring phrase: '{phrase}'

Your task is to categorize the
    referring phrase into one of the
     following types:
- 'exact': matches or semantically
    refers to the full object name
   (e.g., 'lamp' -> 'lamp', or '
      couch' -> 'sofa')
- 'part': refers to a part, feature
    , or something depicted within
    the object
   (e.g., 'handle' of a 'pan', '
      cloud' in a 'painting', or '
      the text' on a 'poster')
- 'pronominal': is a pronoun (e.g.,
     'this', 'it', 'that one', 'they
    ')

Important note:
- If the phrase refers to something
     inside or depicted on a
    painting, poster, or picture,
    classify it as 'part'.
- For example, 'the dog' referring
    to an image in a painting ->
    label: 'part'
- But if the phrase is 'the
    painting', 'the poster', or 'the
     picture' -> label: 'exact'

Only return the label: 'exact', '
    part', or 'pronominal'.
```

B.2.2. Contextual Disambiguation Prompt

We use Qwen2.5-VL for linguistic disambiguation in Stage 1 of the two-stage pipeline (Section 4.3.2). The rewriter operates in *text-only* mode: no image is provided, so disambiguation is pure linguistic reasoning over dialogue history and the visible-object list. The rewritten noun phrase is then passed to the Stage 2 visual grounding model (Florence-2, GroundingDINO, or Qwen2.5-VL).

Listing 2: Prompt used for contextual disambiguation of referring expressions using Qwen2.5-VL. The model receives conversation history and a visible-object list, and outputs a single noun phrase.

```
SYSTEM: You are a referring-
    expression rewriter for visual
    grounding. Given a phrase, its
    utterance, conversation history,
     and a list of visible objects,
    output a single unambiguous noun
     phrase (3-8 words) that
    identifies the intended referent
    .

TASK: Rewrite the phrase "{phrase}"
     from the utterance into a clear
    , specific noun phrase that can
    be grounded in the visible scene
    .

UTTERANCE: {utterance}

CONVERSATION HISTORY:
{conversation_history}

VISIBLE OBJECTS IN SCENE:
{object_list}

CRITICAL RULES:
1. OUTPUT ONLY A NOUN PHRASE - no
    explanations, no meta-commentary
2. If the original phrase doesn't
    match  visible objects exactly,
    output the CLOSEST semantic
    match
3. Add spatial/relational context
    when helpful (e.g., "the desk
    with the laptop" vs just "the
    desk")
4. Use object attributes from the
    visible objects list when
    available
5. NEVER output: "There is no X
    visible", "The scene does not
    contain", "I cannot see"
6. NEVER use quotation marks,
    backticks,
    or formatting in your output
7. Keep it concise but specific
    (typically 3-8 words)

EXAMPLES:
- "the stove" + visible: [FirePlace
    ] -> "the fireplace"
- "it" + context about painting  +
    visible: [Painting2] -> "the
    wall painting"
- "the little black thing"  +
    context: helicopter painting  ->
     "the black element in the
    painting"
- "cloud" + visible: [
    Painting_0db0fb84]  -> "the
    cloud-like shape in the
    painting"

YOUR REWRITE (noun phrase only):
```

C. Rewrite Quality Validation

Validation of rewrites shows 83.4% correctly preserve referent identity. Performance varies by type: pronominals 87.2%, full NPs 84.6%, partitives 67.4%.

Table 10: Grounding success by rewrite quality.

Quality	Acc@0.5	Acc@0.3	Mean IoU
Good	55.0%	64.7%	0.481
Borderline	32.5%	44.6%	0.323
Bad	23.4%	27.2%	0.216
Overall	**50.5%**	**60.0%**	**0.447**

Error-flag impact (Acc@0.5): wrong_object 7.5% | missing_part 20.3% | lost_qualifier 41.8%. Cross-rewriter: GPT-4o 51.9% vs Qwen 55.0% (3.1pp).

