# When Seeing Is Not Enough: Benchmarking Interactive Visual Grounding in LVLMs

**Zhengxiang Wang**
Department of Linguistics & IACS
Stony Brook University
zhengxiang.wang@stonybrook.edu

**Owen Rambow**
Department of Linguistics & IACS
Stony Brook University
owen.rambow@stonybrook.edu

## Abstract

Visual grounding is typically evaluated as a one-shot mapping from an informative referring expression to a visual target. This formulation misses a central property of real-world reference: initial referring expressions are often incomplete or ambiguous, requiring participants to establish shared understanding through interaction. We introduce a controlled evaluation framework for interactive visual grounding in large vision-language models (LVLMs), varying how much target information is provided upfront and how much must be acquired through dialogue. Across four human-grounded visual contexts and four interaction protocols, current LVLMs perform significantly below task-level human baselines. Interaction can help when follow-up questions refine or repair an initial target description. Performance is lowest when no initial description is provided and target information must be acquired through questions, indicating that proactive question-driven grounding remains difficult. LVLMs are also poorly calibrated, often reporting confidence that exceeds their empirical accuracy. Follow-up studies confirm these patterns across varied description sources (human versus AI), reasoning efforts, repeated interactions, description providers, and visual contexts. Overall, interactive visual grounding remains challenging, requiring visual matching, information seeking and synthesis.[1]

## 1 Introduction

Visual grounding is a core vision-language capability that requires a system to identify and localize regions or objects in a visual environment from a textual description (Xiao et al., 2026). It underlies applications such as visual question answering, visual retrieval, human–AI interaction, and embodied AI agents. Enabled by large-scale cross-modal training and alignment (Radford et al., 2021; Alayrac et al., 2022; Liu et al., 2023), recent large vision-language models (LVLMs) have demonstrated strong general-purpose vision-language capabilities (Wang et al., 2024; OpenAI et al., 2024; Gemini Team et al., 2025); in particular, they have largely saturated standard visual grounding benchmarks (Chen et al., 2024; Dong et al., 2026).

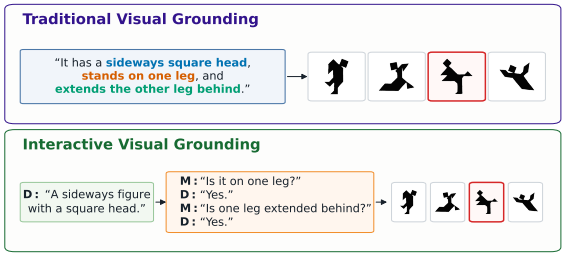


Figure 1: Traditional visual grounding performs one-shot matching, assuming an informative input description. Interactive visual grounding removes this assumption, requiring the Matcher (M) to interact with the Director (D) to seek more information. Examples are adapted from Hawkins et al. (2020) to be illustrative.

However, traditional visual grounding (Figure 1 upper part) relies on a restrictive assumption: the model is given a sufficiently informative referring expression before grounding begins. Most benchmarks therefore cast grounding as a one-shot, non-interactive task, where the model maps an already available description to a visual target (Mao et al., 2016). This setup primarily evaluates *static visual matching*. It does not test whether a model can recognize that a description is incomplete, ask for missing information, or update its grounding decision as new evidence becomes available.

This assumption departs from how reference works in real-world human communication (Figure 1, lower part). Referring expressions are frequently underspecified, ambiguous, or sufficient only relative to a particular visual context. Speakers and listeners establish reference collaboratively,

[1]GitHub project: https://github.com/jaaack-wang/benchmarking-interative-visual-grounding.



*Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 11212–11240
October 24-29, 2026 ©2026 Association for Computational Linguistics

using clarification questions and feedback to build common ground over time (Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996).

We therefore study *interactive visual grounding*, where target information may be ambiguous or underspecified at first and must be established through interaction. We argue that interactive visual grounding requires two broad capabilities. First, models need *visual matching* ability: mapping a referring expression, or an accumulated target representation, to the intended visual item. Second, models need *information seeking and synthesis* ability: assessing whether the current information is sufficient, asking effective questions when it is not, and synthesizing answers across turns. Traditional grounding benchmarks primarily test the former; interactive grounding requires both.

Prior studies evaluate non-interactive referring-expression comprehension or object matching (Tang et al., 2024; Wang et al., 2025a), and interactive reference tasks that involve human–AI communication (Hua and Artzi, 2024; Imai et al., 2025; Zeng et al., 2026; Jones et al., 2026). These studies reveal important limitations, but without systematically controlling whether target information is supplied or must be acquired, they cannot disentangle visual-matching limitations from failures of interactive information seeking. Consequently, aggregate performance can mask fundamentally different failure modes and offer little guidance for model improvement.

To address this issue, we introduce a controlled evaluation framework for interactive visual grounding. Our framework varies how much target information is available upfront and how much must be acquired through dialogue, thereby testing whether models can ground informative descriptions, use optional clarification, recover from underspecification, and construct target representations through their own questions. Across 38,848 target selections from eight LVLMs under diverse experimental conditions, current models perform significantly below task-level human baselines even when additional target information is available. They can use follow-up questions to refine or repair an initial description, but struggle to ground targets through proactive question-driven dialogue despite asking the most questions. Their reported confidence also frequently exceeds empirical accuracy. Together, these findings show that interactive visual grounding remains an important challenge for LVLMs beyond static visual matching.

## 2 Related Work

**Static Visual Grounding** Most visual grounding benchmarks evaluate grounding as a static identification or localization problem: given an image and a referring expression, a model selects the corresponding region or object (Lin et al., 2014; Kazemzadeh et al., 2014; Plummer et al., 2015; Yu et al., 2016; Chen et al., 2024; Dong et al., 2026). This formulation has been central to measuring visual-language alignment, but it assumes that the referring expression is already sufficiently informative. As a result, it primarily tests visual matching, rather than whether a model can recognize missing information or seek clarification.

**Interactive Visual Grounding** Prior work has introduced interaction into visual grounding through goal-oriented visual dialogue and clarification-based object identification. GuessWhat?! frames object identification as a single-target game in which a questioner asks yes/no questions to identify an object (de Vries et al., 2017). INGRESS studies situated referring-expression grounding, where a robot asks clarification questions when a referent is ambiguous (Shridhar et al., 2020). InViG benchmarks open-ended multi-turn grounding for object disambiguation (Zhang et al., 2023). These settings show that interaction is important for grounding, but they typically focus on single-target identification or constrained question-answering. In contrast, we study multi-target grounding under controlled information conditions, allowing us to distinguish visual matching from information-seeking ability.

**LVLMs in Reference Tasks** Recent studies have evaluated LVLMs in repeated, multi-target reference tasks, building on classic work on referential communication and common ground (Krauss and Glucksberg, 1969; Clark and Wilkes-Gibbs, 1986; Schober and Clark, 1989). Some work examines whether LVLMs can comprehend referring expressions in human-human dialogues that become more efficient over repeated interaction (Wang et al., 2025a). Other work studies LVLM behavior in self-play (Hua and Artzi, 2024; Imai et al., 2025) or human–AI interaction (Zeng et al., 2026; Jones et al., 2026), revealing non-human-like patterns in both generation and comprehension. Our work differs by systematically controlling how much target information is provided upfront and how much must be acquired through dialogue. This lets us test whether model failures arise from visual matching

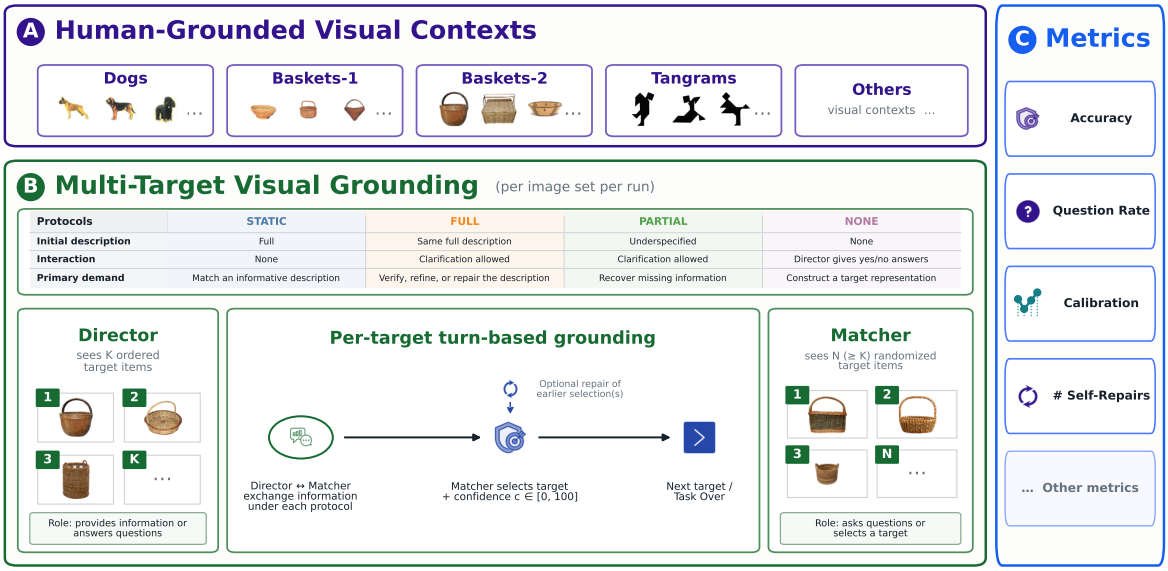


Figure 2: Overview of our proposed framework for controlled interactive visual grounding evaluation. (A) and (C) show the visual contexts and metrics used in our main experiments, respectively; (B) summarizes the interaction protocols and grounding process. The "Others" entries in (A) and (C) highlight the framework's extensibility.

or information seeking or their combination.

