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 MM-Conv—*speak, point, look*—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. The benchmark includes over 4,200 manually verified referring expressions spanning full, partitive, and pronominal types, enabling systematic evaluation of multimodal reference resolution.

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, facial expressions, and 3D scene geometry. We annotated 4,211 naturally occurring referring expressions and designed an evaluation framework spanning explicit nouns to highly contextual 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 over 4,200 annotated expressions from 6.7 hours of con-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9240–9253
11–16 May 2026.

versational data in dynamic 3D environments.

- **A comparative evaluation of state-of-the-art VLMs** against a human baseline, revealing critical failures of current models to leverage conversational context.
- **A modular, context-aware pipeline** that significantly outperforms end-to-end models, demonstrating a more robust architectural approach for interactive AI.
- **A rich, reusable and extendable dataset** with synchronized full-body motion, gaze, facial expression, and egocentric video to support downstream research in gesture generation and 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

Understanding referential expressions is central to multimodal interaction and grounding, but most benchmarks remain static or constrained. Foundational 3D datasets—ScanRefer (Chen et al., 2020) and ReferIt3D (Achlioptas et al., 2020)—spurred progress with single-turn textual descriptions in static scans, while SUNRefer (Liu et al., 2021) (2D RGB-D) and Multi3DRefer (Zhang et al., 2023) extend coverage (e.g., multiple-object grounding). Yet all assume clean, text-only references and single-shot contexts, overlooking spontaneous dialogue where meaning is co-constructed over time. This gap limits the evaluation of models meant for dynamic, interactive settings.

Recognizing these limits, follow-up work moved toward dialogue and embodiment. YouRefIt (Chen et al., 2021) adds multi-turn, multimodal dialogue with gestures but uses third-person videos, thereby missing an egocentric, agent-centric view. ScanERU (Lu et al., 2024) pairs 3D scans with gestures, yet the gestures are synthetic, reducing ecological validity. Interactive benchmarks such as TEACh (Padmakumar et al., 2022) and CEREAL-BAR (Suhr et al., 2019) include dialogue in simulation but are largely text-focused or lack rich, continuous multimodal streams (e.g., synchronized gaze and full-body motion). As a result, current resources do not capture how non-verbal cues resolve ambiguity in real time from an embodied perspective.

Consequently, existing resources fail to model the referential acts of embodied interaction. They struggle to capture the intricacies of conversational grounding, where meaning is shaped by joint attention and perceptual history. As summarized in Table 1, a critical gap remains across existing datasets.

Our work addresses this by integrating temporally continuous modalities, egocentric speech, natural motion capture, and 3D scene representations, from immersive VR scenarios. This new benchmark enables the study of reference resolution in realistic, interaction-rich environments.

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 such as Grounding DINO (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. Unlike CLIP-based approaches that rely on fixed-length embeddings, GroundingGPT conditions its outputs on rich, autoregressively generated context. This enables it to handle long, compositional utterances and resolve ambiguous references by reasoning over dialogue history and visual evidence. Compared to Florence-2, GroundingGPT benefits from deeper language modeling capabilities and more flexible multi-modal generation, making it a promising candidate for grounding tasks in real-world, situated conversational scenarios.

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

Table 1: Comparison of multimodal referential benchmarks and datasets since 2020 highlighting modalities, visual data type, recording settings, embodiment, scale, and dialogue structure.

Dataset	Modal.	Visual Data	Record.	Embod.	Scale	Interaction
ScanRefer (2020) (Chen et al., 2020)	Text, Visual	3D scans	Crowdsourced	None	51.6k utt., 800 scenes	One-shot references
ReferIt3D (2020) (Achlioptas et al., 2020)	Text, Visual	3D scans	Crowdsourced	None	125k utt., 707 scenes	One-turn references
YouRefIt (2020) (Chen et al., 2021)	Text, Visual	RGB video (3rd person)	Real-world dyads	Gesture	4k ref.inst., 432 scenes	Multi-turn, multimodal dialogue
TEACh (2022) (Padmakumar et al., 2022)	Text, Visual, Actions	Simulated 3D (AI2-THOR)	Crowdsourced	Navigation, Manipulation	3k dialogues, 100 scenes	Multi-turn, task-oriented
ScanERU (2023) (Lu et al., 2024)	Text, Visual, Gesture	3D scans	Semi-synthetic	Gesture (synthetic)	46k utt., 706 scenes	Single-turn, gestures

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.

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 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.

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).

Full scenario descriptions are given in Appendix A.1.