Paper G

Paper G

Look and Tell: A Dataset for Multimodal Grounding Across Egocentric and Exocentric Views

ANNA DEICHLER AND JONAS BESKOW

Published in
Workshop on SPACE in Vision, Language, and Embodied AI (SpaVLE), NeurIPS, 2025

Look and Tell: A Dataset for Multimodal Grounding Across Egocentric and Exocentric Views

Anna Deichler
TMH
KTH Royal Institute of Technology
deichler@kth.se

Jonas Beskow
TMH
KTH Royal Institute of Technology
beskow@kth.se

Abstract

We introduce Look and Tell, a multimodal dataset for studying referential communication across egocentric and exocentric perspectives. Using Meta Project Aria smart glasses and stationary cameras, we recorded synchronized gaze, speech, and video as 25 participants instructed a partner to identify ingredients in a kitchen. Combined with 3D scene reconstructions, this setup provides a benchmark for evaluating how different spatial representations (2D vs. 3D; ego vs. exo) affect multimodal grounding. The dataset contains 3.67 hours of recordings, including 2,707 richly annotated referential expressions, and is designed to advance the development of embodied agents that can understand and engage in situated dialogue.

1 Introduction

Understanding how humans coordinate gaze, gestures, and speech in referential communication is critical for building embodied agents that can interact naturally and exhibit spatial intelligence. While gaze has long been recognized as a cue for attentional focus, most research has emphasized its temporal relationship to speech in controlled tasks. Less attention has been given to how spatial representations of the environment—whether in 2D image space, 3D reconstructions, or across egocentric and exocentric perspectives—influence multimodal spatial grounding.

We present a new dataset collected with 25 participants in a naturalistic kitchen setting, where participants recalled recipe ingredients while wearing Meta Aria smart glasses [Engel et al., 2023]. The glasses provided synchronized gaze and speech streams, complemented by exocentric GoPro recordings. This dual-view setup enables both fine-grained analysis of gaze–speech synchrony and systematic evaluation of representation choices in multimodal grounding. Specifically, the dataset supports comparisons between 2D and 3D scene representations, as well as between egocentric and exocentric perspectives, offering a unique testbed for investigating how humans and embodied models organize spatial knowledge during communication. This is crucial for human–robot interaction, where an agent must understand references from a user's point of view and integrate multimodal signals such as gaze, gestures, and speech. This dual-view paradigm is critical for tackling challenges in **shared autonomy and human-robot collaboration**, where an agent must simultaneously interpret a user's first-person intent while maintaining an objective, third-person model of the world state.

Our contributions are threefold: (1) a multimodal dataset of synchronized gaze, speech, and dual-view video designed to study **situated dialogue**; (2) an annotation pipeline for **aligning spatial concepts across language and vision**; and (3) a new **benchmark for evaluating spatial intelligence** in **grounded communication**. By capturing rich, natural behavior, our dataset provides a foundation for future analysis of how gesture, gaze, and speech jointly enable **multimodal spatial grounding**.

2 Related Work

Gaze and speech in communication. Prior work shows that gaze strongly predicts communicative goals [Hanna and Brennan, 2007, Brennan and Hanna, 2007], and has been combined with gestures or speech to improve reference resolution [Renner et al., 2014]. Event synchronization studies [Kaur et al., 2003] highlight the potential of aligning multimodal streams, though mainly in controlled tasks. Recent work emphasizes gaze in instruction and teaching contexts [Wagner et al., 2023] and in virtual environments [Tanriverdi and Jacob, 2000]. However, temporal alignment of gaze and speech in natural referential tasks remains underexplored.

Multimodal communication datasets. Several datasets integrate multiple communicative channels, such as speech, gaze, and gesture in collaborative tasks [Kontogiorgos et al., 2018]. VENUS [Kim et al., 2025] combines speech, facial expressions, and body pose. In vision–language research, gaze has been used for captioning [Sugano and Bulling, 2016], VQA [Vasudevan et al., 2018, Ilaslan et al., 2023]. Despite this progress, datasets combining first-person gaze with concurrent speech in referential tasks are rare.