**Multi-Turn Evaluation** Multi-turn evaluation has been increasingly used to evaluate LLMs beyond isolated single-turn prompts (Zheng et al., 2023; Bai et al., 2024; Kwan et al., 2024; Wang et al., 2025b; Han et al., 2025). Recent work shows that LLMs can degrade across turns and often fail to seek clarification, causing misunderstandings to compound over interaction (Laban et al., 2026; Shaikh et al., 2025). Multimodal multi-turn evaluation extends these challenges to LVLMs, where models must track visual inputs, dialogue history, and evolving user intent (Liu et al., 2024; Lee et al., 2025; Xue et al., 2025; Feng et al., 2023; Tian et al., 2025). Our work instead tests whether models can use dialogue to determine and acquire missing target information and improve grounding accuracy.

## 3 Interactive Visual Grounding

Figure 2 visualizes our proposed evaluation framework, connecting the visual contexts (Panel A), multi-target visual grounding (Panel B), and evaluation metrics (Panel C). This section details the core designs for the overall task, the interaction protocols, and the evaluation metrics.

**Task Formalization** A *Director* sees an ordered target set $D = (d_1, \ldots, d_k)$; a *Matcher* sees a separately indexed candidate set $M = (m_1, \ldots, m_n)$ containing the targets and possibly additional distractors. The players do not see each other's view, and indices are randomized independently.

For each target position $j$, the players exchange utterances until the Matcher selects $\hat{d}_j \in M$. The Matcher may first ask clarification questions and may later revise a selection. The final output is the ordered sequence $\hat{D} = (\hat{d}_1, \ldots, \hat{d}_k)$. Thus, the task adds clarification, cross-turn integration, and revision to one-shot matching under partial observability (Clark and Wilkes-Gibbs, 1986).

**Interaction Protocols** We instantiate four protocols that vary how much target information is provided upfront and how much the Matcher must acquire through interaction.

In STATIC, the Matcher receives one full human-produced target description at a time and selects without interaction. FULL provides the same description but permits optional follow-up questions, testing whether clarification can verify, refine, or repair an initially informative input. "Full" means that the description comes from a high-accuracy human dialogue and is therefore expected to be sufficiently informative, not that it exhaustively specifies the target. PARTIAL instead begins with an underspecified attribute, requiring the Matcher to seek missing information. NONE provides no initial description; the Director returns only yes/no/unclear answers, so the Matcher must decide what to request and construct a target representation through questions (de Vries et al., 2017). Together, the pro-

| Image Set | Source | Dataset Size | | | Context Properties | | | |
|---|---|---|---|---|---|---|---|---|
| | | # Items | # Targets | # Dialogues | Matching | Modality | Object Type | Nameable |
| Dogs | W | 13 | 10 | 10 | Free-form | SPOKEN | REAL | ✓ |
| Baskets 1 | W | 13 | 10 | 10 | Free-form | SPOKEN | REAL | – |
| Baskets 2 | Z | 18 | 12 | 14 | Free-form | TEXT | REAL | – |
| Tangrams | H | 12 | 12 | 14 | Sequential | TEXT | ABSTRACT | – |

Table 1: Dataset statistics and key properties of the four visual contexts used in our interactive visual grounding experiments. Source: W = Wang et al. (2025a); Z = Zeng et al. (2026); and H = Hawkins et al. (2020). In the source human tasks, free-form matching permitted revisions, whereas sequential matching proceeded one target at a time without revision; our simulated interaction rules are described in Figure 2 and Section 3.

tocols progress from matching an informative description (STATIC) to optional clarification (FULL), recovery from underspecification (PARTIAL), and fully question-driven grounding (NONE).

A description drawn from a high-performing human dialogue may still leave an LVLM or even another human uncertain or with an incorrect prediction, motivating optional interaction even in the FULL condition.

**Evaluation Metrics** We primarily consider **accuracy** and **question rate** as our main metrics, reflecting models' visual matching ability and tendency to seek information in interactive settings. Accuracy is the percentage of correctly matched items within a task round. Question rate is the average number of questions asked per item; by definition, it is 0 under STATIC. As a baseline, random accuracy is $1/n$ for a Matcher view with $n$ candidates.

We also analyze **calibration** to assess whether models know when they have enough information to make a selection. Each Matcher is prompted to report a confidence score with its final selection, and we compare these stated confidence scores against empirical accuracy using calibration plots.[2]

Finally, we consider **self-repairs** as a secondary behavioral signal. Self-repairs occur when a Matcher revises an earlier selection after receiving additional information, indicating whether models can recognize and correct earlier grounding decisions during interaction.

## 4 Experimental Setup

### 4.1 Datasets

We evaluate interactive visual grounding across four human-grounded, object-level visual contexts (Table 1): dogs and baskets from Wang et al. (2025a), baskets from Zeng et al. (2026), and tangrams from Hawkins et al. (2020). Appendix H provides more details on the four image sets, and Panel A of Figure 2 shows representative examples.

We select these datasets for two reasons. First, they provide human-to-human reference data over the same visual objects, allowing us to simulate human Directors and compare model behavior against human-grounded baselines. Second, they cover several dimensions central to referential communication: real versus abstract objects, spoken versus text-based references, free matching versus sequential target selection, and objects with versus without conventional names. The less nameable contexts are particularly challenging: baskets require fine-grained attribute descriptions, while tangrams often elicit highly variable descriptions across speakers. We also include two basket image sets with different levels of initial-round human performance: 80% in Zeng et al. (2026) and 100% in Wang et al. (2025a). Together, these image sets provide a diverse, human-grounded testbed for evaluating LVLMs under varying degrees of visual ambiguity, nameability, and referential difficulty.

All four datasets contain repeated-round human reference data, where a round denotes one pass through the same image set by the same dyad. By default, we use only the initial round, where referring expressions are more self-contained and less dependent on dyad-specific common ground.

### 4.2 Referring Expression Processing

We reuse the referring expressions from the four datasets in Table 1, which contain human-produced descriptions for the same visual objects across multiple discourse participants. This supports a realistic and diverse evaluation setting, given the open-ended nature of referring expression generation.

[2]We also compute Expected Calibration Error (ECE), which is strongly correlated with accuracy in our setting (Spearman's $\rho = -0.96$), making it less informative as a separate measure of calibration. See Appendix B for details.

**Filtering** To ensure that the retained descriptions are accurate and sufficiently informative, we keep only initial-round referring expressions from dialogues in which human participants achieved at least 80% matching accuracy for Zeng et al. (2026) and 90% for Hawkins et al. (2020), balancing description quality against sample size. For the two image sets from Wang et al. (2025a), initial-round accuracy is 100%, so we retain all dialogues. Although item-level success could provide a more fine-grained filter, we use dialogue-level accuracy because the referring expressions are drawn from complete interaction transcripts and may therefore depend on the broader dialogue context.

We further control description quality by selecting only participants whose matching performance does not decrease across rounds within each dataset. Since repeated reference games typically become easier as participants establish common ground, this filter removes dialogues with unstable engagement or inconsistent task behavior. It also ensures that the selected dialogues are suitable for our multi-round follow-up studies, where we examine whether LVLMs can handle realistic long-horizon interactive visual grounding. This filtering yields 48 high-quality human dialogues.

**Full Target Item Description Extraction** Following Zeng et al. (2026), we use GPT-5.4 (OpenAI, 2025) to automatically extract object-level referring expressions for each image set in Table 1, except for Tangrams, which already includes manually curated object-level descriptions. This process yields 536 diverse, high-quality referring expressions across the four image sets, which we use in STATIC and FULL.

