LVLMs are Bad at Overhearing Human Referential Communication

Zhengxiang Wang[1,2] **Weiling Li**[3]
Panagiotis Kaliosis[4] **Owen Rambow**[1,2] **Susan E. Brennan**[3]
[1]Department of Linguistics [2]Institute for Advanced Computational Science
[3]Department of Psychology [4]Department of Computer Science, Stony Brook University
zhengxiang.wang@stonybrook.edu

Abstract

During conversation, speakers collaborate on spontaneous referring expressions, which they can then re-use in subsequent conversation with the same partner. Understanding such referring expressions is an important ability for an embodied agent so that it can carry out tasks in the real world. This requires integrating and understanding language, vision, and conversational interaction. We study the capabilities of seven state-of-the-art Large Vision Language Models (LVLMs) as overhearers to a corpus of spontaneous conversations between pairs of human discourse participants engaged in a collaborative object-matching task. We find that such a task remains challenging for current LVLMs, which fail to show a consistent performance improvement as they overhear more conversations from the same discourse participants repeating the same task for multiple rounds. We release our corpus and code[1] for reproducibility and to facilitate future research.

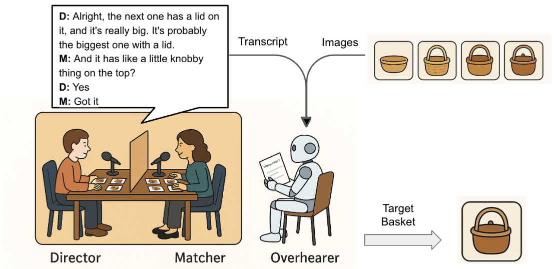

D: Alright, the next one has a lid on it, and it's really big. It's probably the biggest one with a lid.
M: And it has like a little knobby thing on the top?
D: Yes
M: Got it

Figure 1: Our overhearer matching task (after Schober and Clark, 1989): the AI agent (LVLM) reads a transcript from a human referential communication corpus and tries to match the same cards as the matcher to the director's target sequence.

1 Introduction

A crucial skill for embodied AI agents working with humans is *grounding in referential communication*: the ability to resolve which object in the visual environment a speaker is referring to. This is a non-trivial problem for several reasons: there may be no lexicalized label associated with the referent; there may be many ways to refer to it; or there may be multiple objects of the same type in the environment. Moreover, referential communication occurs in different interactive contexts: the referring expression can be part of a single, one-off instruction given to an AI agent; it can unfold over several conversational turns as a human interacts with the AI agent to clarify meaning; or the AI agent may overhear a conversation between two or more humans. The past few years have witnessed rapid advances in large vision language models (LVLMs) (Alayrac et al., 2022; Liu et al., 2023; Dai et al., 2023; OpenAI et al., 2024b; Team et al., 2024, *inter alia*). However, LVLMs still lag behind human capabilities in both comprehending and generating referring expressions (Tang et al., 2024), even in a single-instruction setting.

In this paper, we address the problem of an AI agent overhearing two people engaged in spontaneous referential communication. This scenario is important because the AI agent may be a side participant who stands by to assist when called on; it will need to understand the referential conventions that discourse participants develop over time. For example, assistive robots in the home may need to monitor conversations between residents (with prior consent) and perform tasks (such as manipulating objects) that require integrating and understanding language, vision, and conversational interaction.

To address this issue, we use a previously unpublished corpus of spontaneous conversations between pairs of humans engaged in a collaborative object-matching task over repeated rounds of the same task. As illustrated in Figure 1, we prompt an LVLM, acting as an overhearer, to perform the

[1] https://github.com/jaaack-wang/lvlms-overhearing

same task as the human matcher—that is, matching objects to the director's target sequence.

We ask two research questions: (**1**) How well can LVLMs perform as overhearers in a referential communication task? and (**2**) Can LVLMs improve on their ability to resolve human-generated referring expressions, after witnessing repeated references to the same objects by the same human discourse participants? The second question is particularly important because even though psycholinguistics studies have shown that interacting in a conversation differs from overhearing it (Schober and Clark, 1989; Fox Tree, 1999; Fox Tree and Mayer, 2008; Castano et al., 2023), these two roles have not been defined or distinguished for LLMs. A useful AI agent in the overhearer role should adapt to and learn from the dynamics of language use, much like a human overhearer can. A failure to do so would suggest that the agent cannot effectively accumulate personalized knowledge across interactive contexts, thereby limiting its practical utility. To our knowledge, this study is the first to test LVLMs on a referent-matching task using a corpus of human referential communication.

Our primary contributions are as follows:

1. We release a corpus of spontaneous referential communication dialogues, collected under controlled conditions but previously unpublished, to facilitate future studies.

2. We demonstrate empirically that resolving references to common real-world objects (i.e., baskets and dogs) produced during spontaneous conversation remains challenging for LVLMs, with high unexplained variability.

3. We show that all tested LVLMs, including proprietary LVLMs like GPT-4o, fail to show a consistent performance improvement as they read more conversations from the same human pair matching the same objects over time.

2 Theoretical Background

Conversation, by its very nature, is collaborative. Speakers and addressees tailor their utterances and their interpretations to each other's knowledge, needs, and perspectives, as well as to the *common ground* they share (Clark and Wilkes-Gibbs, 1986). Sources of common ground can include co-membership in a community, as well as perceptual co-presence, which derives from interlocutors' mutual awareness of the shared environment. Most important for our purposes is linguistically-established common ground, or the prior co-presence of interlocutors to what they've said previously and can presume is part of their mutual knowledge (Clark and Marshall, 1981). The process of establishing and updating common ground is an essential engine for collaboration.

Grounding The term *grounding* has been used extensively in cognitive science, psycholinguistics, and AI. Grounding can be described as the "access to or awareness of the physical, perceptual, goal-oriented or social contexts in which language occurs" (Pavlick, 2023), and in human communication more specifically, the interactive process by which interlocutors seek and provide evidence that they understand one another, as they accrue common ground (Clark and Brennan, 1991; Metzing and Brennan, 2003; Brennan, 2005).

This paper focuses on the grounding of linguistic expressions by pairs of people, mapped onto visual representations during a collaborative referential communication task. Although word meanings are informed by linguistic conventions, meanings are not "contained" within words (Reddy, 1993), but can be collaboratively constructed by speakers and addresses, often in service of a shared task or goal (for discussion, see Brennan and Clark, 1996).

Addressees vs Overhearers Conversation is often studied in the lab using variants of a matching task, in which addressees interact with speakers to match a set of picture cards, build something together, or trace a route on a map. The natural behavior of speakers and addressees in such tasks is to collaborate until they have evidence that they've reached a "grounding criterion" sufficient for their current purposes (Clark and Brennan, 1991; Clark and Wilkes-Gibbs, 1986); when the shared goal is to match a set of cards accurately, they continue seeking and providing evidence until they believe they've reached that criterion. The first round of a matching task always takes the longest (more time, words, and turns), becoming more efficient in subsequent rounds with the same objects and partners as they accrue common ground.

The role of an *overhearer* is very different from that of an addressee. A now-classic psycholinguistics study by Schober and Clark (1989) demonstrated that addressees in a matching task perform more accurately than overhearers in the same task, because they are able to ground meaning with speakers, whereas the overhearers cannot. An overhearer sometimes understands the referent *early* (but must wait for the task to move on), sometimes *late* (and falls behind in the task) and sometimes not at all (and selects the wrong picture card). In Schober and Clark (1989), addressees (who could contribute to the conversation) matched the cards nearly perfectly, whereas overhearers (who heard every word but did not interact) reached only about 80% accuracy in Round 1 and about 90% by Round 4. Strikingly, "late overhearers" who listened in to the recorded conversations starting in Round 3 did worst of all, achieving only 68-73% accuracy, even when they could stop and start the recordings to try to keep up (See Figure 3 in their paper).

Variation in Human Language Use There is considerable variation in human language use, including in choices of wording, syntax, prosody, and coordination strategies. This variation is not random, but emerges strategically, such that lexical variability is much greater across conversations than within a conversation. For instance, two partners in dialogue tend to *entrain* on words, consistently using the same referring expression (albeit in shortened form) over the course of repeated referring to the same object, as if to confirm a "conceptual pact" that they're referring to the same thing they discussed previously (Brennan and Clark, 1996; Krauss and Weinheimer, 1964; Metzing and Brennan, 2003).

Challenges for an AI Agent Ultimately, a useful embodied AI agent should be able to maintain, adapt, and build on the common ground accrued in conversation. The agent should (1) be robust enough to understand and track the expressions that different pairs of speakers use to describe the same objects (Brennan and Clark, 1996), and (2) be able to cope with the dynamic variations in human language use, including the choice of wording, syntax, prosody, and coordination strategies (especially those due to the use of common ground). This requires evaluating how well a foundation model (whether LLM or LVLM) performs with spontaneous dialog during repeated discussions of the same objects with the same speakers.

3 Related Work

Machine Comprehension of Referring Expressions Evaluating the visual grounding abilities of language models often involves tasks that require identifying a particular object in an image using a natural language referring expression (Qiao et al., 2021). Conventional methods typically follow a two-stage approach: first generating open-vocabulary object proposals, then selecting the one that best matches the language description. More recent efforts have built upon the emerging capabilities of vision transformers (Dosovitskiy et al., 2020), leading to improved models (Su et al., 2024a). Moreover, LVLMs have demonstrated strong performance on visual grounding tasks even in zero-shot settings (Sui et al., 2023; Subramanian et al., 2022), while specialized approaches have been developed to further improve their zero-shot grounding abilities without task-specific supervision (Han et al., 2024).

Visual Grounding Capabilities of LVLMs LVLMs exhibit strong visual grounding capabilities thanks to their large-scale multi-modal pretraining (Lu et al., 2023). On top of the core vision-language alignment principles established by foundational models, such as CLIP (Radford et al., 2021), LVLMs show remarkable performance on a wide range of tasks, including referring expression comprehension (Su et al., 2024b), visual question answering (Sinha et al., 2025), and instruction following (Bitton et al., 2023). However, spatial and compositional reasoning remains a challenging task for current LVLMs, as, for example, they often struggle with relational cues (Chen et al., 2024) or multiple visual concepts (Zeng et al., 2024).

Another line of research explores the interactive visual grounding capabilities of LVLMs engaged in multi-turn dialogues to resolve ambiguous references and refine their understanding through conversational context (Feng et al., 2023; Tian et al., 2025). Recently, Tang et al. (2024) showed that humans consistently outperform LVLMs in comprehending referring expressions generated to target a dot placed in a shared 3-D environment. Our study extends this work to focus on the capabilities of LVLMs as overhearers, to examine the mapping of referring expressions onto referents discussed repeatedly in spontaneous human conversations, rather than as one-off descriptions.

Finally, although LVLMs may be able to process language produced by humans, they do not exhibit a human-like tendency to spontaneously make the language they generate more efficient over multiple turns (Hua and Artzi, 2024).