Spatial representations for grounding. Recent research highlights the importance of spatial representations for multimodal agents. Ego4D and related egocentric datasets emphasize large-scale video and gaze [Grauman et al., 2022], while 3D QA datasets such as ScanQA [Gao et al., 2021] and EmbodiedQA [Das et al., 2018] explore grounding in reconstructed scenes. Yet, few datasets allow direct comparison between 2D vs. 3D representations or ego vs. exo perspectives in the same task. Our dataset bridges this gap, enabling systematic evaluation of how representational choices affect grounding in situated referential communication.

3 Data Collection & Dataset Description

3.1 Experiment Design

Participants. 25 participants (18 female, 7 male; age range 22–37) were recruited at KTH Kitchen Lab. **Materials.** Aria smart glasses (for eye tracking and audio recording), GoPro cameras (environment capture), recipes, and food ingredients. **Task.** Each participant memorized a step of a recipe and then instructed it to a conversational partner. While speaking, they were asked to identify and refer to the corresponding ingredients in the environment.

Procedure.

1. Participant reads recipe step.
2. Participant recalls step aloud while locating ingredients.
3. Egocentric Aria glasses record synchronized gaze, audio, video, and pointcloud data; exocentric GoPro records audio and video feed.
4. Sessions were recorded across five recipes per participant.

3.2 Dataset Description

The **Look and Tell** dataset, curated by KTH Royal Institute of Technology, offers a unique resource for investigating the interplay between visual attention and spoken language. This dataset was collected using Aria smart glasses, which allowed for the simultaneous and real-time capture of participants' eye-tracking data and speech audio.

The experimental paradigm involved participants identifying various food items from a recipe. This task was designed to elicit natural referential communication, providing rich, ecologically valid data on how individuals visually fixate on objects while verbally describing or identifying them. The recordings from the Aria glasses provide a synchronized stream of high-resolution eye-tracking data and corresponding speech, enabling detailed analysis of gaze-speech synchronization patterns. This is complemented by exocentric GoPro recordings. This dual-view setup enables fine-grained analysis of gaze–speech synchrony and evaluation of representational choices in multimodal grounding (see Fig. 1).

Table 1: KTH-ARIA Referential Dataset Summary Statistics

Participants / Sessions	25 participants / 125 unique sessions
Total RGB Frames	396,208 frames
Total Duration	3.67 h (220.1 min) at 30 fps
Frames per Session	Mean: 3,169.7 $\pm$ 1,328.9 Range: 1,223–7,730
Duration per Session	Mean: 1.8 min $\pm$ 0.7 min Range: 0.7 min–4.3 min

3.3 Data Modalities and Format

- **Eye-tracking data:** fixation events, gaze vectors.
- **Speech audio:** raw speech signal, later transcribed and segmented with WhisperX.
- **Video recordings:**
  - Egocentric view: raw video from the head-mounted Aria cameras (MP4 format), recorded at up to 1408×1408 pixels and 30 fps, synchronized with audio and gaze.
  - Exocentric view: side-view video recordings from stationary GoPro cameras, providing third-person context of the participant and environment.
- **Pointcloud representation** of the scene.
- **Recipe and ingredient metadata**, including surface positions.

3D Room Reconstruction. In addition to synchronized gaze and speech recordings, we also obtained reconstructed 3D models of the kitchen environment from separate video sessions. Using the Meta Project Aria MPS service [Meta Reality Labs Research, 2023], we extracted image frames from room recordings and generated point clouds that were then canonicalized into a shared coordinate system (Fig. 5). This process aligns multiple reconstructions from different recording sessions, yielding a consistent 3D representation of the environment. These reconstructions provide additional spatial context beyond 2D video, enabling research on how embodied agents can exploit 3D scene structure for referential grounding and multimodal interaction.

3.4 Dataset Statistics

We recruited 25 participants (aged 22–37; 18 female, 7 male). Across 125 sessions, the dataset contains a total of **396,208 RGB frames** ($\sim$3.67 hours at 30 fps). Sessions lasted on average 1.8 ± 0.7 minutes (range 0.7–4.3), corresponding to $3{,}170 \pm 1{,}329$ frames per session (range 1,223–7,730). The longest session was 4.3 minutes and the shortest 0.7 minutes. See Table 1 for details.