**Partial Target Item Description Extraction** To construct underspecified referring expressions, we prompt GPT-5.4 with the full descriptions for all target items in the same image set and ask it to extract the least informative attribute from each description. We use these extracted attributes as the initial descriptions in PARTIAL. This LLM-as-a-judge procedure is designed not to find optimal shortened descriptions, but to minimize initial informativeness so that the Matcher must decide whether clarification is needed. We verify that the extracted attributes are lexically supported by the source full expressions (see Appendix C). Our experiments further validate this design: all models ask questions under PARTIAL, indicating that the extracted attributes are underspecified enough to elicit clarification (e.g., Figure 3 in Section 5).

### 4.3 Director Simulation

Because our focus is the Matcher, we use GPT-5.4 as a fixed simulated Director. This choice is motivated by prior evidence that GPT-5.2 reaches about 85% initial-round accuracy with human Matchers (Zeng et al., 2026). Fixing the Director makes information access more comparable across Matchers and protocols; Section 6 tests sensitivity to an alternative Director, showing comparable results.

The Director is always conditioned on the same human-produced full target description, defined in Section 4.2. In STATIC, we programmatically present this preconfigured full description to the Matcher one target at a time. In FULL and PARTIAL, the Matcher first receives a preconfigured full or underspecified description, respectively, and the GPT-5.4 Director is called only when the Matcher asks a clarification question. In NONE, no initial description is provided; instead, the Director answers the Matcher's questions while being prompted with the corresponding full target description. When a question asks about information not explicitly stated in the description, the prompt instructs the Director to answer based only on the provided description and visual input. See Appendix F for the full Director prompts.

### 4.4 Matcher Simulation

**Models** We evaluate LVLMs from different model families and scales to simulate matchers: (1) five proprietary LVLMs, namely GPT-5.4, GPT-5.4-Mini, and GPT-5.5 (OpenAI, 2026a,b), and Gemini-3.1-Pro and Gemini-3.1-Flash-Lite (Google DeepMind, 2026); and (2) three open-weight LVLMs, namely Gemma-4-31B-IT (Google, 2026) and Qwen3-VL-8B/32B-Instruct (Bai et al., 2025).

**Prompting** We use protocol-specific system prompts while keeping the task overview fixed for both the Director and Matcher across conditions. In interactive settings, the Matcher returns a JSON object containing a natural-language utterance and any selections it makes, with a 100-point confidence score for each selection. In the static setting, the Matcher returns a JSON selection with a confidence score. See Appendix F for full prompts.

**Simulation Scale** Across the four datasets, we obtain 536 item–description pairs. Under the four interaction protocols in the main experiments, each LVLM matcher makes 2,144 target selections based

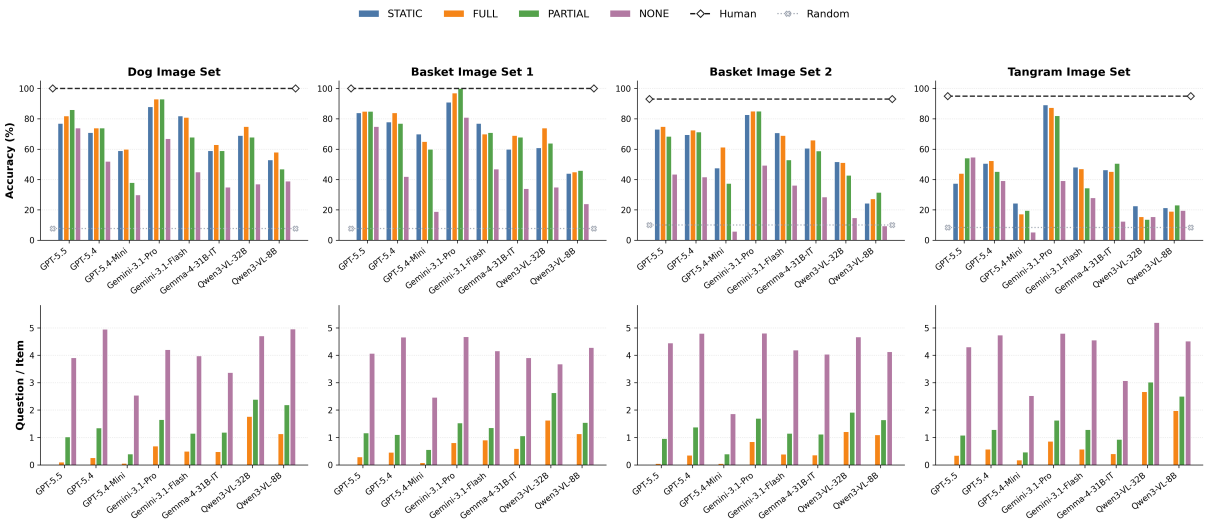


Figure 3: Overall results across the four image sets. Top: average matcher accuracy. Bottom: average questions per target item. Colors denote the four interaction protocols; horizontal lines mark human and random baselines. The human line is a *task-level* reference, but not a protocol-matched baseline, since the interaction conditions differ.

on information from the simulated Director. Across eight LVLM matchers, this yields 17,152 selections. Including the follow-up studies in Section 6, which add another 21,696 selections, our experiments produce 38,848 total target selections for analysis.

### 4.5 Multi-Turn Dialogue Simulation

**Visual Input Format** For each image set, the Director is shown only the target items, while the Matcher is shown the full candidate set, which may include distractors. We concatenate the individual item images into a single indexed grid and provide the grid as one image input. Director indices follow the item-description order in the original dataset, while Matcher indices are randomized to prevent trivial index matching or data contamination.

**Turn Taking** The Director communicates first in STATIC, FULL, and PARTIAL, whereas the Matcher initiates NONE with a yes/no question. In all three interactive protocols, namely, FULL, PARTIAL, and NONE, Matchers may ask questions and revise earlier selections when later information changes their decision. Each turn contains one natural-language utterance capped at 50 words, and interactions are limited to a 10-turn budget. If this budget is exhausted, the Matcher is prompted to make a forced target selection to ensure termination. The Director advances to the next target only after the Matcher selects. The final prediction sequence is computed from the Matcher's finalized selections.

| Model | Dogs | Baskets 1 | Baskets 2 | Tangrams | All |
|---|---|---|---|---|---|
| GPT-5.5 | $-14.0^{**}$ | $-15.0^{**}$ | $-18.5^{***}$ | $-40.5^{***}$ | $-25.3^{***}$ |
| GPT-5.4 | $-26.0^{***}$ | $-16.0^{***}$ | $-20.8^{***}$ | $-42.9^{***}$ | $-27.3^{***}$ |
| GPT-5.4-Mini | $-40.0^{***}$ | $-30.0^{***}$ | $-32.1^{***}$ | $-70.8^{***}$ | $-47.7^{***}$ |
| Gemini-3.1-Pro | $\mathbf{-7.0}^{*}$ | **0.0** | $\mathbf{-8.3}^{*}$ | $\mathbf{-6.0}$ | $\mathbf{-6.8}^{***}$ |
| Gemini-3.1-Flash | $-18.0^{***}$ | $-23.0^{**}$ | $-22.6^{***}$ | $-47.0^{***}$ | $-28.9^{***}$ |
| Gemma-4-31B-IT | $-37.0^{***}$ | $-31.0^{***}$ | $-27.4^{***}$ | $-44.6^{***}$ | $-36.7^{***}$ |
| Qwen3-VL-32B | $-25.0^{***}$ | $-26.0^{**}$ | $-41.7^{***}$ | $-72.6^{***}$ | $-46.2^{***}$ |
| Qwen3-VL-8B | $-42.0^{***}$ | $-54.0^{***}$ | $-61.9^{***}$ | $-72.0^{***}$ | $-61.4^{***}$ |

Table 2: Best paired mean LVLM–human matcher accuracy gap (%) across datasets, selected from the best-performing interaction protocol for each model–dataset pair. The All column is computed over all examples from the four datasets. Darker red denotes larger deficits relative to human matchers, and bold marks the smallest gap within each dataset. Significance is assessed with paired $t$-tests: $^{*}p \leq .05$, $^{**}p \leq .01$, and $^{***}p \leq .001$. See Appendix D.1 for protocol-specific results.