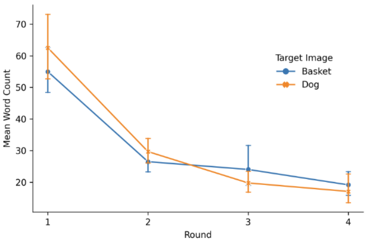
Example Target Object	Example Dialogue (R1-R4)	Mean Word Count Per Object Over Rounds
	Round 1: ***D:*** *The first basket has a handle. And, it's kinda like a rectangular shape. It's kinda long.* ***M:*** *It has two handles?* ***D:*** *No, one handle.* ***M:*** *And it's kinda long?* ***D:*** *Yes. It's the only--it's the longest basket with a handle* ***M:*** *Ok*	
	……	
	Round 4: ***D:*** *And the tenth one is the--the rectangular shaped basket* ***M:*** *Yeah*	

Figure 2: The left panel shows one target basket; the middle panel shows one pair's corresponding dialogue from Round 1 to Round 4, demonstrating entrainment on more concise language (for the perspective "rectangular-shaped"). Here, entrainment occurs after they consider multiple proposals in Round 1. The right panel depicts the mean word count (a measure of efficiency) for baskets and dogs across rounds. Error bars indicate ±1 standard error of the mean across pairs.

Studies of Machines as Overhearers Numerous studies have evaluated LLM/LVLM foundation models in the overhearer/observer role (Castano et al., 2023; Kim et al., 2023; Kosinski, 2024; Jin et al., 2024; Soubki et al., 2024; Street et al., 2024). Most pertain to Theory of Mind (ToM) (Premack and Woodruff, 1978) and examine a model's ToM ability to attribute false beliefs to characters due to their absence at a critical point in the story, following the classic Sally-Anne test (Wimmer, 1983; Baron-Cohen et al., 1985). Our study differs from ToM studies in that we analyze the grounding capabilities of LVLMs without assuming information asymmetry between the overhearer and the discourse participants: all have access to the same words, as in Schober and Clark (1989).

4 Corpus

Overview Our corpus comprises 80 human-to-human dialogues totalling 27,902 words, collected by Calion B. Lockridge and Susan E. Brennan at Stony Brook University in 2001 and not previously published. Ten pairs of native-English-speaking undergraduates (20 speakers in total) did repeated rounds of a referential communication task (Krauss and Glucksberg, 1969; Clark and Wilkes-Gibbs, 1986). During each round, the pairs spoke freely while they matched duplicate sets of picture cards. The dialogues were recorded and manually transcribed. Figure 2 shows a representative example excerpted from the transcripts of two people describing the same target basket in Round 1 and again in Round 4.

Task and Materials Following Clark and Wilkes-Gibbs (1986), speakers were recruited in pairs, with one partner randomly assigned to the role of director (D) and the other to the role of matcher (M); they remained in their assigned role throughout the experiment. Partners sat in separate rooms and communicated via an audio channel.

Each pair completed a total of eight rounds of the referential communication task in a one-hour session—four rounds with the same set of pictures of dogs, and four rounds with the same set of pictures of baskets (counterbalanced for order). These basic-level categories were chosen to vary the difficulty of expressing and identifying referents : dogs are associated with commonly-known subordinate category labels such as breeds, whereas baskets are not (see Figures 5 and 6 in Appendix A for details). D's and M's sets contained duplicates of the same 10 dogs (or baskets), with 3 additional cards only in M's set, to require them to discuss all 10 targets.

Performance and Linguistic Patterns All pairs successfully completed the matching task in all rounds, achieving 100% accuracy. This corpus showed consistent linguistic patterns that align with the findings of Clark and Wilkes-Gibbs (1986) and Schober and Clark (1989), with partners becoming more efficient in their expressions across rounds. That is, objects were described in greatest detail in Round 1, often with multiple proposals for expressions until M acknowledged understanding. By as early as Round 2, word counts dropped sharply, by about 50%. By Rounds 3 and 4, partners typically had entrained on shared conceptualizations, with concise labels for the objects This reflects the accumulation of common ground over repeated interactions. The summary plot in Figure 2 illustrates this in the form of reduced word counts across rounds (see also Figure 7 in Appendix A).

Manual Extraction of Object Descriptions Our experiments use the transcripts from these spontaneous conversations; however, we extended the corpus for a follow-up experiment (see the Object Descriptions test reported in Section 7) by manually extracting 10 complete object descriptions from each transcript, yielding a 800 object descriptions. Each description starts with D describing a target object and ends when M recognizes it. This allows for a finer examination of a system's visual grounding capability on an object level.

Corpus Value and Lack of Data Contamination As this corpus has not been published and is not included in any LVLM training data, it is free from the risk of data contamination (Jacovi et al., 2023; Sainz et al., 2023). It provides an ideal testbed for evaluating LVLMs' ability to adapt to spontaneously produced referring expressions from multiple speakers, grounded in visual images, without the influence of prior exposure or memorization.

5 Methodology

Task Description To prompt an LVLM to perform the overhearer matching task, we provide it with a transcript and an image as inputs. The transcript is a conversation between a director and a matcher at a specific round from our corpus. The input image contains the corresponding 13 objects (baskets or dogs) used during the conversation, randomly arranged in a 3x5 grid, and numbered from 1 to 13 (see Figures 5 and 6 in Appendix A for two examples). The LVLM is instructed to produce the correct sequence of 10 target object indices for the objects as described by the director.

Experimental Procedure Our corpus contains four rounds of conversations between each human director-matcher pair for each object type. To address the two research questions listed in Section 1, we evaluate (1) the performance of an LVLM as overhearer at single rounds and (2) how this performance evolves over multiple rounds of conversation between each director-matcher pair, with the methods and performance of Schober and Clark (1989)'s overhearers in mind. More concretely, we measure LVLM task performance in each round from a starting round to the end round. We tested four starting rounds for each object type, i.e., Round 1 (R1), Round 2 (R2), Round 3 (R3), and Round 4 (R4), and one end round, R4. We prompt the LVLM to perform the matching task for each starting round separately in a multi-turn conversation setting.

The input objects are shuffled for each round, since the human overhearer's display showed the objects in arbitrary order. Our early experiments showed that LVLMs appear to be sensitive to object orderings in their visual input, which we further evaluate in Section 6.3. While this sensitivity to ordering reflects a challenging aspect of our corpus, offering an ideal testbed for robustness testing, it poses an evaluation challenge. To minimize effects of order and obtain more reliable results, we run each LVLM five times with different object orderings via greedy decoding for each experimental configuration throughout the study.

LVLMs We evaluate (**1**) four proprietary LVLMs, namely Claude-3.7-Sonnet (Anthropic, 2025), Gemini-2.0-Flash (Google DeepMind, 2024), and GPT-4o and GPT-4o-mini (OpenAI et al., 2024a), as well as (**2**) three open-weight LVLMs, namely Qwen2.5-VL-32B and Qwen2.5-VL-7B (Bai et al., 2025), and Pixtral-12B (Agrawal et al., 2024). We choose only models that support multiple input images, since our task involves multiple input images from different rounds. See Appendix B.1 for more details about these models.

Prompting In the prompt, we provide all the necessary background information regarding the human director-matcher matching task, as described in Section 4. We explain the overhearer matching task and the task procedure, as illustrated earlier. We prompt all LVLMs with zero-shot chain-of-thought (Kojima et al., 2024), with temperature set to 0 to maximize reproducibility. All the prompt templates can be found in Appendix D.

Evaluation Metric We compute accuracy, i.e., percentage of correctly matched objects, to measure LVLM task performance on the overhearer matching task. In other words, a model getting 9/10 or 0/10 objects correct would score 90% and 0%, respectively.

6 Results

Figure 3 shows the average accuracy (with 95% confidence intervals) of various LVLMs on the

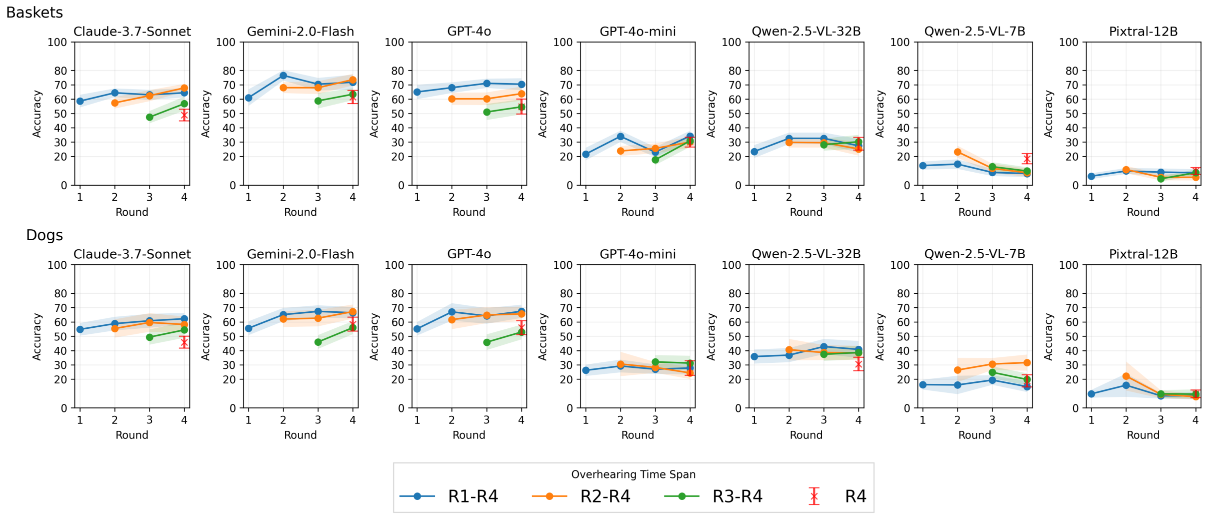


Figure 3: Average accuracy of various LVLMs in the overhearer task over rounds. There are 4 overhearing starting points from Round 1 to Round 4, yielding three lines and one single point. The shaded areas and error bars denote 95% confidence intervals. In this corpus, all human matchers' performance is 100% at every round.

overhearer matching task across multiple rounds. We compare performance among the models for a single round and for sequences of rounds with different starting points (capturing the experience of Schober and Clark (1989)'s "late overhearers").

To interpret these results, we first examine the performance of the LVLMs at single rounds to evaluate how well they resolve real-world object references in spontaneous, interactive conversations. We then analyze how their performance evolves when initialized from different starting rounds, providing insights into their potential as embodied agents. Lastly, we conduct a robustness analysis.

6.1 Performance at Single Rounds

Proprietary LVLMs substantially outperform open-weight models. As Figure 3 shows, the large proprietary LVLMs (i.e. GPT-4o, Gemini-2.0-Flash, and Claude-3.7-Sonnet) achieve an average accuracy ranging from 45.8% (Claude-3.7-Sonnet at R4 with a R4 start for dogs) to 76.5% (Gemini-2.0-Flash at R2 with a R1 start for baskets). In contrast, the open-weight models achieve only 42.8% accuracy at best (Qwen-2.5-VL-32B at R3 with a R1 start for dogs), and 4.4% at worst (Pixtral-12B at R3 with a R3 start for baskets).