Hardware and synchronization. Collection used an OptiTrack motion-capture system with 50-marker skeletons for full-body kinematics (both participants), MANUS Quantum MetaGloves for finger motion on the main actor, and a Meta Quest Pro headset providing binocular gaze and 52 facial blendshape signals. All streams (audio, mocap, gaze/face, simulator) were synchronized via SMPTE timecode using Tentacle Sync E, with the timecode injected into the AI2-THOR simulation. Headset pose was calibrated to the mocap head pose prior to rendering to ensure egocentric alignment between the physical and simulated rigs. The interlocutor was tracked via full-body motion capture only; gaze and facial expression were recorded exclusively for the main actor, who wore the HMD. Further details can be found in Appendix A.3.

Environments and objects. Five apartment rooms from AI2-THOR 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 A.2 for object distributions).

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	Trained	Notable Features
CLIP (Radford et al., 2021)	Transformer (63M)	ViT / ResNet	RGB (2D)	76	No	Dual-encoder with embedding similarity; strong zero-shot alignment but no box output.
Grounding DINO (Liu et al., 2024)	BERT	DETR (Swin)	RGB (2D)	256	Yes	Referring expression grounding via cross-attention; zero-shot localization with box output.
Florence-2 (Xiao et al., 2024)	Transformer (BART)	DaViT	RGB (2D)	1024	Yes	Foundation model with prompt-driven generation; supports grounding using polygon or box output.
GroundingGPT (Li et al., 2024)	Vicuna-v1.5	CLIP ViT-L/14, Q-former (video)	RGB (2D), Video	4096	Yes	Multi-modal model achieving SOTA performance in image, video, and audio grounding.

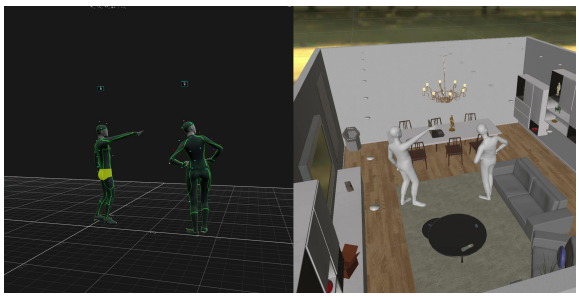


Figure 1: The multimodal data collection environment. A participant in a full-body motion capture suit and VR headset (left) interacts within a virtual scene. Their actions are rendered in real-time in the AI2-THOR simulator (right), enabling the synchronized capture of egocentric vision, speech, motion, and 3D scene geometry.

3.2. Annotation

All modalities were aligned using the shared SMPTE timecode recorded during capture (audio-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 CTC-forced-aligner[1] to obtain word-level timings. Further information on speech transcription can be found in Appendix B.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 AI2-THOR scene—highlighting discourse focus and shifts in attention. Next, referring expressions were classified into three categories using GPT-4o: full noun phrase, partitive/attribute noun phrase, and pronominal (see C.1). 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

[1]For word-level alignment, we used a CTC forced aligner: https://github.com/MahmoudAshraf97/ctc-forced-aligner.

on discourse context and shared perceptual attention.

This categorization follows long-standing distinctions in referring expression research. In computational linguistics, full noun phrases correspond to definite descriptions or explicit referring expressions that uniquely identify entities through head nouns and modifiers (Dale and Reiter, 1995; Kazemzadeh et al., 2014). Partitive/attribute noun phrases capture part–whole and property-based reference—where speakers identify sub-components or perceptual properties of an object—well-documented in multimodal REG (van der Sluis and Krahmer, 2001; Viethen and Dale, 2008). Finally, pronominal references correspond to deictic or anaphoric pronouns (e.g., “it”, “that one”) that depend on discourse context or shared perceptual focus (Hobbs, 1978; Staudte and Crocker, 2011). Together, these three forms span a continuum from lexically explicit to contextually dependent reference, providing a linguistically interpretable basis for evaluating multimodal grounding models.

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

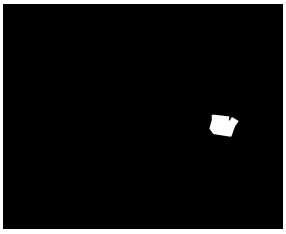
To build a usable multimodal dataset for grounded language modeling, we aligned the referring expressions with the visual stream via word-level timestamps, yielding the exact frame where each reference occurs. For that frame, we extract synchronized egocentric assets: (1) an RGB image, (2) a per-pixel 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 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.