## 5 Results

### 5.1 Interactive Visual Grounding Poses a Challenging Task

**LVLM matchers significantly lag behind human matchers in interactive visual grounding.** The upper panels of Figure 3 show that nearly all tested LVLM matchers underperform the corresponding task-level human baselines across the four image sets, with the largest gaps on the Tangram Image Set. Table 2 further shows that these gaps are often substantial and statistically significant under paired $t$-tests, especially when results are pooled across all datasets. The only case of human parity is Gemini-3.1-Pro on Basket Image Set 2 under PARTIAL, but no model outperforms the human

| Model | STATIC | FULL | PARTIAL | NONE |
|---|---|---|---|---|
| GPT-5.5 | 65.8 | $\underline{69.5}_{+3.6}$ | $\underline{71.4}_{+5.5}$ | $\mathbf{59.7}_{-6.1}$ |
| GPT-5.4 | 66.1 | $69.4_{+3.4*}$ | $65.5_{-0.4}$ | $43.2_{-23.6*}$ |
| GPT-5.4-Mini | 47.9 | $49.0_{+0.6}$ | $37.1_{-11.5*}$ | $13.5_{-35.2*}$ |
| Gemini-3.1-Pro | **87.5** | $\mathbf{89.9}_{+2.9}$ | $\mathbf{89.0}_{+2.3}$ | $\underline{56.7}_{-28.6*}$ |
| Gemini-3.1-Flash | $\underline{67.8}$ | $65.3_{-2.7}$ | $54.5_{-12.9*}$ | $37.9_{-30.4**}$ |
| Gemma-4-31B-IT | 56.0 | $60.0_{+4.3}$ | $58.4_{+2.6}$ | $26.4_{-29.0**}$ |
| Qwen3-VL-32B | 48.8 | $50.5_{+2.8}$ | $44.0_{-4.0}$ | $23.9_{-25.5*}$ |
| Qwen3-VL-8B | 33.6 | $35.0_{+1.6}$ | $35.3_{+1.2}$ | $21.6_{-12.7*}$ |

Table 3: Micro-averaged accuracy (%) by model and interaction protocol. Subscripts report paired mean differences relative to STATIC; green indicates gains and red indicates losses. The best-performing model within each protocol is **bolded**, and the second-best model is underlined. Significance is assessed with paired $t$-tests: $^{*}p \leq .05$, $^{**}p \leq .01$, and $^{***}p \leq .001$.

baselines. Moreover, the consistent gaps under STATIC and FULL suggest that current LVLMs still struggle to comprehend descriptions drawn from high-accuracy human dialogues (Tang et al., 2024; Wang et al., 2025a; Zeng et al., 2026).

**The task is challenging because it requires both visual matching and effective information seeking.** GPT-5.4-Mini consistently asks the fewest questions. It outperforms Qwen3-VL-8B under STATIC and FULL, but its advantage nearly disappears under PARTIAL and reverses under NONE (Table 3), where question asking matters more. This suggests that limited question asking can hurt when the task requires acquiring missing information. However, asking more questions alone is not sufficient. Gemini-3.1-Pro asks more questions than GPT-5.5, yet GPT-5.5 achieves higher overall accuracy under NONE. These contrasts suggest that success requires not only visual matching or more questions, but useful questions and effective integration of answers across turns.

### 5.2 Stronger Models Perform Better, but Question-Only Grounding Remains Hard

**Model performance exhibits a clear hierarchy.** Figure 3 and Table 3 show that larger models generally outperform their smaller counterparts from the same model family, suggesting benefits from scaling, and proprietary models tend to outperform open-weight models. Gemini-3.1-Pro is strongest overall, achieving the best performance under STATIC, FULL, and PARTIAL, with especially large gains on the Tangram Image Set. GPT-5.5 is also competitive, particularly under NONE, suggesting stronger question-driven information seeking and integration capabilities.

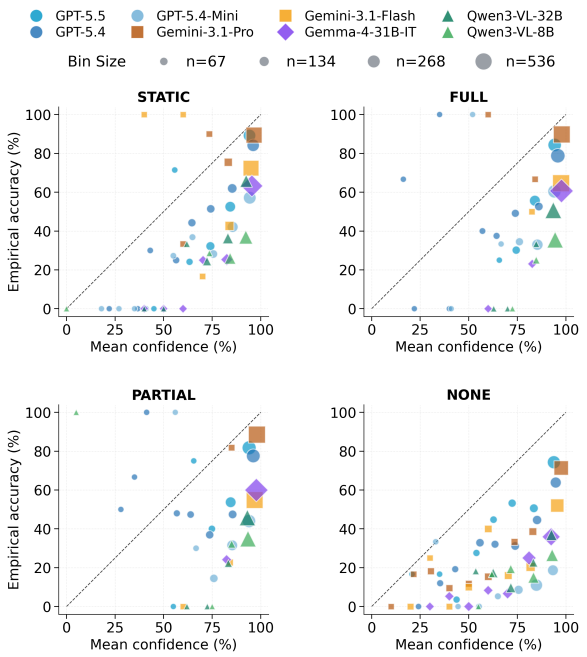


Figure 4: Protocol-level calibration. Points show model-specific confidence bins from 10 evenly spaced bins between 0 and 100; point size is proportional to bin count. The dashed diagonal denotes perfect calibration.

**LVLMs can benefit from interaction, but question-driven grounding remains challenging.** Table 3 shows that FULL improves over STATIC for seven of the eight models, suggesting that optional follow-up questions can help refine or repair initial descriptions. Under PARTIAL, four models improve over STATIC despite receiving an underspecified description, consistent with successful information recovery, while four decline. Compared with FULL, however, six of the eight models perform worse under PARTIAL, indicating that interaction often does not fully compensate for missing initial information. The clearest limitation appears under NONE, where all models perform below STATIC despite asking the most questions. This suggests that current LVLMs struggle to construct and ground target representations through proactive question-driven dialogue.

### 5.3 LVLMs Are Not Yet Calibrated for Interactive Visual Grounding

**Current LVLMs are sensitive to information availability, but remain overconfident.** All models ask more questions when less initial information is available: NONE > PARTIAL > FULL > STATIC (= 0 by design; lower panels of Figure 3). Figure 4 shows a similar shift in confidence: models tend to report lower confidence in

more interaction-heavy protocols. However, most points still fall below the diagonal, indicating that reported confidence often exceeds empirical accuracy. This overconfidence is especially pronounced under PARTIAL and NONE, suggesting that LVLMs are less reliable at judging information sufficiency when the initial referring expression is underspecified or must be constructed through interaction. Overall, this suggests that LVLMs adjust their behavior to the amount of available information, but their confidence does not yet reliably reflect whether they have enough evidence for accurate grounding.

**Self-repairs are rare and inconsistently reliable.** Revising a selection could indicate that a model recognizes and acts on uncertainty, but repairs occur only under FULL and PARTIAL and account for under 2% of matches. Among the 49 observed repairs, 26 (53.1%) correct a wrong selection, 5 (10.2%) change a correct selection to a wrong one, and 18 (36.7%) replace one wrong answer with another. Outcomes are strongly model-dependent: Gemini-3.1-Pro corrects 13 of its 17 revisions (76.5%), whereas GPT-5.4 corrects none of six and nevertheless raises its confidence by 11.2 points on average. Thus, models can occasionally exploit later evidence to repair an error, but revision remains too sparse and its confidence changes too inconsistent to serve as reliable error awareness. Full model- and protocol-level results appear in Appendix D.2.

### 5.4 Summary

Overall, interactive visual grounding remains challenging for current LVLMs. Models ask more questions when information is limited and can sometimes benefit from interaction, especially when follow-up questions refine or repair an initial description. However, they still lag behind human matchers, struggle with proactive question-driven dialogue, and remain poorly calibrated for the task.

## 6 Follow-Up Studies

To test whether our conclusions depend on description source, test-time reasoning, repeated-reference structure, Director choice, or the original visual contexts, we conduct five follow-up studies. We (i) replace human descriptions with GPT-5.4-generated descriptions across all Matchers and protocols; (ii) vary reasoning effort for GPT-5.4 and GPT-5.4-Mini on tangrams; (iii) evaluate three LVLMs over four rounds of repeated reference, using the corresponding human-produced expressions from the original human–human dialogues; (iv) replace GPT-5.4 with Gemini-3.1-Flash as the Director on Basket Set 2 and Tangrams; and (v) evaluate five LVLM Matchers (i.e., GPT-5.4, GPT-5.4-Mini, Gemini-3.1-Flash, Gemma-4-31B-IT, and Qwen3-VL-8B) on six public face image sets with GPT-5.4-generated descriptions (Holland et al., 2018), each containing 12 images of one individual with varied facial expressions.