Model scaling appears beneficial for language grounding. Model sizes of proprietary LVLMs are not publicly disclosed. But available information suggests that size and performance are correlated. Specifically, GPT-4o consistently outperforms GPT-4o-mini, and Qwen-2.5-VL-32B outperforms Qwen-2.5-VL-7B. Furthermore, proprietary LVLMs are likely larger than other models, suggesting that larger models perform better.

LVLMs underperform human matchers. Recall that in our corpus, all human director-matcher pairs completed the matching task with 100% accuracy in every round, substantially outperforming all tested LVLMs, regardless of LVLM starting round. Prior research has shown that human overhearers also perform worse than interacting partners in matching tasks (Schober and Clark, 1989; Fox Tree, 1999; Fox Tree and Mayer, 2008; Wilkes-Gibbs and Clark, 1992). But unlike human overhearers, LVLM overhearers in our experiments can access the entire conversation for every matching decision, thanks to their built-in attention mechanisms. Despite this advantage, even state-of-the-art LVLMs fail to reliably exploit this additional conversational context to match human-level performance. This suggests that *grounding spontaneous, naturalistic descriptions to visual referents remains a substantial challenge for LVLMs.*

6.2 Performance Dynamics Across Rounds

LVLMs fail to show consistent improvement on the same matching task over time. To measure how each model's performance evolves over time, we use ordinary least squares (OLS) regression to model the overall performance trend shown in Figure 3. A run shows improvement only if the regression line has a positive coefficient with a p-value less than 0.05. The results in Table 1 confirm

	Starting Round	R1	R2	R3
Source	Model			
Baskets	Claude-3.7-Sonnet	1.6	5.2***	9.4**
	Gemini-2.0-Flash	2.6*	2.8	4.6
	GPT-4o	1.9*	1.8	3.6
	GPT-4o-mini	2.7**	3.1*	13.2***
	Qwen-2.5-VL-32B	1.2	-2.3	2.0
	Qwen-2.5-VL-7B	-2.3***	-7.3***	-3.0
	Pixtral-12B	0.6	-2.7**	4.2**
Dogs	Claude-3.7-Sonnet	2.4*	1.4	5.0
	Gemini-2.0-Flash	3.5***	2.6	10.0**
	GPT-4o	3.4**	2.0	7.1
	GPT-4o-mini	0.3	-3.1	-0.8
	Qwen-2.5-VL-32B	2.1	-1.1	1.2
	Qwen-2.5-VL-7B	-0.1	2.6	-5.0
	Pixtral-12B	-1.2	-7.2**	-0.2

Table 1: Overall performance trend (slope) over rounds for each LVLM starting at R$_i$, using ordinary least squares (OLS) regression. Significant positive and negative trends are highlighted, along with significance ("*" is $p < 0.05$, "**" is $p < 0.01$, and "***" is $p < 0.001$.) See Figure 3 for a visualization of the performance trend.

the performance gap between proprietary LVLMs and open-weight LVLMs, with the former showing more desirable performance trends. However, even for the proprietary LVLMs, overall performance does not consistently improve for every starting round (the only exception is, surprisingly, the smallest proprietary model, GPT-4o-mini, for baskets). Often, the open-weight models even decrease. We also analyze the overall performance trend using Spearman and Kendall rank correlations, which further validate these findings and show that even for a positive performance trend, the overall correlations remain small (see Appendix C.1).

We further measure the performance trend for the dialogues of each human pair, using the same method. We then compute the percentage of human pairs for whom the LVLMs show consistent improvement over the starting rounds. The results show that the models tested fail to show an overall improvement on at least 70% of the human pairs (see also Table 7 in Appendix C.1).

Lastly, as a sanity check, we simply compute the percentage of times an LVLM shows monotonically increasing performance for each starting round across the two datasets. The results show that all LVLMs struggle to achieve a smooth, incremental performance improvement over time, since, for example, when starting from R1, the best model exhibits a monotonically increasing performance curve only 46% of the time. See Table 6 in Appendix C.1 for details.

LVLMs do not consistently benefit from an early start, unlike humans. Given the characteristics of our corpus where object descriptions used in R1 tend to be much more elaborate than the subsequent rounds (see Figure 2), we expected LVLMs to perform better in their starting round if they begin with R1 than with R2 (or other later rounds). However, no models show a significantly better performance in R1 than R2 as their starting rounds in paired t-tests (see Table 9 in Appendix C). In fact, several LVLMs perform significantly better in their starting round when beginning with R2 rather than R1, including Gemini-2.0-Flash for baskets (7% mean accuracy gain for R2, $p < 0.05$.)

Source Model	Baskets	Dogs
Claude-3.7-Sonnet	-3.9***	-3.3**
Gemini-2.0-Flash	-0.8	-0.8
GPT-4o	-4.0***	-1.3
GPT-4o-mini	1.9*	0.6
Qwen-2.5-VL-32B	1.5	-1.9
Qwen-2.5-VL-7B	0.4	0.6
Pixtral-12B	0.4	-1.2

Table 2: Overall performance trend (slope) across the four starting points (R1, R2, R3, and R4) for each LVLM, using ordinary least squares (OLS) regression. A significant negative slope is more desired here as that indicates that an model benefits from an earlier start.

To study the overall performance differences between starting at an earlier round versus a later round, we use OLS regression to analyze the overall performance trend across the four starting points (i.e., R1, R2, R3, and R4). Table 2 shows that two proprietary LVLMs (GPT-4o and Claude-3.7-Sonnet) generally perform better the earlier overhearing starts, but none of the other LVLMs show such a performance pattern.

In principle, with an earlier start, an LVLM reads in more conversation between the human director and matcher, so it should better understand the subsequent rounds of conversations, compared to when it begins at a later round. We compare the mean accuracy differences of overlapping rounds between an early start and a late start using paired t-tests (see Table 8 in Appendix C). Proprietary LVLMs tend to benefit more from an early start, in terms of performance gain over subsequent rounds, than open-weight models. This suggests that these open-weight models are less able to use previous discourse effectively to capture the dynamics of language use over time.

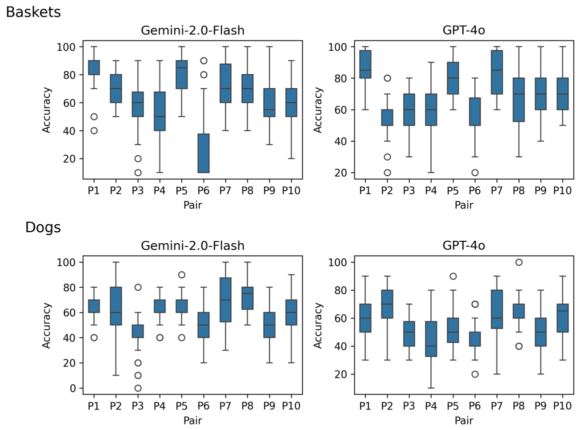


Figure 4: Accuracy boxplots of two best-performing LVLMs in the overhearer task for Round 1 conversations across 10 human pairs (whiskers denote 25th and 75th percentiles). Each boxplot represents 30 runs of a model, each with a different object ordering.

6.3 Robustness Analysis

An AI agent should understand not only the dynamically-evolving entrained-upon expression used by a human pair for a referent, but also that different pairs of human discourse participants entrain upon different expressions for the same referent. The agent should also be robust to object ordering in the input image. We tested matching performance on R1 for the two best-performing LVLMs, GPT-4o and Gemini-2.0-Flash, across human pairs, using different object orderings in each of 30 runs for each pair.

Figure 4 shows substantial performance variation of model performance with dialogue from different pairs of people and different object orderings. We use the difference between the 25th and 75th percentile of each boxplot as a proxy to measure performance variations and find that for the baskets (dogs) datasets, there is at least a 10.0% (10.0%) difference in these percentiles, with an average difference of 21.5% (18.5%).

Furthermore, both models perform better on some pairs' conversations than others. For example, GPT-4o performs better with human Pair 1 than Pair 2 for their R1 conversations about baskets, with a significant difference between the means of 32.0% ($p < 0.001$). See Table 10 in Appendix C.3 for exhaustive pairwise comparison across pairs.

We hypothesize that these performance variations may be caused by different levels of information density in transcripts of different human pairs. However, we find no significant correlation between the average model performance and five proxy features we use, namely, number of words, number of sentences, number of utterances, number of director turns, and number of matcher turns. See Table 11 in Appendix C.3 for details. We leave further analysis for future studies.

7 Follow-Up Experiments

We perform a series of follow-up studies to analyze factors that may affect LVLM's performance on our corpus, both in terms of a single round (R1) and over successive rounds (R1-R4). Given the relatively small size of our corpus, we focus on evaluating the *out-of-box* capabilities of current LVLMs, instead of performing finetuning. For the best-performing closed and open LVLMs, namely GPT-4o and Qwen-2.5-VL-32B, we run follow-up experiments that vary each of the factors below. Unless otherwise stated, we run each experiment from R1 through R4 and five times for each human pair, aligning with Section 5.

Textual Inputs We test this factor in four experiments. In the first two, we vary the text that the LVLM sees; in the next two we vary how the transcript is parceled out to the LVLM.

First, we remove colloquial features to make the transcripts read more like formal written text (**+Formal**). Second, we replace the text produced incrementally during interaction with summaries of object descriptions, removing any interactive features (**-Interaction**). We rewrite the transcripts using GPT-4.1 (OpenAI, 2025) for the two conditions and run the two LVLMs on the rewritten transcripts from R1 to R4. We inspected the output and found the generated quality acceptable for our intended use. The prompt templates for **+Formal** and **-Interaction** can be found in Appendix D.2.1 and Appendix D.2.2, respectively.

Third, instead of providing the entire transcript, we provide each LVLM with complete object descriptions one at a time, manually extracted as described in Section 4 (**ObjectDesc**). In contrast with the condition **-Interaction**, we maintain the interactive dialogue, but the LVLM only sees one object description at a time. This provides a control experiment with isolated descriptions, which helps disentangle whether the model performance failures are due to the overhearer setting or a more fundamental weakness in the models' basic grounding ability. Since 10 times more API calls are needed for this experiment, we run it only on R1.

Fourth, to test whether LVLMs can benefit from the “foresight” of accessing all transcripts at once (**AllTranscripts**), we prompt LVLMs to do the matching task for all R1-R4 transcripts at once, along with the corresponding 4 input images appended after each transcript.

Visual Inputs In the main experiments, LVLMs see 13 objects each time, in shuffled orders. In the follow-up experiments, the LVLMs see only the 10 target objects in same shuffled order (**OnlyTargets**), or the 13 objects with fixed order (**FixedOrder**). This is to determine whether either condition makes the task easier.

Feedback After an LVLM produces its answer at the end of each matching round, we provide it with the correct answers for all 10 objects in the set and prompt it to self reflect if its answers are not correct (**+Feedback**). This tests whether LVLMs can learn from feedback and improve over time.