Participants reported the following native languages: Korean (1), Indian languages (2), Icelandic (1), Chinese Mandarin (9), Spanish (2), Swedish/English bilingual (1), Swedish (4), Turkish (1), Nigerian (1), Indonesian (1), Portuguese (1), and Russian (1).

4 Annotation Pipeline and Analysis

To enable fine-grained analysis of gaze–speech synchrony, we developed a multi-stage annotation pipeline that integrates audio transcription, language-based reference extraction, and multimodal object detection.

Audio acquisition. Audio data was obtained from the **Meta Project Aria MPS service**, which allowed us to request synchronized audio streams from the original `.vrs` files alongside the corresponding image frames. The images were pre-processed to remove fisheye distortion, yielding undistorted RGB frames aligned with the audio.

Speech transcription. We transcribed the audio using **WhisperX** [Bain et al., 2023], which provides robust automatic speech recognition together with **word-level timestamps**. These timestamps form the backbone of our temporal alignment between speech and other modalities.

Figure 1: Example of synchronized video feeds. The exocentric (left) view provides situational context, while the egocentric (right) view captures the participant's first-person perspective, including their gaze target (green circle, overlaid for visualization).

Figure 2: Example of reconstructed 3D room point cloud used as the canonical reference space. All point clouds extracted from Aria recordings are aligned and canonicalized to this shared coordinate system.

Reference extraction. Using the transcribed text, we employed **GPT-based prompting** to identify mentions of ingredients and distractor objects referenced in the recipe instructions, as well as additional objects spontaneously mentioned by participants [OpenAI, 2023]. Here, we use *additional object* to denote non-ingredient items from the recipe/distractor lists that participants sometimes referred to explicitly (e.g., *fridge*, *sink*, *stove*). This step produced a set of candidate **mentions** linked to precise word spans in the transcripts. The full prompting templates used for this step are provided in the Appendix.

4.1 GPT-based Mention Linking

To process raw speech transcripts into structured data for multimodal analysis, we automatically convert WhisperX word-level transcripts into span-level ingredient, utensil, and object mentions using a hybrid pipeline (Fig. 3). This process leverages a large language model augmented with deterministic rules to ensure consistency.

Inputs. The pipeline uses three sources of information: (a) tokenized transcripts with word-level timings from WhisperX, (b) recipe metadata and ingredient lists from `recipes.json`, and (c) a pre-defined set of scene distractors (which are ingredients not used in current recipe).

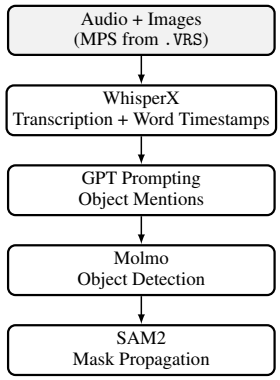


Figure 3: Annotation pipeline for the egocentric camera: synchronized audio and images are processed with WhisperX for word-level transcripts, GPT for object mentions, Molmo for object detection, and SAM2 for mask propagation.

LLM Labelling. We prompt a GPT model with the lower-cased transcripts and a list of candidate items (ingredients + distractors), constraining its output to a JSON schema. The model identifies and labels mention spans, providing attributes such as match type, confidence, and optional coreference links. Coreference resolution for pronouns and demonstratives (*it, this, that, them*) is guided by nearby cooking-related actions.

Post-processing. To ensure accuracy and consistency, we apply a series of deterministic rules: (1) normalization of predicted strings to canonical recipe or distractor names; (2) alias enrichment to handle synonyms (e.g., *tap* $\rightarrow$ *sink*); (3) precise alignment of timings and surface text; and (4) construction of a mention graph with unique IDs and antecedent links for coreferences. Automated checks are used to flag unmapped items, missing antecedents, or timestamp errors.

Full details on the prompt schema, processing rules, and handling of edge cases are documented in Appx. A.1–A.