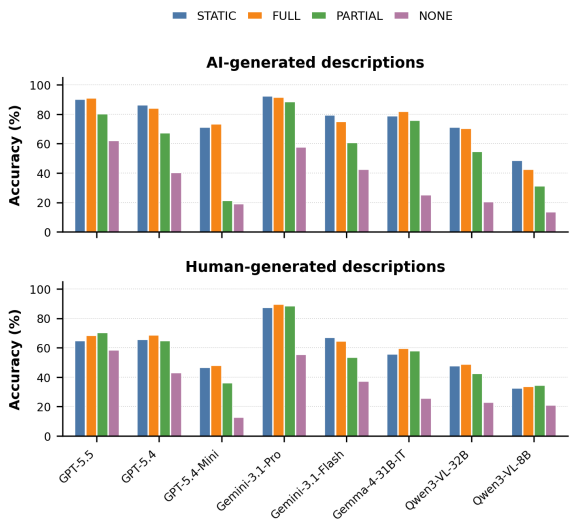


Figure 5: Accuracy (%) on AI-generated and human-generated descriptions across models and interaction protocols, averaged over all four datasets.

| Reasoning | GPT-5.4 | | GPT-5.4-Mini | |
|---|---|---|---|---|
| | STATIC | FULL | STATIC | FULL |
| None | 50.6 | 52.4 | 24.4 | 17.3 |
| Medium | $52.4_{+1.8}$ | $60.7_{+8.3}$ | $26.2_{+1.8}$ | $30.4_{+13.1}$ |
| High | $48.8_{-1.8}$ | $61.3_{+8.9}$ | $15.5_{-8.9}$ | $31.0_{+13.7}$ |

Table 4: Accuracy (%) under different reasoning-effort settings on tangrams. Subscripts report differences relative to default no reasoning.

**LVLMs perform better with AI-generated descriptions, but result patterns remain comparable.** Figure 5 shows that LVLMs generally achieve higher accuracy with AI-generated descriptions than with human-generated descriptions, especially under STATIC and FULL. This suggests that AI-generated descriptions are easier for LVLMs to interpret, while human referring expressions pose a distinct challenge. Importantly, however, the overall protocol-level patterns remain similar: STATIC and FULL are strongest, PARTIAL is more variable, and NONE remains most difficult.

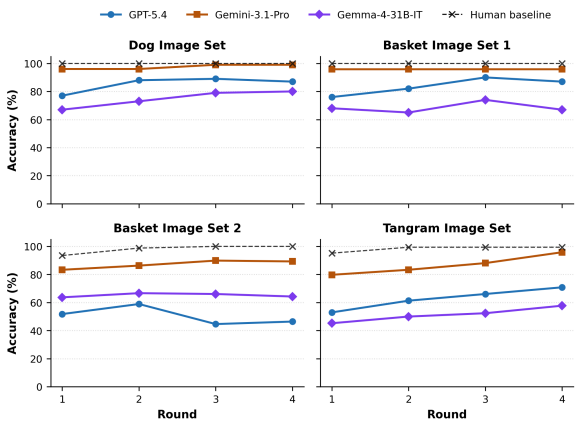


Figure 6: Accuracy (%) across rounds on all four image sets under FULL. See Appendix E.3 for STATIC results.

| **A. Alternative Director robustness** | | |
|---|---|---|
| Director / comparison | FULL | PARTIAL |
| GPT-5.4 | 51.5 | 48.3 |
| Gemini-3.1-Flash | 50.4 | 44.8 |
| Δ (Gemini − GPT) | −1.1 | −3.5 |
| Rank correlation $\rho$ | .929 | .976 |

| **B. Face-domain robustness** | | | | |
|---|---|---|---|---|
| Metric | STATIC | FULL | PARTIAL | NONE |
| Accuracy (%) | 48.7 | 47.8 | 41.9 | 30.7 |
| Δ vs. STATIC | – | −0.9 | −6.8 | −18.0 |

Table 5: Aggregate alternative-Director (A) and face-domain (B) results. Panel A averages across Matchers, datasets, and description sources; Panel B across Matchers and six face sets.

**Reasoning helps in interactive settings, but scaling plateaus quickly.** Reasoning yields little benefit in STATIC, but substantially improves accuracy in FULL (Table 4). This suggests that stronger reasoning helps models gather and synthesize visual evidence through conversation. However, gains largely plateau from medium to high reasoning, indicating that additional test-time compute alone is not sufficient. Reasoning also modestly improves calibration, with similar plateaus from medium to high effort, as discussed in Appendix E.2.

**LVLMs can benefit from repeated interaction, but the gap to human matchers remains.** Figure 6 shows that models can improve across rounds under FULL. Results for STATIC show similar patterns and are reported in Appendix E.3. These findings suggest that LVLMs can benefit from increasingly efficient referring expressions produced through entrainment between humans. However, improvements are inconsistent across datasets, and human matchers continue to outperform LVLMs. While Gemini-3.1-Pro approaches human baselines in some settings, long-horizon interactive visual grounding remains a broad challenge for LVLMs.

**Matcher rankings are stable across Director models.** Table 5 A shows that replacing the GPT-5.4 Director with Gemini-3.1-Flash reduces mean Matcher accuracy by only 1.1 points in FULL (51.5% to 50.4%), but by 3.5 points in PARTIAL (48.3% to 44.8%). The larger PARTIAL decrease is consistent with the Director carrying more responsibility when the initial description is underspecified. Despite these absolute changes, Matcher rankings remain highly correlated across Directors (Spearman's $\rho = .929$ for FULL and .976 for PARTIAL). Thus, interaction quality affects accuracy, especially under underspecification, but the central Matcher hierarchy is not an artifact of a single simulated Director (Appendix E.4).

**The protocol-level pattern extends to fine-grained face matching.** Across six face sets with AI-generated descriptions, Table 5B shows that FULL remains within 0.9 points of STATIC (47.8% vs. 48.7%), whereas PARTIAL and NONE fall 6.8 and 18.0 points below STATIC, respectively. Thus, optional clarification largely preserves performance when an informative description is available, but interaction does not fully compensate as initial information is removed. This graded pattern matches Figure 5 and extends the main protocol-level observation to a distinct fine-grained domain, confirming the core finding in Section 5 (Appendix E.5).

## 7 Conclusion

We introduce a controlled framework for evaluating interactive visual grounding beyond one-shot matching. Across four human-grounded visual contexts, current LVLMs perform significantly below task-level human references. Interaction can help when follow-up questions refine or repair an initial description, but models consistently struggle with proactive question-driven grounding. They are also poorly calibrated, often reporting confidence that exceeds accuracy. These findings are further validated across follow-up studies using AI-generated descriptions, test-time scaling, exposure to repeated interaction, an alternative simulated Director, and a fine-grained face context. Overall, interactive visual grounding remains an important challenge for LVLMs, with potential implications for future embodied human–AI interaction.

## Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) 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 NSF.

We are grateful for support from the Institute for Advanced Computational Science (IACS) at Stony Brook University. We also thank the three anonymous reviewers for their valuable feedback.

## AI Usage Disclosure

We used generative AI tools, particularly Codex, to assist with developing experimental and analysis code and with language polishing. All methodological and interpretive decisions were made by the authors, who reviewed and verified all AI-assisted outputs.

## Limitations

**Simulated Directors and Director Dependence.** We use simulated rather than human Directors for scale and experimental control. Repeating FULL and PARTIAL with Gemini-3.1-Flash produces highly stable Matcher rankings, reducing concern that the main conclusions are specific to GPT-5.4. That said, this robustness check does not establish equivalence to human Directors, and individual model–protocol outcomes can still vary with the Director.

**Attribution.** The four protocols vary how much information must be acquired interactively. They are designed to create a controlled way to separate visual matching from information seeking and question-driven grounding, but do not uniquely isolate question selection, answer integration, dialogue memory, and visual matching etc. Our claim is also limited: these protocols increase the need for information seeking and synthesis beyond static visual matching, rather than attributing each failure to one precise subcomponent, which we leave to future studies.

**Visual Context Coverage.** The main testbed contains four object-level contexts, and the face follow-up adds another fine-grained within-category setting. Neither covers cluttered scenes, spatial or relational grounding, or open-world visual search. Our conclusions therefore concern proactive interactive grounding in controlled, fine-grained matching contexts rather than visual grounding in all settings.

**Constrained Interaction under NONE.** The Director provides yes/no answers under NONE to create a controlled information-seeking setting, following de Vries et al. (2017). Open-ended Director answers may support different questioning strategies and should be studied separately.

## Ethical Considerations

**Data Use and Privacy.** This work evaluates LVLMs on interactive visual grounding using existing research datasets and public or previously released visual materials. We do not collect new human-subject data. The face-domain robustness study uses public research images only for within-set visual matching; we neither infer identity nor predict sensitive attributes. The main evaluated images involve object-level visual contexts such as dogs, baskets, and abstract tangrams, which reduces risks related to privacy or demographic stereotyping.

## References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine 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. In *Advances in Neural Information Processing Systems*.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand. Association for Computational Linguistics.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

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):1482–1493.

Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S. H. Gary Chan, and Hongyang Zhang. 2024. Revisiting referring expression comprehension evaluation in the era of large multimodal models. *Preprint*, arXiv:2406.16866.

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

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qihua Dong, Kuo Yang, Lin Ju, Handong Zhao, Yitian Zhang, Yizhou Wang, Huimin Zeng, Jianglin Lu, and Yun Fu. 2026. Ref-adv: Exploring MLLM visual reasoning in referring expression tasks. In *The Fourteenth International Conference on Learning Representations*.