Observations & Findings Table 3 shows the performance change under each condition relative to the baseline model performance from Section 6. Removing informal and interactive features of spontaneous conversations does not yield a significant performance difference for GPT-4o and Qwen-2.5-VL-32B, but including all rounds significantly *hurts* model performance from Round 2 on. This means that these two models cannot effectively use information from all transcripts to resolve references across different rounds and thus do not benefit from the “foresight” of seeing all transcripts upfront. This is also true when feedback is provided to help the model reflect on mistakes, which makes no significant difference. The last two points potentially explain the lack of a consistent performance improvement over rounds observed in Section 6. That is, they show that LVLMs do not accumulate knowledge across rounds, even in light of all information presented or feedback that reveals true answers.

Moreover, both models tend to show a significant and large performance gain when given object descriptions. This shows that identifying individual objects may not be the bottleneck that causes LVLMs to perform much worse than human matchers and fail to improve over rounds.

Finally, our follow-up experiments with visual inputs demonstrate that repeating the same input image for multiple rounds typically yields no significant difference, but, as expected, models do consistently benefit from doing the matching task with only the target objects (since they can choose from 10 objects rather than 13).

	Model	GPT-4o		Qwen-2.5-VL-32B	
	Source	Baskets	Dogs	Baskets	Dogs
Condition	Round				
+Formal	1	-4.0	+3.6	-0.2	-2.8
	2	+0.8	-2.4	-3.2	+1.4
	3	-3.0	-1.2	-2.8	-0.2
	4	-2.6	0.0	-0.2	+0.4
-Interaction	1	+4.2	+2.8	+5.2	+5.8**
	2	+4.8*	-6.2	-0.8	-4.8
	3	+0.6	-2.4	-2.4	-2.8
	4	+1.6	-1.0	-1.8	-6.0
ObjectDesc	1	+4.4	+15.8***	+17.0***	+13.4***
AllTranscripts	1	-0.8	-0.6	-2.2	-2.0
	2	-15.4***	-16.4***	-6.6**	-11.8***
	3	-19.8***	-14.0***	-9.2***	-17.2***
	4	-25.8***	-19.8***	-15.2***	-11.2**
OnlyTargets	1	+11.2**	+8.8*	+7.6**	-1.2
	2	+10.4**	-1.0	+7.0**	+5.4
	3	+5.2	+13.8***	+2.4	+4.0
	4	+13.6***	+12.8***	+12.4**	+8.4*
FixedOrder	1	-3.0	+0.8	-0.8	-1.4
	2	-0.6	-1.6	-6.8**	+2.4
	3	-5.0*	+1.8	-4.8	-4.0
	4	-2.2	-1.4	+1.6	0.0
+Feedback	1	+3.2	+1.6	+1.0	-0.2
	2	+1.8	+3.8	-0.2	0.0
	3	-0.4	-1.0	+0.4	-3.4
	4	-1.2	-1.8	-1.4	-2.0

Table 3: Results for the seven follow-up experiments in Section 7, each differing from the main experiments (baselines) in Section 6 by one factor. We highlight significant findings for both performance increase and decrease, relative to the baseline performance (see Figure 3) based on paired t-tests for each condition. See Section 7 for details of each condition.

8 Conclusion

Our findings demonstrate that modern LVLMs still struggle to resolve referring expressions to real-world objects produced during spontaneous conversation, a task that humans excel at when they can ground meanings together. Overhearers, whether human or LVLM, perform more poorly in a matching task than human addressees, even when they are present to every word of a conversation. LVLMs in the overhearer role, even state-of-the-art models, fail to exploit the dynamic nature of conversation and do not improve over repeated referring, unlike human overhearers. These limitations constrain the practical utility of LVLMs as embodied agents, while also highlighting clear directions for future improvement. Given that our primary goal is to benchmark current LVLM capabilities in this novel overhearing setting, providing mechanistic insights or finding pathways to solutions is beyond the scope of our paper; that, we leave to future studies. We release our corpus for reproducibility and to support continued research in this area.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2125295 and by a seed grant from Stony Brook University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Zhengxiang Wang and Owen Rambow were supported in part by funding from the Defense Advanced Research Projects Agency (DARPA) under Contracts No.HR01121C0186, No. HR001120C0037, and PR No. HR0011154158. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

Zhengxiang Wang and Owen Rambow are grateful for support from the Institute for Advanced Computational Science (IACS) at Stony Brook University, in particular the free GPT access it provides. Zhengxiang Wang was supported by IACS's Junior Researcher Award for the academic year of 2025.

The work benefited greatly from discussions with Gregory Zelinsky, Dimitris Samaras, Ritik Raina, and Amie Paige. We also thank the five anonymous reviewers for their valuable feedback.

Limitations

Due to resource constraints, our corpus is relatively small (80 dialogues from 10 pairs of human conversation partners). However, our corpus represents high-quality human-human spontaneous conversational data collected in a controlled experiment with a meaningful range of referential complexity (dogs and baskets) and without data contamination.

For the same reason, we were unable to include human baselines by conducting experiments with participants acting as overhearers for the same tasks. While we agree that this would have been ideal, we make no direct statistical comparisons between LVLMs and humans as overhearers in the paper. We also note that the observations that human overhearers improve in efficiency over time, and yet perform more poorly than actual participants in the dialogue, is well established in existing psycholinguistic literature (Schober and Clark, 1989; Fox Tree, 1999; Fox Tree and Mayer, 2008; Wilkes-Gibbs and Clark, 1992).

Additionally, the non-deterministic nature of large vision-language models (LVLMs) introduces challenges for evaluation; while we ran each experiment five times to reduce variability, further repetitions might yield different results.

Our study also does not exhaustively cover all existing LVLMs—particularly some open-weight models that may perform well on these tasks—but we believe the models we selected are representative of the current state-of-the-art.

Ethical Considerations

Human Subjects The corpus described here was collected with approval from Stony Brook University's Institutional Review Board, CORIHS (Committe on Research Involving Human Subjects) and with informed consent provided by the participants. The corpus contains no personally identifiable information.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Anthropic. 2025. Claude 3.7 sonnet and claude code. `https://www.anthropic.com/news/claude-3-7-sonnet`. Accessed: 2025-05-11.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Joshua P Gardner, Rohan Taori, and Ludwig Schmidt. 2023. VisIT-bench: A dynamic benchmark for evaluating instruction-following vision-and-language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Susan E. Brennan. 2005. How conversation is shaped by visual and spoken evidence. In J. Trueswell and M. Tanenhaus, editors, *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-Action Traditions*, pages 95–129. MIT Press, Cambridge, MA.

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482.

Emanuele Castano, Alison Jane Martingano, Gabriana Basile, Elly Bergen, and Evelyn Hye Kyung Jeong. 2023. Listening in to a conversation enhances theory of mind. *Current Research in Ecological and Social Psychology*, 4:100108.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In L. B. Resnick, J. Levine, and S. D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington, DC. Reprinted in R. M. Baecker (Ed.), *Groupware and Computer-Supported Cooperative Work: Assisting Human-Human Collaboration* (pp. 222–233). San Mateo, CA: Morgan Kaufman Publishers, Inc.

Herbert H. Clark and Catherine R. Marshall. 1981. Definite reference and mutual knowledge. In Aravind K. Joshi, Bonnie Webber, and Ivan Sag, editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge University Press, Cambridge.

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2023. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7348–7363, Toronto, Canada. Association for Computational Linguistics.

Jean E Fox Tree. 1999. Listening in on monologues and dialogues. *Discourse processes*, 27(1):35–53.

Jean E Fox Tree and Sarah A Mayer. 2008. Overhearing single and multiple perspectives. *Discourse Processes*, 45(2):160–179.

Google DeepMind. 2024. Introducing gemini 2.0: our new ai model for the agentic era. `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/`. Accessed: 2025-04-30.

Zeyu Han, Fangrui Zhu, Qianru Lao, and Huaizu Jiang. 2024. Zero-shot referring expression comprehension via structural similarity between images and captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14364–14374.

Yilun Hua and Yoav Artzi. 2024. Talk less, interact better: Evaluating in-context conversational adaptation in multimodal LLMs. In *First Conference on Language Modeling*.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. *arXiv preprint arXiv:2305.10160*.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. 2024. MMToM-QA: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16077–16102, Bangkok, Thailand. Association for Computational Linguistics.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2024. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45).

Robert M Krauss and Sam Glucksberg. 1969. The development of communication: Competence as a function of age. *Child development*, pages 255–266.

Robert M. Krauss and Sidney Weinheimer. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: a preliminary study. *Psychonomic Science*, 1(1–12):113–114.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. 2023. Evaluation and enhancement of semantic grounding in large vision-language models.

Charles Metzing and Susan E. Brennan. 2003. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49(2):201–213.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Introducing gpt-4.1 in the api. `https://openai.com/index/gpt-4-1/`. Accessed: 2025-05-11.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 381(2251):20220041.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Yanyuan Qiao, Chaorui Deng, and Qi Wu. 2021. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Michael J. Reddy. 1993. *The conduit metaphor: A case of frame conflict in our language about language*, page 164–201. Cambridge University Press.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *arXiv preprint arXiv:2310.18018*.

Michael F Schober and Herbert H Clark. 1989. Understanding by addressees and overhearers. *Cognitive Psychology*, 21(2):211–232.

Neelabh Sinha, Vinija Jain, and Aman Chadha. 2025. Guiding vision-language model selection for visual question-answering across tasks, domains, and knowledge types. In *Proceedings of the First Workshop of Evaluation of Multi-Modal Generation*, pages 76–94, Abu Dhabi, UAE. Association for Computational Linguistics.

Adil Soubki, John Murzaku, Arash Yousefi Jordehi, Peter Zeng, Magdalena Markowska, Seyed Abolghasem Mirroshandel, and Owen Rambow. 2024. Views are my own, but also yours: Benchmarking theory of mind using common ground. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14815–14823, Bangkok, Thailand. Association for Computational Linguistics.

Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Blaise Aguera y Arcas, and Robin I. M. Dunbar. 2024. Llms achieve adult human performance on higher-order theory of mind tasks. *Preprint*, arXiv:2405.18870.

Wei Su, Peihan Miao, Huanzhang Dou, and Xi Li. 2024a. Scanformer: Referring expression comprehension by iteratively scanning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13449–13458.

Wei Su, Peihan Miao, Huanzhang Dou, and Xi Li. 2024b. Scanformer: Referring expression comprehension by iteratively scanning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13449–13458.

Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. 2022. ReCLIP: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5198–5215, Dublin, Ireland. Association for Computational Linguistics.

Xiuchao Sui, Shaohua Li, Hong Yang, Hongyuan Zhu, and Yan Wu. 2023. Language models can do zero-shot visual referring expression comprehension.

Zineng Tang, Lingjun Mao, and Alane Suhr. 2024. Grounding language in multi-perspective referential communication. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19727–19741, Miami, Florida, USA. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1331 others. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Yunjie Tian, Tianren Ma, Lingxi Xie, and Qixiang Ye. 2025. Chatterbox: Multimodal referring and grounding with chain-of-questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7401–7409.

Deanna Wilkes-Gibbs and Herbert H Clark. 1992. Coordinating beliefs in conversation. *Journal of memory and language*, 31(2):183–194.

H Wimmer. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. 2024. Investigating compositional challenges in vision-language models for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14141–14151.