4.2 Object detection and tracking.

For each candidate mention, we applied **Molmo** Deitke et al. [2024], a multimodal vision–language model, to localize the referenced item (ingredient or object; see definition above) in the undistorted frames. We prompted Molmo with concise imperatives (e.g., *"Point to the {description} object."*) and evaluated it on frames sampled from the mention interval, using at least $\max(n_{\text{frames}}, n_{\text{min}})$ frames. Molmo returned 2D point coordinates (in percent of image width/height), which we converted into seed locations to initialize the **SAM2**Ravi et al. [2024] tracker. SAM2 then propagated segmentation masks across the full mention interval, yielding dense per-frame masks aligned to each mention. When Molmo did not yield a reliable localization, we performed *manual annotation* to seed or correct the tracker, particularly for tiny items (e.g., spice containers) and visually similar look-alikes (e.g., sugar vs. wheat-flour jars; see Fig. 4). In total, 747 mentions required manual handling: 106 cases were skipped because the object was not visible in the frames, while 641 mentions were manually annotated after inspection identified them as either too small (Molmo never reliable) or belonging to ambiguous object categories.

Together, the pipeline yields synchronized **speech transcripts**, **object mentions** (tokens $\rightarrow$ mentions $\rightarrow$ chains), **frame-level points**, **bounding boxes**, and **per-frame masks** for each referential episode.

4.3 Exo and Ego Spatial and Temporal Alignment

Spatial alignment. To register egocentric Aria trajectories to the COLMAP world frame, we localize sparse Aria frames using hloc Sarlin et al. [2019] with SuperPoint features and LightGlue

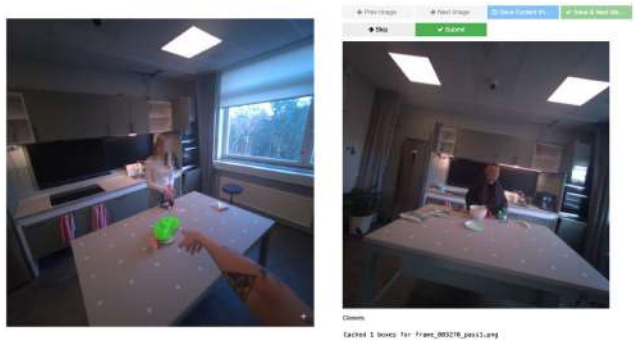


Figure 4: Examples of annotation strategies: (a) Molmo detection based automated SAM2 mask propagation, (b)manual seeding of tracker with annotation interface.

matching against the reconstructed 3D model. We then estimate a Sim(3) transformation (scale, rotation, translation) per recording via RANSAC-based Umeyama alignment Umeyama [1991]. Spatial alignment between egocentric Aria trajectories and the COLMAP world model achieved $\geq$70% inlier correspondence for 109 of 125 sessions (87%), with mean alignment error of 9.1 cm. Sessions with lower alignment quality are retained for analyses that do not require precise 3D localization (e.g., gaze-speech synchrony, 2D tracking).

Temporal alignment. Ego and exo video streams were synchronized by computing the cross-correlation of their audio tracks, sampled at 16 kHz. The detected offset was used to trim or pad the exo stream to align with the ego reference, producing synchronized outputs at a constant 30 fps.

4.4 Human Mesh Recovery from Exocentric Videos

Human motion is recovered from static exocentric GoPro recordings captured at 30 fps, providing a stable third-person view of the participant. Camera poses and intrinsics are obtained from a COLMAP Schönberger [2018] reconstruction and define the shared world coordinate system. For each exocentric frame, we estimate a frame-wise 3D human mesh using SAM-3D-Body Yang et al. [2025] conditioned on the calibrated camera intrinsics. Each mesh is then lifted into the COLMAP world frame using the corresponding camera pose. To ensure physically plausible placement, meshes are vertically aligned by projecting foot keypoints to the floor plane and correcting the root translation accordingly. To obtain temporally consistent motion tracks, we perform explicit identity tracking in world space. Person identities are initialized by lifting detected foot keypoints onto the estimated floor plane and clustering detections by spatial proximity. The final output consists of per-person, temporally ordered mesh sequences in world coordinates, including mesh vertices, joint trajectories, and identity metadata, enabling joint analysis of human motion and scene geometry.