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.

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, and Angeliki Lazaridou and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Google. 2026. Gemma 4 model overview. `https://ai.google.dev/gemma/docs/core/model_card_4`. Accessed: 2026-05-19.

Google DeepMind. 2026. Gemini 3.1 Pro model card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`. Accessed: 2026-05-19.

Chi Han, Xin Liu, Haodong Wang, Shiyang Li, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, Qingyu Yin, Liang Qiu, Changlong Yu, Yifan Gao, Zheng Li, Bing Yin, Jingbo Shang, and Heng Ji. 2025. Can language models follow multiple turns of entangled instructions? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25445–25460, Suzhou, China. Association for Computational Linguistics.

Robert D. Hawkins, Michael C. Frank, and Noah D. Goodman. 2020. Characterizing the dynamics of learning in repeated reference games. *Cognitive Science*, 44(6).

Catherine A. C. Holland, Natalie C. Ebner, Tian Lin, and Gregory R. Samanez-Larkin. 2018. Emotion identification across adulthood using the dynamic faces database of emotional expressions in younger, middle aged, and older adults. *Cognition and Emotion*, 33(2):245–257.

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

Saki Imai, Mert Inan, Anthony B. Sicilia, and Malihe Alikhani. 2025. Measuring how (not just whether) VLMs build common ground. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 441–451, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Cameron R. Jones, Agnese Lombardi, Kyle Mahowald, and Benjamin K. Bergen. 2026. Llms and people both learn to form conventions – just not with each other. *Preprint*, arXiv:2602.08208.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara 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, Doha, Qatar. Association for Computational Linguistics.

Robert M. Krauss and Sam Glucksberg. 1969. The development of communication: Competence as a function of age. *Child Development*, 40(1):255.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. MT-eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20153–20177, Miami, Florida, USA. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*.

Young-Jun Lee, Byung-Kwan Lee, Jianshu Zhang, Yechan Hwang, Byungsoo Ko, Han-Gyu Kim, Dongyu Yao, Xuankun Rong, Eojin Joo, Seung-Ho Han, Bowon Ko, and Ho-Jin Choi. 2025. Multiverse: A multi-turn conversation benchmark for evaluating large vision and language models. *Preprint*, arXiv:2510.16641.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. *Microsoft COCO: Common Objects in Context*, page 740–755. Springer International Publishing.

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.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. MMDU: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for LVLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20.

OpenAI. 2025. Gpt-5 is here. https://openai.com/gpt-5/. Accessed: 2026-05-20.

OpenAI. 2026a. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/. Accessed: 2026-05-19.

OpenAI. 2026b. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/. Accessed: 2026-05-19.

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, and Jeff Belgum and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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

Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2025. Navigating rifts in human-LLM grounding: Study and benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20832–20847, Vienna, Austria. Association for Computational Linguistics.

Mohit Shridhar, Dixant Mittal, and David Hsu. 2020. Ingress: Interactive visual grounding of referring expressions. *Int. J. Rob. Res.*, 39(2–3):217–232.

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.

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.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Keqin Chen, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Cogvlm: Visual expert for pretrained language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 121475–121499. Curran Associates, Inc.

Zhengxiang Wang, Weiling Li, Panagiotis Kaliosis, Owen Rambow, and Susan Brennan. 2025a. LVLMs are bad at overhearing human referential communication. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16758–16782, Suzhou, China. Association for Computational Linguistics.

Zhengxiang Wang, Veronika Makarova, Zhi Li, Jordan Kodner, and Owen Rambow. 2025b. LLMs can perform multi-dimensional analytic writing assessments: A case study of L2 graduate-level academic English writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8637–8663, Vienna, Austria. Association for Computational Linguistics.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. 2026. Toward visual grounding: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2749–2771.

Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, Yutong Xie, Imran Razzak, Zongyuan Ge, Jionglong Su, Junjun He,

and Yu Qiao. 2025. MMRC: A large-scale benchmark for understanding multimodal large language model in real-world conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22477–22503, Vienna, Austria. Association for Computational Linguistics.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. *Preprint*, arXiv:1608.00272.

Peter Zeng, Weiling Li, Amie Paige, Zhengxiang Wang, Panagiotis Kaliosis, Dimitris Samaras, Gregory Zelinsky, Susan Brennan, and Owen Rambow. 2026. Lvlms and humans ground differently in referential communication. *Preprint*, arXiv:2601.19792.

Hanbo Zhang, Jie Xu, Yuchen Mo, and Tao Kong. 2023. Invig: Benchmarking interactive visual grounding with 500k human-robot interactions. *Preprint*, arXiv:2310.12147.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

## Appendix Table of Contents

**A LVLM Details** **13**

**B Expected Calibration Error** **13**

**C Extracted Partial Description Validation** **14**

**D Results** **14**
- D.1 LVLM and Human Matcher Performance Gap across Interaction Protocols . . . 14
- D.2 Self-Repair Analysis . . . 15

**E Follow-Up Studies** **15**
- E.1 AI-Generated Descriptions . . . 15
- E.2 Reasoning Effort . . . 17
- E.3 Multi-Round Repeated References 18
- E.4 Alternative Director Robustness . 18
- E.5 Face-Domain Robustness . . . 18

**F Prompts for Visual Grounding Experiments** **21**
- F.1 Task Overview . . . 21
- F.2 Matcher Output Format Instructions 21
- F.3 STATIC Protocol . . . 22
- F.4 FULL Protocol . . . 22
- F.5 PARTIAL Protocol . . . 23
- F.6 NONE Protocol . . . 24

**G Prompts for Referring Expression Processing** **25**
- G.1 Referring Expression Extraction . 25
- G.2 Partial Description Extraction . . . 26
- G.3 Object Description Generation . . 27

**H Dataset Details** **27**

## A LVLM Details

Table 6 lists the exact LVLM variants used in our experiments. We evaluate proprietary GPT and Gemini models through their respective APIs, using the latest available variants at the time of experimentation. For open-weight models, we use the Hugging Face checkpoints listed in the table and run local inference with vLLM (Kwon et al., 2023). Unless otherwise specified, we use the default decoding configuration across models.

## B Expected Calibration Error

We compute Expected Calibration Error (ECE) to summarize the alignment between the Matcher's stated confidence and empirical accuracy. For each final selection, the Matcher is prompted to report a confidence score $c_j \in [0, 100]$, and we define correctness as whether the selected item matches the target item at the corresponding position. We partition predictions into 10 evenly spaced confidence bins over $[0, 100]$: $[0, 10), [10, 20), \ldots, [90, 100]$.

For each bin $B_\ell$, we compute the average empirical accuracy and the average stated confidence of all predictions in that bin. ECE is then computed as the weighted average absolute gap between accuracy and confidence across bins:

$$\mathrm{ECE} = \sum_{\ell=1}^{10} \frac{|B_\ell|}{N} \left|\mathrm{acc}(B_\ell) - \mathrm{conf}(B_\ell)\right|,$$

where $N$ is the total number of predictions, $|B_\ell|$ is the number of predictions in bin $B_\ell$, $\mathrm{acc}(B_\ell)$ is the percentage of correct predictions in the bin, and $\mathrm{conf}(B_\ell)$ is the average stated confidence in the bin. Lower ECE indicates better calibration.

In the main text, we report calibration primarily through calibration plots rather than ECE. This is because ECE is strongly negatively correlated with accuracy in our setting (Spearman's $\rho = -0.96$), making it less informative as a separate measure of calibration. Intuitively, models with lower accuracy tend to have larger confidence–accuracy gaps

| Model | Provider / Family | Variant used (snapshot / HF repo) | Inference backend |
|---|---|---|---|
| GPT-5.5 (OpenAI, 2026b) | OpenAI | `gpt-5.5-2026-04-23` | OpenAI API |
| GPT-5.4 (OpenAI, 2026a) | OpenAI | `gpt-5.4-2026-03-05` | OpenAI API |
| GPT-5.4-Mini (OpenAI, 2026a) | OpenAI | `gpt-5.4-mini-2026-03-17` | OpenAI API |
| Gemini-3.1-Pro (Google DeepMind, 2026) | Google Gemini | `gemini-3.1-pro-preview` | Gemini API |
| Gemini-3.1-Flash (Google DeepMind, 2026) | Google Gemini | `gemini-3.1-flash-lite-preview` | Gemini API |
| Gemma-4-31B-IT (Google, 2026) | Gemma | `google/gemma-4-31b-it` | vLLM |
| Qwen3-VL-32B (Bai et al., 2025) | Qwen | `Qwen/Qwen3-VL-32B-Instruct` | vLLM |
| Qwen3-VL-8B (Bai et al., 2025) | Qwen | `Qwen/Qwen3-VL-8B-Instruct` | vLLM |

Table 6: LVLM variants used in our experiments. API models are identified by their provider-facing model names, and open-weight models are identified by their Hugging Face repository IDs. Open-weight models are run locally with vLLM.

even when their confidence patterns are qualitatively similar. Calibration plots therefore provide a more interpretable view of whether models are overconfident or underconfident across confidence levels.