A Corpus

In addition to the 80 human-to-human transcribed dialogues, our corpus also comes with 13 pictures of baskets (Figure 5) and 13 pictures of dogs (Figure 6) that were discussed during the original human experiments and deployed in the LVLM overhearer experiments in this paper.

To quantify how referring becomes more efficient over time, we analyzed the average number of words used by directors and matchers across the four rounds for each object category (see Figure 7). Linear trend analyses revealed a significant decrease in word count as the rounds progressed. For basket images, directors' word counts significantly decreased from 39.94 words in the first round to 15.31 words in the final round, $F(1, 38) = 36.05$, $p < .001$. Matchers' word counts exhibited a similar pattern, decreasing from 15.04 to 3.83 words, $F(1, 38) = 19.21$, $p < .001$. Comparable trends were observed for dog images. Directors' word counts dropped from 46.05 to 13.84 words, $F(1, 38) = 37.96$, $p < .001$, while matchers' word counts decreased from 16.47 to 3.22 words, $F(1, 38) = 39.90$, $p < .001$.

We also examined whether the object category influenced word counts (a proxy for effort) by comparing the average words used per round for baskets and dogs. On average, pairs used 31.16 words per round to describe baskets and 32.25 for dogs. A paired sample t-test found this difference was not statistically significant, $t(9) = -0.34$, $p = 0.74$.

B Experiments

B.1 Details about LVLMs

The specific versions of LVLMs used in our study are as follows. We used the APIs from the respective LVLM providers for the first four proprietary LVLMs. The open-weight models were deployed locally through Hugging Face and run on 3 Nvidia RTX A6000s.

- GPT-4o: GPT-4O-2024-08-06
- GPT-4o-mini: GPT-4O-MINI-2024-07-18
- Gemini-2.0-Flash: Accessed in April 2025.
- Claude-3.7-Sonnet: CLAUDE-3-7-SONNET-20250219.
- Qwen-2.5-VL-7B: Hugging Face model card, QWEN/QWEN2.5-VL-7B-INSTRUCT
- Qwen-2.5-VL-32B: Hugging Face model card, QWEN/QWEN2.5-VL-32B-INSTRUCT
- Pixtral-12B: Hugging Face model card, MISTRALAI/PIXTRAL-12B-2409

B.2 Input Images

We generated 30 different orderings of objects for both baskets and dogs datasets, in addition to the two orderings shown in Figure 5 and Figure 6, respectively. These additional images share the same layout as the ones in Figures 5 and 6 and differ from the latter only in the specific order of the 13 objects.

To ensure that all LVLMs see the same input images across multiple rounds as well as across different human pairs to allow for paired t-tests, we create a playbook that stores the the specific input image for each run number and for each round of conversation (1-4). In the main experiments, we run each model on each transcript from each

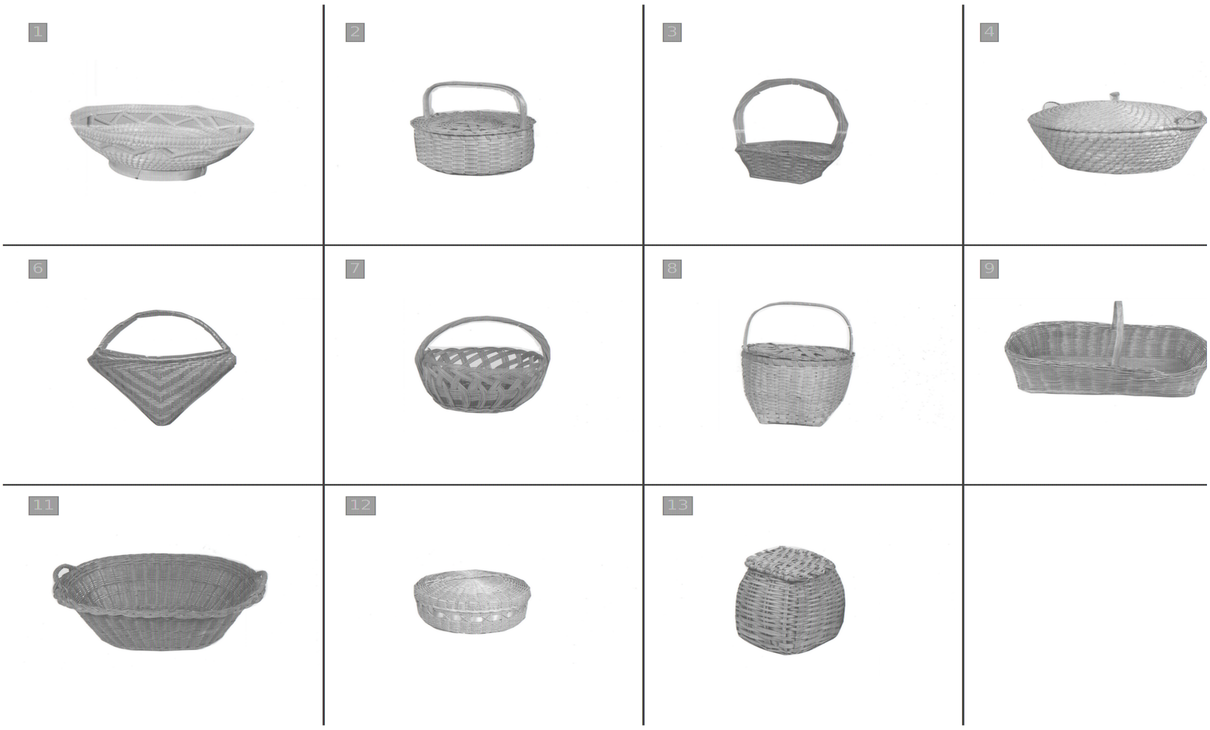


Figure 5: The 13 basket pictures from our corpus and an example input image for our experiments. The 10 target baskets are placed in the first two rows, numbered from 1 to 10, for illustration. Speakers in the original task did not see the numbers.

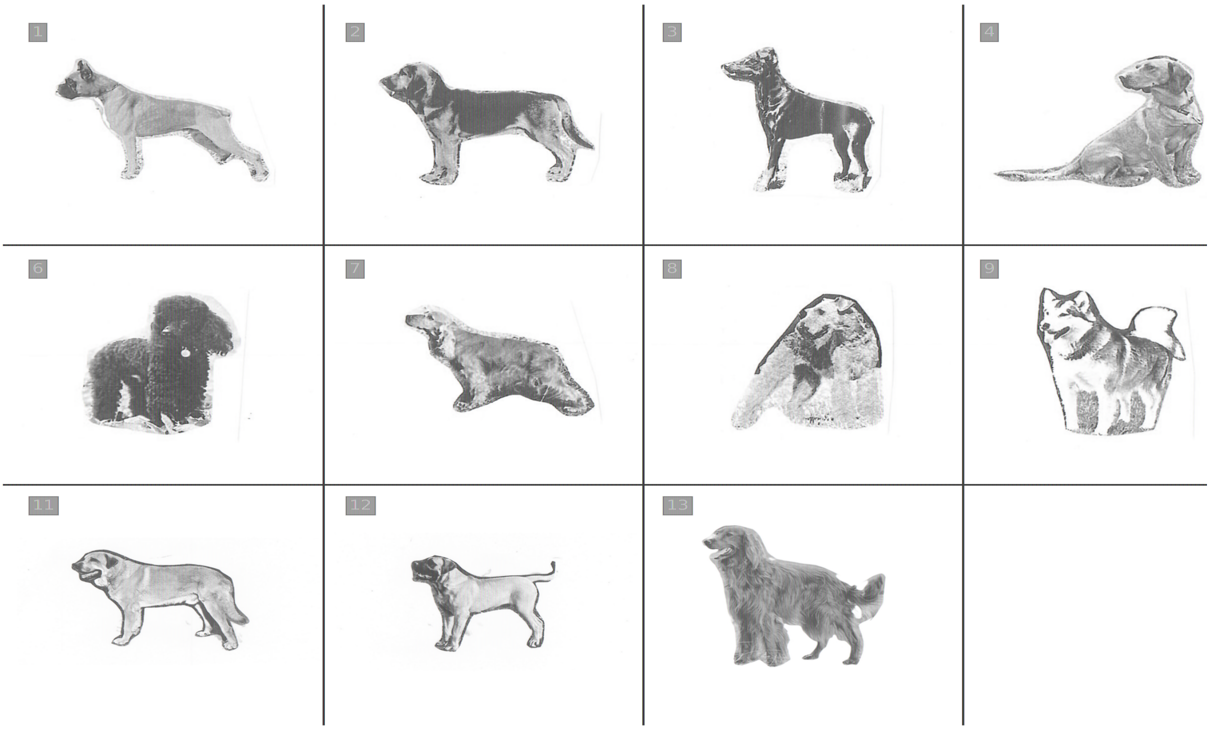


Figure 6: The 13 dog pictures from our corpus and an example input image for our experiments. The 10 target dogs are placed in the first two rows, numbered from 1 to 10, for illustration. Speakers in the original task did not see the numbers.

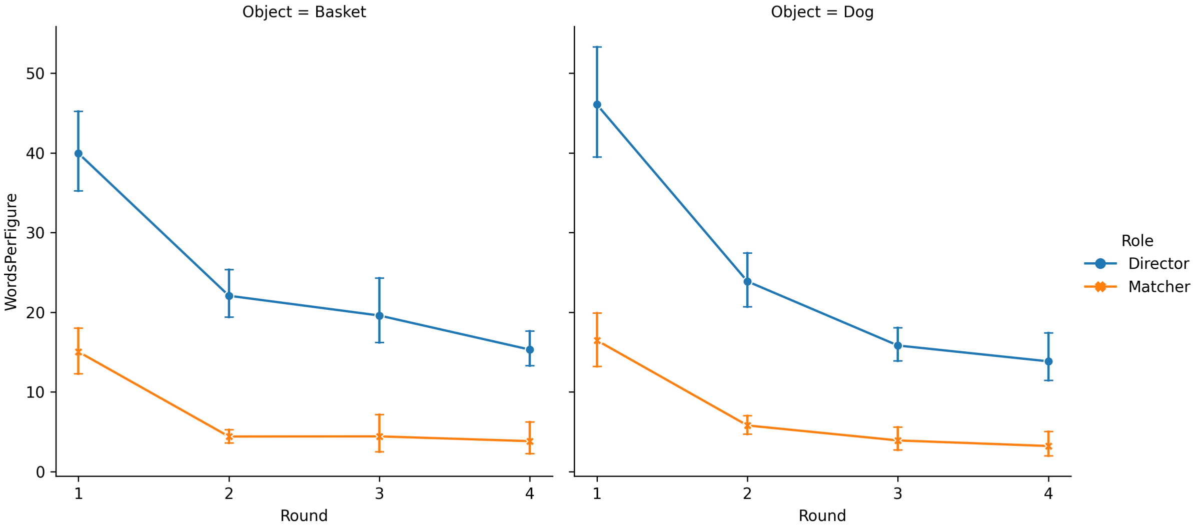


Figure 7: Mean number of words used per object across rounds for both Directors and Matchers. The plots for basket images (left) and dog images (right) both show a clear decrease in word count over time, demonstrating improved communicative efficiency. Error bars indicate ±1 standard error of the mean.