4.5 Analysis of Mention Annotations

Across all 25 participants and 125 sessions, we annotated **2,707 mentions** ($\sim$22 per session). Table 2 shows the distribution: ingredients dominate (62%), while **23% are pronouns or coreferential forms**, highlighting the prevalence of indirect reference. Objects (6%) and distractors ($< 2\%$) provide additional ecological variation.

Coreference chains averaged **6.7 mentions**, indicating repeated reference to the same item once introduced. Ingredient coverage was high: nearly all target items appeared at least once (mean coverage $> 90\%$).

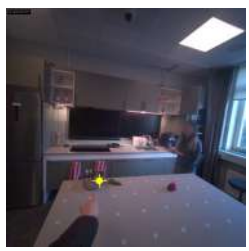
Figure 5: Visualization of aligned ego trajectories (blue), exo camera poses (green), human meshes, and 3D scene reconstruction in a shared world coordinate system using viser Yi et al. [2025].

Mention type	**Count**	**Proportion**
Ingredients	1,680	62.1%
Pronoun/coref	614	22.7%
Additional objects	154	5.7%
Distractors	28	1.0%
Other (utensils, unresolved)	231	8.5%
Total	2,707	100%

Table 2: Distribution of annotated mention types across the dataset.

Although modest in scale, these annotations capture diverse referential behavior, including explicit naming, pronouns, long chains, and variable timing—features crucial for studying multimodal reference resolution. Examples of annotated mentions are shown in Fig. 6.

Paprika mention
(“slice the paprika”)



Pot mention
(“put some water in it”)

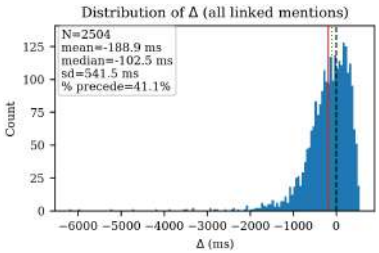
Knife mention
(“grab the knife”)

Figure 6: Examples of annotated mentions: (left) paprika, (middle) pot, (right) knife.

4.6 Gaze–Speech Synchrony Analysis

To quantify the temporal relationship between gaze and speech, we linked 2,504 mentions (92.5% of all 2,707 mentions) to at least one fixation event. We define Δ as fixation offset minus mention onset, so that $\Delta < 0$ indicates the fixation ended before the mention began (gaze fully precedes speech), while $\Delta > 0$ indicates the fixation extended past mention onset (temporal overlap). Figure 7(a) shows the distribution of Δ. The mean lag was -189 ms (median -102 ms, SD $= 541$ ms), with values ranging from -6159 ms to $+531$ ms. Confidence intervals around the mean were $[-211, -168]$ ms. In 41.1% of mentions, gaze fully preceded speech ($\Delta < 0$; 95% CI [39.2%, 43.1%]); in the remaining 58.9%, the fixation was still ongoing when the mention began. The average temporal overlap between gaze and mention was 352 ms (median 367 ms).

The negative mean reflects a heavy left tail: in cases where gaze fully preceded speech, fixations sometimes ended several seconds before the mention. However, in the majority of instances (58.9%), the fixation extended past mention onset, indicating that speakers frequently began naming the referent while still fixating on it. This pattern suggests that gaze functions as a concurrent rather than strictly anticipatory cue in naturalistic referential tasks.



(a)Distribution of gaze–speech lag Δ (fixation offset minus mention onset). Negative values ($\Delta < 0$) indicate that the fixation ended before mention onset; positive values indicate temporal overlap.

Metric	Value	95% CI
Total mentions	2707	–
Linked mentions	2504 (92.5%)	–
Δ mean (ms)	-188.9	$[-210.7, -167.9]$
Δ median (ms)	-102.5	–
Δ SD (ms)	541.5	–
Δ range (ms)	$[-6159, +531]$	–
% gaze precedes	41.1%	[39.2%, 43.1%]
Overlap mean (ms)	351.9	–
Overlap median (ms)	366.6	–

(b) Summary statistics for gaze–speech synchrony.

Figure 7: Gaze–speech synchrony: (a) distribution of Δ = fixation offset $-$ mention onset, and (b) summary statistics.