## C Extracted Partial Description Validation

**Metric definitions.** Let $d$ denote the full description and let $A = \{a_1, \ldots, a_n\}$ denote the generated attribute list for an item. Text is normalized by lowercasing and tokenized with the regular expression `[A-Za-z0-9]+`, so punctuation is ignored and only alphanumeric tokens are retained.

For each attribute $a_i$, we compute an attribute support score $s(a_i, d)$ as the maximum of exact substring match and token overlap:

$$s(a_i, d) = \max\left(\mathrm{substr}(a_i, d), \mathrm{overlap}(a_i, d)\right).$$

An attribute is marked as supported if $s(a_i, d) \geq 0.45$.

The attribute support rate for an item is the fraction of generated attributes whose support score exceeds this threshold:

$$\mathrm{support_rate}(A, d) = \frac{\sum_{i=1}^{n} \mathbb{1}[s(a_i, d) \geq 0.45]}{n}.$$

Description coverage measures how much of the original full description is recovered by the generated attributes:

$$\mathrm{coverage}(A, d) = \frac{\sum_{w \in V} \min\left(c_A(w), c_d(w)\right)}{|T(d)|}.$$

For the selected underspecified expression, we compute the least-attribute support score as $s(a_{\text{least}}, d)$, which verifies that the initial partial expression is grounded in the original description despite being intentionally underspecified.

**Validation results.** The extracted attributes are strongly supported by the original full descriptions. Across all items, the mean attribute support rate is 0.998, indicating that nearly all generated attributes can be matched back to the source description. The mean attribute support score is 0.980, and the mean description coverage is 0.895, suggesting that the decomposition preserves most of the descriptive content while keeping the attributes grounded in the original expression. The selected least-informative attributes are also well supported, with a mean least-attribute support score of 0.981.

These results suggest that our procedure produces source-grounded partial descriptions: the initial expressions used in PARTIAL are intentionally underspecified, but not hallucinated or detached from the original human descriptions.

## D Results

### D.1 LVLM and Human Matcher Performance Gap across Interaction Protocols

Tables 7–10 report the full LVLM–human matcher accuracy gaps for each interaction protocol, complementing the best-over-protocol summary in Table 2. Across STATIC, FULL, and PARTIAL, nearly all LVLMs significantly lag behind human matchers, although the size of the gap varies by model, dataset, and protocol. The only non-significant cases occur for Gemini-3.1-Pro, including one case in which it reaches parity with human matchers.

Under NONE, all models significantly underperform human matchers. This comparison should be interpreted with caution, however, because the human reference data were not collected under an explicit NONE setting. Instead, the human interactions more closely resemble a mixture of the other protocols, depending on how much detail human Directors provided in their object descriptions and

| Model | Dogs | Baskets 1 | Baskets 2 | Tangrams | All |
|---|---|---|---|---|---|
| GPT-5.5 | $-23.0^{***}$ | $-16.0^{**}$ | $-20.2^{***}$ | $-57.7^{***}$ | $-30.9^{***}$ |
| GPT-5.4 | $-29.0^{***}$ | $-22.0^{***}$ | $-23.8^{***}$ | $-44.6^{***}$ | $-30.6^{***}$ |
| GPT-5.4-Mini | $-41.0^{***}$ | $-30.0^{***}$ | $-45.8^{***}$ | $-70.8^{***}$ | $-48.8^{***}$ |
| Gemini-3.1-Pro | $\mathbf{-12.0}^{*}$ | $\mathbf{-9.0}^{***}$ | $\mathbf{-10.7}^{*}$ | $\mathbf{-6.0}$ | $\mathbf{-9.2}^{***}$ |
| Gemini-3.1-Flash | $-18.0^{***}$ | $-23.0^{**}$ | $-22.6^{***}$ | $-47.0^{***}$ | $-28.9^{***}$ |
| Gemma-4-31B-IT | $-41.0^{***}$ | $-40.0^{***}$ | $-32.7^{***}$ | $-48.8^{***}$ | $-40.7^{***}$ |
| Qwen3-VL-32B | $-31.0^{***}$ | $-39.0^{***}$ | $-41.7^{***}$ | $-72.6^{***}$ | $-47.9^{***}$ |
| Qwen3-VL-8B | $-47.0^{***}$ | $-56.0^{***}$ | $-69.0^{***}$ | $-73.8^{***}$ | $-63.1^{***}$ |

Table 7: Mean LVLM–human matcher accuracy gap (%) under the STATIC protocol. Darker red denotes larger deficits relative to human matchers, and bold marks the smallest gap within each dataset. Significance is assessed with paired $t$-tests: $^{*}p \leq .05$, $^{**}p \leq .01$, $^{***}p \leq .001$.

| Model | Dogs | Baskets 1 | Baskets 2 | Tangrams | All |
|---|---|---|---|---|---|
| GPT-5.5 | $-18.0^{***}$ | $-15.0^{**}$ | $-18.5^{***}$ | $-51.2^{***}$ | $-27.2^{***}$ |
| GPT-5.4 | $-26.0^{***}$ | $-16.0^{***}$ | $-20.8^{***}$ | $-42.9^{***}$ | $-27.3^{***}$ |
| GPT-5.4-Mini | $-40.0^{***}$ | $-35.0^{***}$ | $-32.1^{***}$ | $-78.0^{***}$ | $-47.7^{***}$ |
| Gemini-3.1-Pro | $\mathbf{-7.0}^{*}$ | $\mathbf{-3.0}$ | $\mathbf{-8.3}^{*}$ | $\mathbf{-7.7}^{*}$ | $\mathbf{-6.8}^{***}$ |
| Gemini-3.1-Flash | $-19.0^{***}$ | $-30.0^{***}$ | $-24.4^{***}$ | $-48.2^{***}$ | $-31.4^{***}$ |
| Gemma-4-31B-IT | $-37.0^{***}$ | $-31.0^{***}$ | $-27.4^{***}$ | $-50.0^{***}$ | $-36.7^{***}$ |
| Qwen3-VL-32B | $-25.0^{***}$ | $-26.0^{**}$ | $-42.3^{***}$ | $-79.8^{***}$ | $-46.2^{***}$ |
| Qwen3-VL-8B | $-42.0^{***}$ | $-55.0^{***}$ | $-66.1^{***}$ | $-76.2^{***}$ | $-61.7^{***}$ |

Table 8: Mean LVLM–human matcher accuracy gap (%) under the FULL protocol. Darker red denotes larger deficits relative to human matchers, and bold marks the smallest gap within each dataset. Significance is assessed with paired $t$-tests: $^{*}p \leq .05$, $^{**}p \leq .01$, $^{***}p \leq .001$.

how well Matchers understood those descriptions.

### D.2 Self-Repair Analysis

Tables 12 and 11 summarize self-repair behavior for models that revise their selections at least once. Self-repairs are relatively rare, accounting for no more than 2% of the 536 matches in any interaction protocol, and occur only under FULL and PARTIAL, with no observed cases under NONE. When models do revise, outcomes vary substantially: some repairs change initially wrong selections to correct ones, especially for Gemini-3.1-Pro, while others

| Model | Dogs | Baskets 1 | Baskets 2 | Tangrams | All |
|---|---|---|---|---|---|
| GPT-5.5 | $-14.0^{**}$ | $-15.0^{**}$ | $-25.0^{***}$ | $-41.1^{***}$ | $-25.3^{***}$ |
| GPT-5.4 | $-26.0^{***}$ | $-23.0^{***}$ | $-22.0^{***}$ | $-50.0^{***}$ | $-31.2^{***}$ |
| GPT-5.4-Mini | $-62.0^{***}$ | $-40.0^{***}$ | $-56.0^{***}$ | $-75.6^{***}$ | $-59.6^{***}$ |
| Gemini-3.1-Pro | $\mathbf{-7.0}$ | $\mathbf{0.0}$ | $\mathbf{-8.3}^{*}$ | $\mathbf{-13.1}^{**}$ | $\mathbf{-7.7}^{***}$ |
| Gemini-3.1-Flash | $-32.0^{***}$ | $-29.0^{***}$ | $-40.5^{***}$ | $-60.7^{***}$ | $-42.2^{***}$ |
| Gemma-4-31B-IT | $-41.0^{***}$ | $-32.0^{***}$ | $-34.5^{***}$ | $-44.6^{***}$ | $-38.3^{***}$ |
| Qwen3-VL-32B | $-32.0^{***}$ | $-36.0^{***}$ | $-50.6^{***}$ | $-81.5^{***}$ | $-52.7^{***}$ |
| Qwen3-VL-8B | $-53.0^{***}$ | $-54.0^{***}$ | $-61.9^{***}$ | $-72.0^{***}$ | $-61.4^{***}$ |