	Starting Round	R1	R2	R3
Source	Model			
Baskets	Claude-3.7-Sonnet	0.2*	0.3***	0.3**
	Gemini-2.0-Flash	0.2*	0.1	0.1
	GPT-4o	0.1*	0.1	0.1
	GPT-4o-mini	0.2**	0.2*	0.4***
	Qwen-2.5-VL-32B	0.1	-0.1	0.0
	Qwen-2.5-VL-7B	-0.3***	-0.4***	-0.1
	Pixtral-12B	0.1	-0.3***	0.3**
Dogs	Claude-3.7-Sonnet	0.2*	0.1	0.1
	Gemini-2.0-Flash	0.2**	0.1	0.3**
	GPT-4o	0.2**	0.0	0.2
	GPT-4o-mini	0.0	-0.0	-0.0
	Qwen-2.5-VL-32B	0.1	0.0	-0.0
	Qwen-2.5-VL-7B	0.0	0.2*	-0.2
	Pixtral-12B	-0.0	-0.0	-0.1

Table 4: Overall performance trend for each LVLM starting at R_i, using Spearman rank correlation. We highlight both **significant** positive and negative trends and use asterisks to denote different levels of significance, where "*" means $p < 0.05$, "**" means $p < 0.01$, and "**" means $p < 0.001$.

	Starting Round	R1	R2	R3
Source	Model			
Baskets	Claude-3.7-Sonnet	0.1*	0.3***	0.3**
	Gemini-2.0-Flash	0.1*	0.1	0.1
	GPT-4o	0.1*	0.1	0.1
	GPT-4o-mini	0.2**	0.2*	0.4***
	Qwen-2.5-VL-32B	0.1	-0.1	0.0
	Qwen-2.5-VL-7B	-0.2***	-0.4***	-0.1
	Pixtral-12B	0.1	-0.2***	0.3**
Dogs	Claude-3.7-Sonnet	0.1*	0.1	0.1
	Gemini-2.0-Flash	0.2**	0.1	0.2**
	GPT-4o	0.2**	0.0	0.1
	GPT-4o-mini	0.0	-0.0	-0.0
	Qwen-2.5-VL-32B	0.1	0.0	-0.0
	Qwen-2.5-VL-7B	0.0	0.2*	-0.2
	Pixtral-12B	-0.0	-0.0	-0.1

Table 5: Overall performance trend for each LVLM starting at R_i, using Kendall rank correlation. We highlight both **significant** positive and negative trends, using "*" to denote different levels of significance, where "*" means $p < 0.05$, "**" means $p < 0.01$, and "**" means $p < 0.001$.

human pair for five times and the model see exactly the same input image for each round number across 10 human pairs for both baskets and dogs datasets. The playbook was simply created by randomly sampling four distinct images from the generated images for 100 times, although we only ran each model up to 30 times in our study (see Section 6.3).

C Results

C.1 Overall Performance Trend

Table 4 and Table 5 shows the overall performance trend for each LVLM across baskets and dogs using Spearman rank correlation and Kendall rank correlation, respectively. The overall results are consistent with Table 1 in Section 6.

Table 7 shows the percentage of human pairs for whom LVLMs show a consistent improvement when starting at R_i. We use OLS regression to measure if there is a consistent improvement (i.e., significant positive coefficient).

Table 6 shows the percentage of time an LVLM's performance monotonically increases from a starting round (R1-R3) to the end round (R4) over all runs of each model. We report three levels of "monotonically increases": (1) monotonically increasing or non-decreasing, (2) monotonically increasing with a positive slope, meaning that model performance at the end round must surpass the starting round, and (3) strictly monotonically increasing, where we require that model performance on a round must be better than that of a previous round. As can be seen in Table 6, if we use the strictest measurement, none of the tested LVLMs show more than 6% strictly monotonically increasing performance.

C.2 Starting Early Versus Starting Late

Table 8 shows the pairwise mean differences of overlapping rounds between an early start and a late start. Here, the overlap rounds can be one round or multiple rounds. For example, the overlapping rounds between R1-R4 and R2-R4 are R2-R4, whereas the verlapping round between R1-R4 and R4 is just R4.

Table 9 shows the pairwise mean differences between an early start and a late start. Here, we are comparing two single starting rounds, such as R1 versus R2, R1 versus R4 etc.

C.3 Performance Difference between Human Pairs

Table 10 shows the mean model performance difference between different human pairs when doing the overhearer matching task based on the first rounds of conversations. We use paired t-tests to determine a significant difference.

We hypothesize that the performance variations may be due to different levels of information density in transcripts of different human pairs. We use five simple features to measure information density of a transcript: number of words, number of sentences, number of utterances, number of director turns, and number of matcher turns. We correlate these features with the average model performance and find that none of these correlations are statistically significant. See Table 11 for details.

Source	Starting Round Model	R1	R2	R3
Baskets	Claude-3.7-Sonnet	36.0 / 24.0 / 4.0	56.0 / 50.0 / 12.0	86.0 / 58.0
	Gemini-2.0-Flash	22.0 / 22.0 / 2.0	34.0 / 32.0 / 6.0	76.0 / 44.0
	GPT-4o	26.0 / 26.0 / 6.0	42.0 / 32.0 / 2.0	72.0 / 46.0
	GPT-4o-mini	18.0 / 16.0 / 0.0	36.0 / 32.0 / 6.0	84.0 / 66.0
	Qwen-2.5-VL-32B	10.0 / 10.0 / 2.0	26.0 / 16.0 / 6.0	62.0 / 40.0
	Qwen-2.5-VL-7B	6.0 / 4.0 / 0.0	14.0 / 2.0 / 0.0	62.0 / 22.0
	Pixtral-12B	18.0 / 16.0 / 2.0	22.0 / 8.0 / 0.0	82.0 / 44.0
Dogs	Claude-3.7-Sonnet	46.0 / 30.0 / 2.0	54.0 / 30.0 / 4.0	84.0 / 48.0
	Gemini-2.0-Flash	36.0 / 32.0 / 2.0	56.0 / 38.0 / 6.0	88.0 / 60.0
	GPT-4o	22.0 / 20.0 / 0.0	52.0 / 36.0 / 8.0	85.4 / 52.1
	GPT-4o-mini	12.0 / 12.0 / 4.0	26.0 / 26.0 / 6.0	60.4 / 31.2
	Qwen-2.5-VL-32B	14.0 / 14.0 / 2.0	22.0 / 20.0 / 8.0	60.0 / 40.0
	Qwen-2.5-VL-7B	10.0 / 10.0 / 0.0	44.0 / 34.0 / 18.0	54.0 / 22.0
	Pixtral-12B	28.0 / 28.0 / 0.0	42.0 / 30.0 / 2.0	64.0 / 36.0

Table 6: Percentage of time an LVLM's performance monotonically increases from a starting round (R1-R3) to the end round (R4). We report the following numbers in the table, separated by "/" in each cell: percentage of of monotonically increasing, percentage of of monotonically increasing with a positive slope (the model performance on the end round must be greater than the starting round), and percentage of strictly monotonically increasing (model performance on each round is strictly greater than that on the previous round). For a R3 start, there are only 2 rounds (R3 and R4), so the last two numbers are identical and we only report one in the R3 column. Again, proprietary LVLMs are overall more likely to show monotonically increasing performance.

Source	**Baskets**			**Dogs**		
	R1	R2	R3	R1	R2	R3
Claude-3.7-Sonnet	20	30	20	10	0	0
Gemini-2.0-Flash	30	10	0	10	0	10
GPT-4o	10	0	0	10	0	0
GPT-4o-mini	30	20	20	0	0	0
Qwen-2.5-VL-32B	0	0	0	20	0	10
Qwen-2.5-VL-7B	0	0	0	0	20	0
Pixtral-12B	0	0	10	0	0	0

Table 7: Percentage of human pairs for which LVLMs show a consistent improvement when starting at R_i.

C.4 Error Analyses

Figure 8 and Figure 9 shows the percentage of time an LVLM fails to identify a target object placed at a position index i across 13 positions and the specific object among 10 target objects for the two datasets, respectively. The two types of analyses are based only on experiments from R1 through R4 to control the effect of different starting rounds. However, our visualizations based on all the main experiments described in Section 5 show similar patterns.

		Late Start →	R2-R4	R3-R4	R4
Source	Model	Early Start ↓			
Baskets	Claude-3.7-Sonnet	R1-R4	1.4	11.6***	15.6***
		R2-R4	-	13.0***	19.0***
		R3-R4	-	-	8.0*
	Gemini-2.0-Flash	R1-R4	3.1*	10.2***	10.4**
		R2-R4	-	9.7***	12.2***
		R3-R4	-	-	2.0
	GPT-4o	R1-R4	8.4***	17.9***	15.6***
		R2-R4	-	9.2***	9.0**
		R3-R4	-	-	-0.2
	GPT-4o-mini	R1-R4	3.9**	4.4*	4.2
		R2-R4	-	3.6	0.0
		R3-R4	-	-	0.8
	Qwen-2.5-VL-32B	R1-R4	2.7	0.8	-1.4
		R2-R4	-	-1.8	-3.6
		R3-R4	-	-	1.4
	Qwen-2.5-VL-7B	R1-R4	-4.0**	-2.9*	-10.4***
		R2-R4	-	-1.2	-9.8***
		R3-R4	-	-	-8.6***
	Pixtral-12B	R1-R4	1.9	2.3	-1.2
		R2-R4	-	-1.1	-4.4**
		R3-R4	-	-	-1.2
Dogs	Claude-3.7-Sonnet	R1-R4	2.9	9.6***	16.4***
		R2-R4	-	7.0**	12.4***
		R3-R4	-	-	8.6**
	Gemini-2.0-Flash	R1-R4	2.3	15.9***	8.0**
		R2-R4	-	13.9***	8.8*
		R3-R4	-	-	-2.4
	GPT-4o	R1-R4	2.2	16.1***	11.4***
		R2-R4	-	15.9***	9.6**
		R3-R4	-	-	-2.1
	GPT-4o-mini	R1-R4	0.3	-4.4*	0.0
		R2-R4	-	-5.6**	-3.4
		R3-R4	-	-	2.9
	Qwen-2.5-VL-32B	R1-R4	0.9	3.8	10.2**
		R2-R4	-	0.5	7.8*
		R3-R4	-	-	8.0**
	Qwen-2.5-VL-7B	R1-R4	-12.8***	-5.2***	-4.0
		R2-R4	-	8.8***	12.8***
		R3-R4	-	-	1.0
	Pixtral-12B	R1-R4	-2.3	-1.3	-1.4
		R2-R4	-	-1.1	-2.0
		R3-R4	-	-	-0.2

Table 8: Pairwise mean differences of performance on overlapping rounds between an early start (R1-R4-R3-R4 in the rows) and a late start (R2-R4-R4 in the columns). For example, the overlapping rounds for R1-R4 and R2-R4 are R2-R3-R4, and for R2-R4 and R3-R4 they are R3-R4. We use paired t-tests to determine if there is a significant difference between two different starts and highlight both **significant** positive and negative mean differences. We indicate significance level using asterisks: "*" means $p < 0.05$, "**" means $p < 0.01$, and "**" means $p < 0.001$. In each round, there are 50 experiments (10 human pairs times 5 runs of an LVLM), so the degree of freedom is 100 times "number of overlapping rounds" minus 1.