5 Dataset Card and Ethical Considerations

Dataset Maintenance. The Look and Tell dataset, including all annotations and supplementary materials, will be made publicly available upon publication. It will be hosted on Hugging Face Datasets [Lhoest et al., 2021] to ensure long-term accessibility. We commit to maintaining the dataset, addressing issues raised by the community, and providing a clear versioning system for any future updates, such as the gesture annotations mentioned in our future work. A website with documentation and tutorials will also be provided at: `https://huggingface.co/datasets/annadeichler/KTH-ARIA-referential`.

Ethical Considerations. Participants provided informed written consent for the recording of their video, audio, and gaze, and for the use of this data in anonymized form for research purposes.

Participant Privacy. To protect participant privacy, all faces in the egocentric and exocentric video streams have been blurred using a standard face detection and blurring algorithm. While voices are not altered to preserve the speech data, we have reviewed the transcripts to ensure no personally identifiable information beyond what is inherent in the task was mentioned.

Limitations and Bias. Our dataset, while valuable, has limitations. The data was collected in a single kitchen environment (KTH Kitchen Lab), which may introduce environmental biases. The participant pool of 25 individuals, while diverse in native language, is primarily composed of university students and staff (aged 22-37), and may not be representative of the broader population in terms of age or

background. Furthermore, the tasks are centered around specific recipes, which may favor individuals with some cooking familiarity. These factors should be considered when generalizing findings from this dataset.

6 Discussion

A core challenge in developing embodied AI is achieving robust multimodal spatial grounding, the ability to connect language to objects and locations within a shared 3D environment. Our dataset provides a unique opportunity to study how gaze, gestures, and speech unfold together in situated referential communication to achieve this. By uniquely combining egocentric (first-person) and exocentric (third-person) recordings with metadata supporting both 2D and 3D scene representations, our dataset offers a controlled **benchmark for evaluating spatial intelligence**. It is explicitly designed to facilitate research into **multimodal spatial grounding**, supporting analyses of how gaze can act as a predictive cue for resolving referential expressions and how different spatial representations may influence grounding performance. We hope this resource will foster the cross-disciplinary dialogue necessary to build embodied agents that can truly understand and communicate about space.

Future Work. The recorded sessions contain rich gestural behavior, including pointing and co-speech gestures, which are essential for multimodal reference resolution. Future extensions of this dataset will therefore include systematic gesture annotation, enabling analysis of how gaze, gesture, and speech jointly contribute to referential communication. This will allow embodied agents to be benchmarked not only on gaze-speech synchrony but also on their ability to integrate gestural cues into grounding and dialogue.

Acknowledgments

This work was funded by the Digital Futures project *Adaptive Intelligent Homes*. We thank Kristín Hafsteinsdóttir and Yu Lu for their assistance with data collection and annotation.

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*, 2023. URL `https://arxiv.org/abs/2303.00747`.

Susan E Brennan and Joy E Hanna. Coordinating attention in dialogue. *Discourse Processes*, 43(1): 51–77, 2007.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv e-prints*, pages arXiv–2409, 2024.

Jakob Engel, Thomas Whelan, Richard Newcombe, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. URL `https://arxiv.org/abs/2308.13561`.

Chen Gao, Jiarui Liu, Xin Wang, Peng Zhao, Yujing Shen, Jiajun Zhu, Yanwei Fu, and Qi Liu. Scanqa: 3d question answering with spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Kristen Grauman, Alex Westbury, Eugene Zhang, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Joy E Hanna and Susan E Brennan. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57(4):596–615, 2007.

Muhammet Ilaslan, Chenan Song, Joya Chen, Weixian Gao, Difei /and Lei, Qianli Xu, Joo Lim, and Mike Shou. Gazevqa: A video question answering dataset for multiview eye-gaze task-oriented collaborations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7406–7420, 2023.

Jasleen Kaur, Marc Eaddy, and W Keith Edwards. Synchronization of gaze, speech, and gesture in multimodal human-computer interaction. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, pages 1–8. ACM, 2003.