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 participant dyads (groups 3–7, corresponding to five dyads; pid values in Tables 3–6) and recording environments (Table 4), reflecting diverse interaction styles and the trade-offs between scene complexity and referential clarity. In the per-type breakdowns (Tables 5 and 6), "single" and "multiple" indicate whether the expression targets one or more than one object.

(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.

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	1554
3	8	432	220
4	13	974	296
5	12	925	391
6	10	1008	322
7	11	872	325

A key finding is the prevalence of context-dependent language, which is often underrepresented in existing multimodal datasets. As shown in Tables 5 and 6, pronominal references (e.g., “it,” “that one”) are especially common. In the main (egocentric) view, pronouns account for 2,078 expressions (49.4% of 4,211 total; 1,591 single + 487 multiple). In the inter (partner) view, pronouns are

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

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 (Appendix C.1) 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 (Tables 5–6). We note that the interlocutor's high rate of pronominal reference is likely amplified by the instruction not to introduce new referents, which naturally encourages anaphoric rather than full noun phrase form.

Table 5: Distribution of referring expression types by participant (pid) and plurality (main).

	Single			Multiple		
pid	full	part	pron	full	part	pron
Total	1215	435	1591	348	135	487
3	125	41	202	32	11	21
4	335	99	261	117	40	122
5	268	103	373	59	33	89
6	279	85	411	71	15	147
7	208	107	344	69	36	108

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

Table 6: Distribution of referring expression types by participant (pid) and plurality (inter).

	Single			Multiple		
pid	full	part	pron	full	part	pron
Total	248	48	901	137	44	169
3	22	9	160	13	4	12
4	54	14	146	21	16	42
5	70	10	217	47	16	30
6	51	3	195	21	3	48
7	51	12	183	35	5	37

that crowd-sourcing 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 (Florence-2, GroundingGPT, GPT-4o) 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. We focus on the main actor's egocentric perspective because referring expressions originate from this speaker and the egocentric view provides the visual input required for VLM evaluation. The interlocutor's perspective is available in the released dataset for future work on addressee-side reference resolution.

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.

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

Table 7: VLM Grounding Performance Comparison with Multiple IoU Thresholds. We report Match Rate at IoU thresholds of 0.5 and 0.3 (M@.5 / M@.3), along with mean IoU (mIoU@.5). The best performance in each row is in **bold**.

		Florence-2		GroundingGPT		Ours (GPT-4o + F-2)	
Expression Type	**Context**	**M@.5 / M@.3**	**mIoU@.5**	**M@.5 / M@.3**	**mIoU@.5**	**M@.5 / M@.3**	**mIoU@.5**
Full NP	w/o ctx	**61.4% / *70.1%***	0.860	31.4% / *39.8%*	0.722	51.5% / *60.2%*	0.886
	w/ ctx	**59.6% / *62.9%***	0.863	19.1% / *31.4%*	0.675	49.8% / *58.5%*	0.897
Partitive NP	w/o ctx	24.3% / *31.5%*	0.826	9.2% / *15.4%*	0.710	**40.6% / *51.2%***	0.906
	w/ ctx	38.0% / *41.0%*	0.856	9.5% / *13.2%*	0.692	**42.5% / *53.8%***	0.899
Pronominal	w/o ctx	22.0% / *28.4%*	0.851	12.6% / *19.1%*	0.782	**47.1% / *59.3%***	0.895
	w/ ctx	40.9% / *42.4%*	0.852	8.4% / *12.3%*	0.722	**48.4% / *61.0%***	0.894

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. In addition to that, each evaluation run had three semi-randomized attention checks, which were between 3.3-4.6% of the total set of stimuli. The attention checks were prompting a participant to click on one of the four corners of the image; the corner selection in the attention check prompt was also randomized.

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.

The crowd-sourced evaluation (Table 8) 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, however, benefits underspecified forms. Partitive NPs improved from 47.93% to 60.99% accuracy and from 47.01% to 61.96% agreement, while pronominals—lowest overall—rose from 37.43% to 55.42% accuracy and from 33.33% to 55.16% agreement.

Table 8: 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

4.3. VLM evaluation

To better 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 two monolithic, end-to-end models, Florence-2 (Xiao et al., 2024), and GroundingGPT (Li et al., 2024), against a novel modular pipeline designed to explicitly handle conversational context. These models were chosen for their ability to generate bounding box outputs, which aligns with our evaluation methodology.