		Late Start →	R2	R3	R4
Source	Model	Early Start ↓			
Baskets	Claude-3.7-Sonnet	R1	1.2	11.2***	9.8**
		R2	-	10.0***	8.6**
		R3	-	-	-1.4
	Gemini-2.0-Flash	R1	-7.0*	2.2	-0.4
		R2	-	9.2**	6.6**
		R3	-	-	-2.6
	GPT-4o	R1	4.8	14.0**	10.2**
		R2	-	9.2*	5.4
		R3	-	-	-3.8
	GPT-4o-mini	R1	-2.2	4.0	-8.4**
		R2	-	6.2*	-6.2*
		R3	-	-	-12.4***
	Qwen-2.5-VL-32B	R1	-6.4*	-4.8	-5.4
		R2	-	1.6	1.0
		R3	-	-	-0.6
	Qwen-2.5-VL-7B	R1	-9.6***	0.8	-4.8*
		R2	-	10.4***	4.8
		R3	-	-	-5.6*
	Pixtral-12B	R1	-4.6*	1.8	-3.6*
		R2	-	6.4***	1.0
		R3	-	-	-5.4**
Dogs	Claude-3.7-Sonnet	R1	-0.6	5.4	9.0**
		R2	-	6.0	9.6*
		R3	-	-	3.6
	Gemini-2.0-Flash	R1	-6.4	9.6**	-2.8
		R2	-	16.0***	3.6
		R3	-	-	-12.4***
	GPT-4o	R1	-6.4	9.4**	-0.8
		R2	-	15.8***	5.6
		R3	-	-	-10.2**
	GPT-4o-mini	R1	-4.4	-5.9*	-1.6
		R2	-	-1.5	2.8
		R3	-	-	4.3
	Qwen-2.5-VL-32B	R1	-4.8	-1.6	5.2
		R2	-	3.2	10.0*
		R3	-	-	6.8*
	Qwen-2.5-VL-7B	R1	-10.2	-8.6**	-2.6
		R2	-	1.6	7.6
		R3	-	-	6.0*
	Pixtral-12B	R1	-12.4*	0.0	0.0
		R2	-	12.4*	12.4*
		R3	-	-	0.0

Table 9: Pairwise mean differences between an early start (R1-R3 in the rows) and a late start (R2-R4 in the columns). We use paired t-tests to determine if there is a significant difference between two different starts and highlight both **significant** positive and negative mean differences. We indicate significance level using asterisks: “*” means $p < 0.05$, “**” means $p < 0.01$, and “**” means $p < 0.001$. In each round, there are 50 experiments (10 human pairs times 5 runs of an LVLM), so the degree of freedom is 99.

Source	Model	Pair2 → Pair1 ↓	P2	P3	P4	P5	P6	P7	P8	P9	P10
Baskets	GPT-4o	P1	32.0***	29.0***	25.0***	6.0**	31.7***	3.7	19.7***	17.7***	15.7***
		P2	-	-3.0	-7.0	-26.0***	-0.3	-28.3***	-12.3**	-14.3***	-16.3***
		P3	-	-	-4.0	-23.0***	2.7	-25.3***	-9.3*	-11.3**	-13.3***
		P4	-	-	-	-19.0***	6.7	-21.3***	-5.3	-7.3	-9.3*
		P5	-	-	-	-	25.7***	-2.3	13.7***	11.7***	9.7**
		P6	-	-	-	-	-	-28.0***	-12.0**	-14.0**	-16.0***
		P7	-	-	-	-	-	-	16.0***	14.0***	12.0***
		P8	-	-	-	-	-	-	-	-2.0	-4.0
		P9	-	-	-	-	-	-	-	-	-2.0
	Gemini-2.0-Flash	P1	12.7***	26.3***	28.3***	3.7	55.0***	13.0***	14.7***	25.7***	23.7***
		P2	-	13.7**	15.7**	-9.0*	42.3***	0.3	2.0	13.0***	11.0**
		P3	-	-	2.0	-22.7***	28.7***	-13.3**	-11.7**	-0.7	-2.7
		P4	-	-	-	-24.7***	26.7***	-15.3**	-13.7**	-2.7	-4.7
		P5	-	-	-	-	51.3***	9.3**	11.0*	22.0***	20.0***
		P6	-	-	-	-	-	-42.0***	-40.3***	-29.3***	-31.3***
		P7	-	-	-	-	-	-	1.7	12.7***	10.7*
		P8	-	-	-	-	-	-	-	11.0**	9.0*
		P9	-	-	-	-	-	-	-	-	-2.0
Dogs	GPT-4o	P1	-10.0**	10.0**	13.7***	5.0	12.3***	-6.0	-5.3	8.0*	-3.3
		P2	-	20.0***	23.7***	15.0***	22.3***	4.0	4.7	18.0***	6.7
		P3	-	-	3.7	-5.0	2.3	-16.0***	-15.3***	-2.0	-13.3***
		P4	-	-	-	-8.7**	-1.3	-19.7***	-19.0***	-5.7*	-17.0***
		P5	-	-	-	-	7.3*	-11.0*	-10.3**	3.0	-8.3**
		P6	-	-	-	-	-	-18.3***	-17.7***	-4.3	-15.7***
		P7	-	-	-	-	-	-	0.7	14.0**	2.7
		P8	-	-	-	-	-	-	-	13.3***	2.0
		P9	-	-	-	-	-	-	-	-	-11.3***
	Gemini-2.0-Flash	P1	6.3	26.0***	2.4	3.3	18.6***	0.0	-7.0**	16.7***	8.3*
		P2	-	19.7***	-3.1	-3.0	12.4**	-6.3	-13.3**	10.3*	2.8
		P3	-	-	-22.1***	-22.7***	-8.6*	-26.0***	-33.0***	-9.3*	-17.6***
		P4	-	-	-	0.3	16.1***	-2.1	-9.7**	13.8***	5.4
		P5	-	-	-	-	15.2***	-3.3	-10.3***	13.3***	5.5
		P6	-	-	-	-	-	-18.6***	-25.5***	-1.4	-8.9**
		P7	-	-	-	-	-	-	-7.0*	16.7***	8.3
		P8	-	-	-	-	-	-	-	23.7***	14.8***
		P9	-	-	-	-	-	-	-	-	-9.0*

Table 10: Mean model performance difference between human pair 1 and pair 2 when doing the overhearer matching task based on the first rounds of conversations for 30 runs, each with a different object ordering. We use paired t-tests to determine if there is a significant difference between each human pair and highlight both **significant** positive and negative mean differences. We indicate significance level using asterisks: “*” means $p < 0.05$, “**” means $p < 0.01$, and “**” means $p < 0.001$. In each comparison, there are 30 experiments for each human pair, so the degree of freedom is 59.

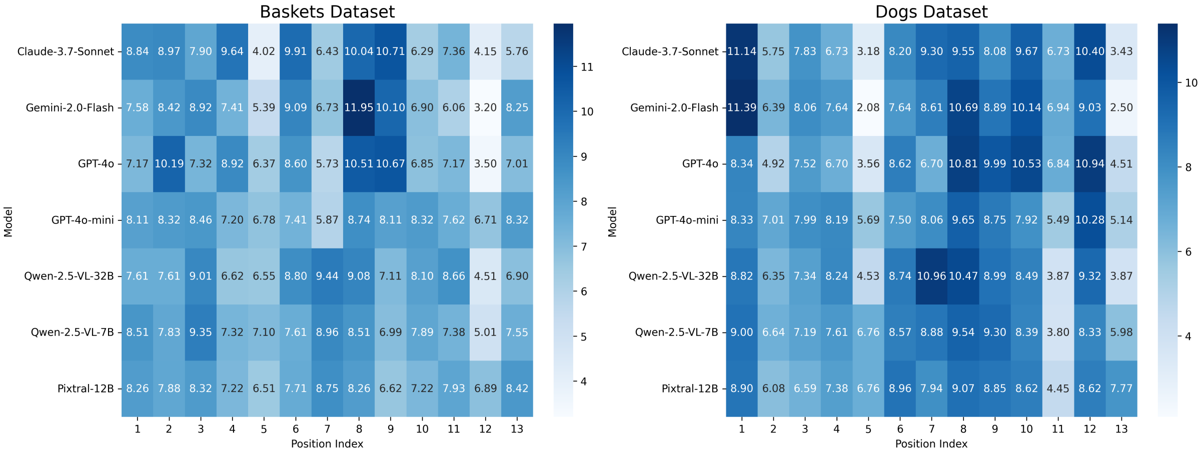


Figure 8: Percentage of times an LVLM fails to identify the correct object at index i for the two datasets. The position index ranges from 1 to 13, since there are 13 objects for the overhearer and the target objects are shuffled across these positions.

Source	Model	Feature	Kendall Tau	Kendall p-value	Spearman R	Spearman p-value
Baskets	Gemini-2.0-Flash	# Words	-0.02	1.00	0.02	0.96
		# Sentences	-0.38	0.16	-0.52	0.13
		# Utterances	-0.16	0.60	-0.18	0.63
		# Director Turns	-0.37	0.15	-0.46	0.18
		# Matcher Turns	-0.40	0.11	-0.49	0.15
	GPT-4o	# Words	0.11	0.73	0.19	0.60
		# Sentences	-0.16	0.60	-0.18	0.63
		# Utterances	-0.02	1.00	-0.04	0.91
		# Director Turns	-0.18	0.47	-0.24	0.51
		# Matcher Turns	-0.22	0.37	-0.29	0.41
Dogs	Gemini-2.0-Flash	# Words	0.18	0.47	0.29	0.41
		# Sentences	-0.04	0.86	-0.01	0.99
		# Utterances	0.18	0.47	0.24	0.50
		# Director Turns	0.18	0.47	0.25	0.49
		# Matcher Turns	0.18	0.47	0.24	0.50
	GPT-4o	# Words	0.38	0.16	0.53	0.12
		# Sentences	0.24	0.38	0.26	0.47
		# Utterances	0.38	0.16	0.50	0.14
		# Director Turns	0.20	0.48	0.28	0.43
		# Matcher Turns	0.16	0.53	0.21	0.57

Table 11: Correlations between average model performance based the 30 runs of an LVLM on R1 transcripts and the features of the related transcripts. None of these correlations are significant.