Youngmin Kim, Jiwan Chung, Jisoo Kim, Sunghyun Lee, Sangkyu Lee, Junhyeok Kim, Cheoljong Yang, and Youngjae Yu. Speaking beyond language: A large-scale multimodal dataset for learning nonverbal cues from video-grounded dialogues. In *ACL (Long Papers)*, 2025.

Dimitris Kontogiorgos, Jörg Bergmann, Giorgio Caldarola, and Stefan Kopp. A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction. In *Proceedings of the LREC Workshop on Multimodal Corpora of Communication (MMC)*, 2018.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, and et al. Datasets: A community library for natural language processing. In *EMNLP: System Demonstrations*, 2021. URL `https://arxiv.org/abs/2109.02846`.

Meta Reality Labs Research. Project aria tools. `https://github.com/facebookresearch/projectaria_tools`, 2023. See citation guidance in docs.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Patrick Renner, Thies Pfeiffer, Nadine Pfeiffer-Leßmann, and Ipke Wachsmuth. Gaze and speech as multimodal input for reference resolution in cooperative tasks. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 84–91. ACM, 2014.

Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019.

Johannes Lutz Schönberger. *Robust Methods for Accurate and Efficient 3D Modeling from Unstructured Imagery*. PhD thesis, ETH Zürich, 2018.

Yusuke Sugano and Andreas Bulling. Seeing with humans: Gaze-assisted neural image captioning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 148–156, 2016.

Vildan Tanriverdi and Robert JK Jacob. Interacting with eye movements in virtual environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 265–272. ACM, 2000.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE TPAMI*, 13(4):376–380, 1991.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Object referring in videos with language and human gaze. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4129–4138, 2018.

Petra Wagner, Lars Schillingmann, and Stefan Kopp. Multimodal cues in instructional interactions: The role of gaze, gesture, and prosody. In *Proceedings of the 25th ACM International Conference on Multimodal Interaction*, pages 123–132. ACM, 2023.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollar, and Kris Kitani. Sam 3d body: Robust full-body human mesh recovery. *arXiv preprint; identifier to be added*, 2025.

Brent Yi, Chung Min Kim, Justin Kerr, Gina Wu, Rebecca Feng, Anthony Zhang, Jonas Kulhanek, Hongsuk Choi, Yi Ma, Matthew Tancik, and Angjoo Kanazawa. Viser: Imperative, web-based 3d visualization in python, 2025. URL `https://arxiv.org/abs/2507.22885`.

A Mention Linking Details

A.1 Prompt Schema

The GPT model is constrained to output a JSON object with the following fields:

- `start`, `len`: token index and length of the mention span
- `ingredient`: string or `null`
- `match_type` $\in$ {`exact`, `synonym`, `hypernym`, `brand`, `coref`, `desc`, `object`, `none`}
- `confidence` $\in [0, 1]$
- `antecedent_start`, `antecedent_len`, `antecedent_text` (optional, for coreference)

A.2 Coreference Rules

Pronouns and demonstratives (*it, this, that, them, these, those*) are linked to the nearest plausible prior mention within approximately 25 tokens, guided by cooking actions. Examples include:

- *open* $\rightarrow$ jar or can
- *drain* $\rightarrow$ pasta or pot
- *put on the stove* $\rightarrow$ pan or pot
- *cut* $\rightarrow$ food item or package

Ambiguous cases default to `ingredient = null, match_type = none`.

A.3 Post-processing Steps

1. **Normalization:** canonicalize predicted strings to recipe or distractor names; non-matches become `None`.
2. **Enrichment:** add common object aliases (e.g., *hob/cooktop* $\rightarrow$ *stove*) and mark `is_object`.
3. **Timing and surface:** attach token indices and timestamps (`start_ns`, `end_ns`, `start_s`, `end_s`) and surface text.
4. **Mention graph:** assign `mention_id` and `chain_id`, infer `mention_type` (including a small utensil lexicon), and link `antecedent_id` for coreference.

A.4 Automated Checks

We flag the following cases:

- unmapped ingredients or objects,
- missing antecedents for coreference mentions,
- non-monotone or overlapping timestamps.

These rules ensure consistent and temporally grounded mentions for each recording `ParXrecY`.