We evaluated the monolithic grounding models by providing the full referring utterance as a prompt

and tasking the model with returning a bounding box for the referent. The results, shown in Table 7, reveal a clear weakness. While Florence-2 achieves a respectable match rate on explicit Full Noun Phrases (61.4% M@.5), its performance drops sharply for Partitive NPs (24.3%) and especially for Pronominal references (22.0%), which are highly dependent on discourse history. This high failure rate for ambiguous expressions underscores a critical limitation of current end-to-end models: they struggle to leverage conversational context to resolve ambiguity.

To address this limitation, we propose a modular, two-stage pipeline that decouples high-level conversational reasoning from low-level visual localization. This approach is designed to first resolve linguistic ambiguity before attempting to ground the reference in the visual scene.

1. **Contextual Disambiguation (GPT-4o):** Given an utterance and its dialogue history, GPT-4o rewrites ambiguous references (e.g., that one") into explicit forms (e.g., the painting of the ship above the fireplace"), using a custom prompt (see supplementary material).
2. **Visual Grounding (Florence-2):** The disambiguated expression is then passed to Florence-2, which performs referring expression segmentation on the corresponding egocentric RGB image without fine-tuning.

As shown in Table 7, our pipeline (GPT-4o + F-2) consistently outperforms monolithic baselines. The largest gain occurs for pronominal references, where the match rate more than doubles (22.0% $\rightarrow$ 47.1%), highlighting the value of resolving discourse history before grounding. Partitive NPs also improve markedly, showing better spatial and part–whole reasoning. Including dialogue history benefits ambiguous types (partitive, pronominal), while for explicit full NPs, it slightly reduces match rate but increases mean IoU, indicating more precise grounding despite added complexity. Overall, the results confirm that decoupling linguistic reasoning from visual perception is highly effective for grounding in conversational contexts.

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.

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.

Haolin Liu, Anran Lin, Xiaoguang Han, Lei Yang, Yizhou Yu, and Shuguang Cui. 2021. Refer-it-in-RGBD: A bottom-up approach for 3D visual grounding in RGBD images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6032–6041.

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.

Ziyang Lu, Yunqiang Pei, Guoqing Wang, Peiwei Li, Yang Yang, Yinjie Lei, and Heng Tao Shen. 2024. ScanERU: Interactive 3D visual grounding based on embodied reference understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3936–3944.

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.

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.

Alane Suhr, Claudia Yan, Charlotte Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019. Executing instructions in situated collaborative interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2119–2130, Hong Kong, China. Association for Computational Linguistics.

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.

Yiming Zhang, ZeMing Gong, and Angel X. Chang. 2023. Multi3DRefer: Grounding text description to multiple 3D objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15225–15236.

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

M. Bain et al. 2023. Whisperx. https://github.com/m-bain/whisperX.

E. Kolve et al. 2017. Ai2-thor simulator. 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 **Speak, Point, Look**, 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	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.

Modality Summary.

Format. Each RE is aligned to its egocentric frame; we provide JSON annotations (utterance, token indices, category, referent ID), RGB, mask, depth (normalized in 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.

[2] https://github.com/allenai/ai2thor

– *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.

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.

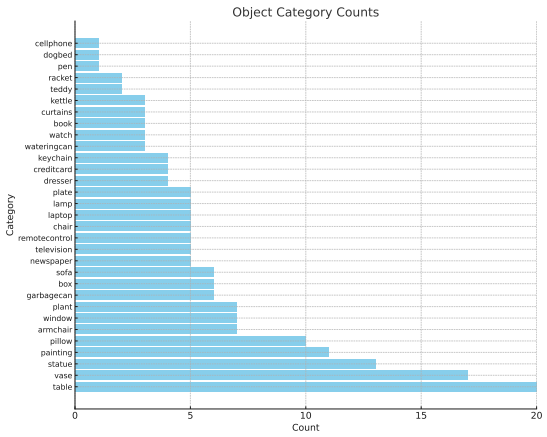


Figure 4: Object category distribution across simulated environments used in the experiments.

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

[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

WhisperX's ASR output, providing a fast starting point for correction. Annotators focused on preserving spontaneous speech characteristics (e.g., 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".

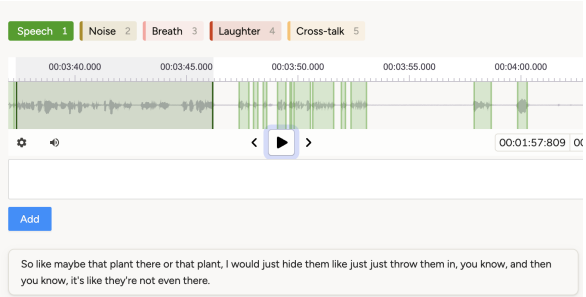


Figure 6: Label Studio interface for speech annotation. Annotators reviewed WhisperX transcriptions aligned to VAD-based audio segments and edited text within the linked text box.