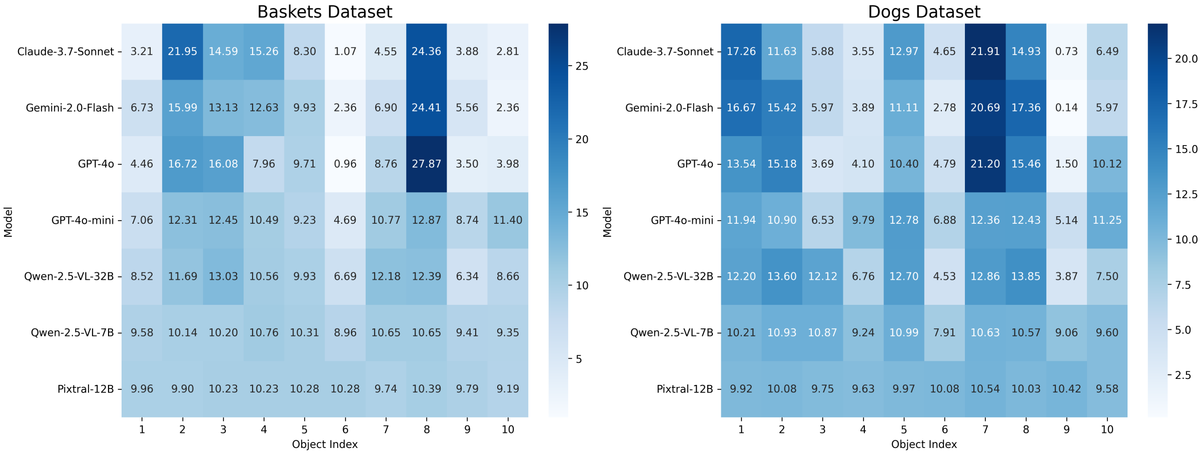


Figure 9: Percentage of times an LVLM fails to identify each one of the 10 target objects for the two datasets. The object index corresponds to those in Figure 5 for baskets and in Figure 6 for dogs.

D Prompt Templates

This section provides prompt templates for all experiments in this study. We use “$” followed by a word to denote a placeholder.

D.1 Prompt Template for Main Experiments

Figure 10 shows the prompt template used for the main experiments in Section 6. Each prompt contains a system prompt and a user prompt, where the system prompt is provided only once in the very beginning, whereas the user prompt may be used multiple times in case of a multi-turn conversation (the starting round is other than R4). Note that in the system prompts, “$example_sequence” (default=“3, 7, 1, 12, 5, 2, 13, 8, 10, 4, 6, 9, 11”) is fixed for a given value of “$num_of_objects” (default=13) to maximize prompt similarity across different prompting conditions.

D.2 Prompt Templates for Follow-Up Experiments

We prompted GPT-4.1-2025-04-14 to remove the colloquial and interactive features from our corpus.

D.2.1 Prompt Template for Removing Colloquial Features

> You are given an excerpt from a transcribed, spontaneous conversation between two individuals. Your task is to revise the excerpt to produce a clear, polished version of the dialogue that reads like formal written text. Transform any standalone words or phrases into complete, grammatically correct sentences where appropriate. Do not add any additional information or context or change the meaning of the text. Do not output anything other than the revised excerpt.
>
> Here is the excerpt: $excerpt
>
> Revised Excerpt:

D.2.2 Prompt Template for Creating Object Summaries

> You are given an excerpt from a transcribed, spontaneous conversation between two individuals. Your task is to extract and concisely summarize all descriptions used to characterize a specific object mentioned in the excerpt. You must follow the instructions below:
>
> 1. Preserve all relevant descriptive details.
>
> 2. Do not alter the meaning, add context, or introduce new information.
>
> 3. Your response must only include the final summary—do not include the original excerpt or any explanatory text.
>
> Excerpt:
>
> $excerpt
>
> Summary of Object Descriptions:

D.2.3 Prompt Template for Providing Transcripts without Colloquial Features

We re-use the same prompt template from Figure 10 when providing LVLMs with transcripts with colloquial features removed (**+Formal**).

Appendix D.2.1 shows the prompt template used for removing colloquial features from a transcript.

D.2.4 Prompt Template for Providing Object Summaries

Figure 11 shows the prompt template we used for providing LVLMs with all 10 target object summaries at once, instead of the original transcript **-Interaction**). Each prompt contains a system prompt and a user prompt, where we repeat the user prompt four times, containing 10 object summaries and the related input image from each round in order. We then concatenate these repeated user prompts together to prompt the LVLMs.

Appendix D.2.2 shows the prompt template used for creating object summaries from a given transcript.

D.2.5 Prompt Template for Providing One Object Description at a time

Figure 12 shows the prompt template we used for providing a complete and manually segmented description of a target object (**ObjectDesc**) one at a time.

D.2.6 Prompt Template for Providing All Transcripts At Once

Figure 13 shows the prompt template we used for providing LVLMs with all transcripts at once (**AllTranscripts**). Each prompt contains a system prompt and a user prompt, where we repeat the user prompt four times, containing the transcript and the related input image from each round in order. We then concatenate these repeated user prompts together to prompt the LVLMs.

D.2.7 Prompt Template for Providing LVLMs with Feedback

We re-use the same prompt we used in the main experiments in Section 6, as shown in Figure 10. After an LVLMs produces its answer to each round, we insert the following prompt with the correct target sequence before proceeding to the next round of the matching task.

You are an overhearer of a conversation between two participants engaged in a collaborative object-matching task for one or multiple rounds. Each participant is in a separate room and has a duplicate set of pictures arranged in different random orders. They cannot see each other's sets and communicate solely via an audio link. During the task, one participant acts as the Director (D) and the other as the Matcher (M). The Director describes the pictures one at a time, and the Matcher selects the corresponding picture from their own set. Please note that it is the same two participants playing the same roles for all the rounds if there are multiple rounds.

As the overhearer and for each round, you are provided with:

- The full transcript of their conversation for that round.
- An image showing all pictures used in the task, randomly arranged and labeled with indices from 1 to $num_of_objects. The image for each round may be different.

Your goal is to determine the correct sequence of picture indices as described by the Director during each round. To do this:

1. Carefully analyze the transcript to understand which pictures the Director refers to, in the order they were described.
2. Use the image to match each described picture to its corresponding index.
3. Think step by step and revise your reasoning and answers as needed. However, you may not ask questions or make assumptions beyond the given materials.

When you reach your conclusion, output your response in the following format:

Final Answer: [$num_of_objects picture indices in correct order, separated by commas]

Example: $example_sequence

The transcript of the current conversation is as follows:

$transcript

The image for the current round showing the pictures is as follows:

<$image_path>

Figure 10: Prompt template for the experiments where the transcripts are presented one at a time in a multi-turn conversation. The user prompt at the bottom is repeated up to 4 times before each transcript/image pair in our experiments.

Here is correct sequence of picture indices as described by the Director: $answer. Reflect on your previous answer if it was wrong. We will proceed after your reflection.

You are an overhearer of a conversation between two participants engaged in a collaborative object-matching task for one or multiple rounds. Each participant is in a separate room and has a duplicate set of pictures arranged in different random orders. They cannot see each other"s sets and communicate solely via an audio link. During the task, one participant acts as the Director (D) and the other as the Matcher (M). The Director describes the pictures one at a time, and the Matcher selects the corresponding picture from their own set. Please note that it is the same two participants playing the same roles for all the rounds if there are multiple rounds.

As the overhearer and for each round, you are provided with:

- 10 object summaries based on the Director's description of the target pictures for that round.
- An image showing all pictures used in the task, randomly arranged and labeled with indices from 1 to $num_of_objects. The image for each round may be different.

Your goal is to determine the correct sequence of picture indices as described by the Director during each round. To do this:

1. Carefully analyze the transcript to understand which pictures the Director refers to, in the order they were described.
2. Use the image to match each described picture to its corresponding index.
3. Think step by step and revise your reasoning and answers as needed. However, you may not ask questions or make assumptions beyond the given materials.

When you reach your conclusion, output your response in the following format:

Final Answer: [$num_of_objects picture indices in correct order, separated by commas]

Example: $example_sequence

The 10 object summaries based on the Director's description are as follows:

$summaries

The image for the current round showing the pictures is as follows:

<$image_path>

Figure 11: Prompt template for the experiments where the summaries of 10 target object summaries from a given transcript are presented, instead of the original transcript (**-Interaction**), one at a time in a multi-turn conversation. The user prompt at the bottom is repeated up to 4 times before each transcript/image pair in our experiments.

You are an overhearer of an ongoing conversation between two participants engaged in a collaborative object-matching task. Each participant is in a separate room and has a duplicate set of pictures arranged in different random orders. They cannot see each other"s sets and they can communicate solely via an audio link. During the task, one participant acts as the Director (D) and the other as the Matcher (M). The Director describes the pictures one at a time, and the Matcher selects the corresponding picture from their own set.

As the overhearer and for each target picture, you are provided with:

- An image showing all pictures used in the task, randomly arranged and labeled with indices from 1 to $num_of_objects.
- Conversation between the Director (D) and Matcher (M) where the Matcher indicates that they have selected a target picture.

Your goal is to determine the correct sequence of picture indices as described by the Director during the task. To do this:

1. Carefully analyze each conversation to understand which pictures the Director refers to, in the order they were described.
2. Use the image to match each described picture to its corresponding index.

3. Think step by step and revise your reasoning and answers as needed. However, you may not ask questions or make assumptions beyond the given materials.

You should produce a target picture index for each conversation presented to you as your current best guess. Once all the 10 pictures have been selected by the Matcher, you should reach a final conclusion and output your response in the following format:

Final Answer: [$num_of_objects picture indices in correct order, separated by commas]

Example: $example_sequence

The image showing the pictures is as follows:

<$image_path>

The conversation between the Director (D) and Matcher (M) for the first target picture is as follows:

$conversation

Figure 12: Prompt template for the experiments where a complete and manually segmented description of a target object (**ObjectDesc**) is provided one at a time in a multi-turn conversation. The user prompt at the bottom is repeated up to 10 times before each one of the 10 target objects' descriptions extracted from the transcript.

You are an overhearer of a conversation between two participants engaged in a collaborative object-matching task for multiple rounds. Each participant is in a separate room and has a duplicate set of pictures arranged in different random orders. They cannot see each other's sets and communicate solely via an audio link. During the task, one participant acts as the Director (D) and the other as the Matcher (M). The Director describes the pictures one at a time, and the Matcher selects the corresponding picture from their own set. Please note that it is the same two participants playing the same roles for all the rounds.

As the overhearer, you are provided with:

- The full transcripts of their conversation for each round.
- Images showing all pictures used in the task for each round, randomly arranged and labeled with indices from 1 to $num_of_objects.

Your goal is to determine the correct sequence of picture indices as described by the Director for each round. To do this:

1. Carefully analyze the transcript to understand which pictures the Director refers to, in the order they were described.
2. Use the image to match each described picture to its corresponding index.
3. Think step by step and revise your reasoning and answers as needed. However, you may not ask questions or make assumptions beyond the given materials.

When you reach your conclusion, output your response in the following JSON format for each round:

Final Answer: {“Round i”: [$num_of_objects picture indices in correct order, separated by commas]}

Example: {“Round 1”: [$example_sequence], ...}

The transcript of the conversation during round#$ix is as follows:

$transcript

The image for the round#$ix showing the pictures is as follows:

<$image_path>

Figure 13: Prompt template for the experiments where multiple transcripts are presented together (**AllTranscript**) in a single-turn conversation. While we repeat the user prompt at the bottom four times for the four rounds, they are concentrated together and passed to an LVLM all at once.