C. GPT-4o prompts

C.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'.
```

C.2. Contextual Disambiguation Prompt

We use GPT-4o for linguistic disambiguation, followed by the Florence-2 model. In case of the without context case, the conversation history is omitted from the prompt.

Listing 2: Prompt used for contextual disambiguation of referring expressions using GPT-4o. The system provides conversational history and object context, and the model generates a resolved, explicit phrase.

```
SYSTEM: You are a helpful
    assistant that creates brief,
    clear descriptions of objects
    and their locations.

USER:
Create a brief, single sentence
    that describes the referent {
    phrase} and its location. Be
    concise and direct.

Examples:
- The black chair sits in the
    corner by the table.
- A blue coffee table stands in
    front of the couch.
- The tennis racquet rests by the
     TV.
- The lamp hangs above the dining
     table.

Previous utterances:
{conversation_history}

Object list in scene:
{object_list}

Current utterance: "{utterance}"
Referenced phrase: "{phrase}"
Object type: {topic}

Return only a short, single
    sentence. Focus on essential
    details only.
```

D. Inter-Annotator Agreement

We assessed annotation reliability through a crowdsourced study with 78 workers (recruited via Prolific, filtered by attention checks), providing 3 independent annotations per stimulus. Participants viewed an egocentric image with the target utterance and clicked on the referred object.

Table 10 reports object-level agreement using 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. Krippendorff's α ranged from 0.42 (pronominal, no context) to 0.66 (full NP, no context), indicating moderate agreement by conventional standards.

Importantly, the gap between inter-annotator agreement and ground-truth accuracy (Table 8) reflects task difficulty rather than annotation noise. Analysis of unanimous misses—where all three annotators clicked outside the ground-truth mask (Table 11)—reveals three patterns:

- **Near-misses** (15–20%): Tight clusters ($\leq$50px spread) with centroids close to the mask ($\leq$30px), reflecting conservative mask boundaries rather than identification errors.

Table 10: Object-level inter-annotator agreement across expression types and context conditions. We report unanimous agreement (all 3 annotators clicked the same object within 30px), Krippendorff's α (nominal), and the conditional click distance when all annotators hit the ground-truth object.

Type	Context	Unanimous	α	Cond. Dist.
Full NP	w/o ctx	76.7%	0.66	23.6px
	w/ ctx	72.1%	0.58	32.8px
Partitive	w/o ctx	79.2%	0.55	44.2px
	w/ ctx	82.7%	0.64	30.8px
Pronominal	w/o ctx	75.0%	0.42	23.7px
	w/ ctx	80.2%	0.57	27.2px

Table 11: Analysis of unanimous misses, where all three annotators clicked outside the ground-truth mask. Near-misses indicate tight annotation masks rather than annotator error; coherent wrong clicks show strong agreement on an alternative referent; scattered responses indicate genuine ambiguity.

Category	Context	Full NP	Partitive	Pronominal
Near-miss	w/o ctx	18–24%	11–16%	9–13%
	w/ ctx		*(similar pattern)*	
Coherent wrong	w/o ctx	33–39%	28–38%	21–30%
	w/ ctx		*(similar pattern)*	
Scattered	w/o ctx	24–32%	27–40%	38–51%
	w/ ctx		*(reduced scatter)*	

Near-miss: cluster $\leq$50px, centroid $\leq$30px from mask. *Coherent wrong*: cluster $\leq$50px, centroid $>$50px from mask. *Scattered*: spread $>$100px.

- **Coherent disagreements** ($\sim$30%): Tight clusters ($\leq$50px) centered on alternative objects, representing genuine referential ambiguity (e.g., "the lamp" with multiple lamps visible).
- **Scattered responses** (24–51%): Click spread exceeding 100px with no consensus, most prevalent for pronominal expressions without context (51%), confirming that bare pronouns are unresolvable without discourse history.

The median pairwise distance among unanimous misses (11–64px) remains far below chance ($\sim$270px), indicating consistent interpretations even when diverging from ground truth. Adjusting for near-misses and coherent alternatives, effective agreement exceeds 80% for most conditions. This validates our annotation quality while underscoring the benchmark's difficulty: references unambiguous in situated interaction become genuinely ambiguous when reduced to text and a single image frame.