Table 9: Mean LVLM–human matcher accuracy gap (%) under the PARTIAL protocol. Darker red denotes larger deficits relative to human matchers, and bold marks the smallest gap within each dataset. Significance is assessed with paired $t$-tests: $^{*}p \leq .05$, $^{**}p \leq .01$, $^{***}p \leq .001$.

| Model | Dogs | Baskets 1 | Baskets 2 | Tangrams | All |
|---|---|---|---|---|---|
| GPT-5.5 | $\mathbf{-26.0}^{***}$ | $-25.0^{***}$ | $-50.0^{***}$ | $\mathbf{-40.5}^{***}$ | $\mathbf{-37.0}^{***}$ |
| GPT-5.4 | $-48.0^{***}$ | $-58.0^{***}$ | $-51.8^{***}$ | $-56.0^{***}$ | $-53.5^{***}$ |
| GPT-5.4-Mini | $-70.0^{***}$ | $-81.0^{***}$ | $-87.5^{***}$ | $-89.9^{***}$ | $-83.2^{***}$ |
| Gemini-3.1-Pro | $-33.0^{***}$ | $\mathbf{-19.0}^{**}$ | $\mathbf{-44.0}^{***}$ | $-56.0^{***}$ | $-40.0^{***}$ |
| Gemini-3.1-Flash | $-55.0^{***}$ | $-53.0^{***}$ | $-57.1^{***}$ | $-67.3^{***}$ | $-58.8^{***}$ |
| Gemma-4-31B-IT | $-65.0^{***}$ | $-66.0^{***}$ | $-64.9^{***}$ | $-82.7^{***}$ | $-70.3^{***}$ |
| Qwen3-VL-32B | $-63.0^{***}$ | $-65.0^{***}$ | $-78.6^{***}$ | $-79.8^{***}$ | $-72.8^{***}$ |
| Qwen3-VL-8B | $-61.0^{***}$ | $-76.0^{***}$ | $-83.9^{***}$ | $-75.6^{***}$ | $-75.1^{***}$ |

Table 10: Mean LVLM–human matcher accuracy gap (%) under the NONE protocol. Darker red denotes larger deficits relative to human matchers, and bold marks the smallest gap within each dataset. Significance is assessed with paired $t$-tests: $^{*}p \leq .05$, $^{**}p \leq .01$, $^{***}p \leq .001$.

| Model | FULL | PARTIAL | Total |
|---|---|---|---|
| GPT-5.5 | 7 | 10 | 17 |
| GPT-5.4 | 4 | 2 | 6 |
| GPT-5.4-Mini | 0 | 1 | 1 |
| Gemini-3.1-Pro | 9 | 8 | 17 |
| Gemini-3.1-Flash | 1 | 4 | 5 |
| Gemma-4-31B-IT | 2 | 1 | 3 |

Table 11: Distribution of self-repair cases by model and interaction protocol. Counts indicate the number of observed self-repairs under FULL and PARTIAL. We did not observe self-repairs under NONE.

introduce errors or replace one wrong answer with another, as in about half of GPT-5.5 cases. Confidence changes are also mixed, suggesting that revisions do not always correspond to increased certainty. Overall, self-repair is observable but limited in current LVLM matchers.

## E Follow-Up Studies

This section provides additional implementation details and results for the follow-up studies conducted in Section 6.

### E.1 AI-Generated Descriptions

**Experimental Procedure** For experiments on AI-generated descriptions, we replace human-generated descriptions with GPT-5.4-generated ones and rerun the main experiments across all models and interaction protocols. We repeat each experiment three times using the same description map (see below) but with shuffled description order. This controls for variation in item ordering while isolating the effect of description source, allowing us to test whether the main results depend on the use of human referring expressions.

| Model | Count | Wrong → Correct | Correct → Wrong | Wrong → Wrong | Δ Confidence |
|---|---|---|---|---|---|
| GPT-5.5 | 17 | 47.1 | 5.9 | 47.1 | −14.9 |
| GPT-5.4 | 6 | 0.0 | 50.0 | 50.0 | +11.2 |
| GPT-5.4-Mini | 1 | 100.0 | 0.0 | 0.0 | −15.0 |
| Gemini-3.1-Pro | 17 | 76.5 | 0.0 | 23.5 | −4.3 |
| Gemini-3.1-Flash | 5 | 60.0 | 20.0 | 20.0 | +1.0 |
| Gemma-4-31B-IT | 3 | 33.3 | 0.0 | 66.7 | −16.7 |

Table 12: Self-repair outcomes by model. Count denotes the number of self-repair cases. Transition columns report the percentage of repairs that change an initially wrong selection to correct, correct selection to wrong, or wrong selection to another wrong answer. Δ Confidence denotes the mean change in confidence after repair; positive values indicate increased confidence, and negative values indicate decreased confidence.

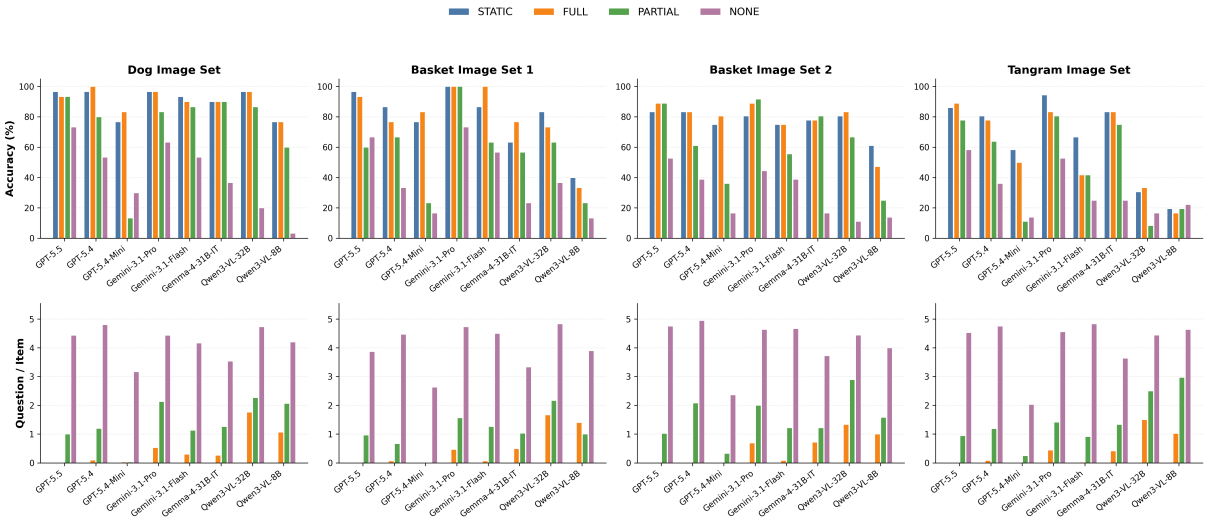


Figure 7: Overall results across the four image sets from the follow-up experiments with AI-generated descriptions. Top: average matcher accuracy. Bottom: average questions per target item. Colors denote the four interaction protocols; horizontal lines mark human and random baselines.

**Additional Result Discussion** Figure 7 reports the full results for each dataset. The overall patterns are comparable to those in the main experiments with human-generated descriptions (Figure 3; Section 5). LVLMs also remain overconfident and undercalibrated even with AI-generated descriptions, as shown in Figure 8.

For easier comparison, we aggregate results across the four datasets and visualize the differences in accuracy and question-asking behavior between AI- and human-generated descriptions in Figure 9. Figure 9 shows that LVLMs generally achieve higher accuracy with AI-generated descriptions than with human-generated descriptions, especially under STATIC and FULL. Models also tend to ask fewer questions under FULL, suggesting that AI-generated descriptions are easier for LVLMs to interpret. This contrast indicates that human referring expressions pose a distinct challenge, even when they are sufficiently informative for human matchers. Importantly, however, the overall protocol-level patterns remain similar: STATIC and FULL are strongest, PARTIAL is more variable, and NONE remains difficult.

| Reasoning | GPT-5.4 | GPT-5.4-Mini |
|---|---|---|
| None | 0.6 | 0.2 |
| Medium | $0.8_{+0.2}$ | $0.1_{-0.1}$ |
| High | $0.9_{+0.3}$ | $0.0_{-0.2}$ |

Table 13: Average number of questions per target item under FULL for different reasoning-effort settings. Subscripts report differences relative to the default no-reasoning setting; green indicates increases and red indicates decreases.

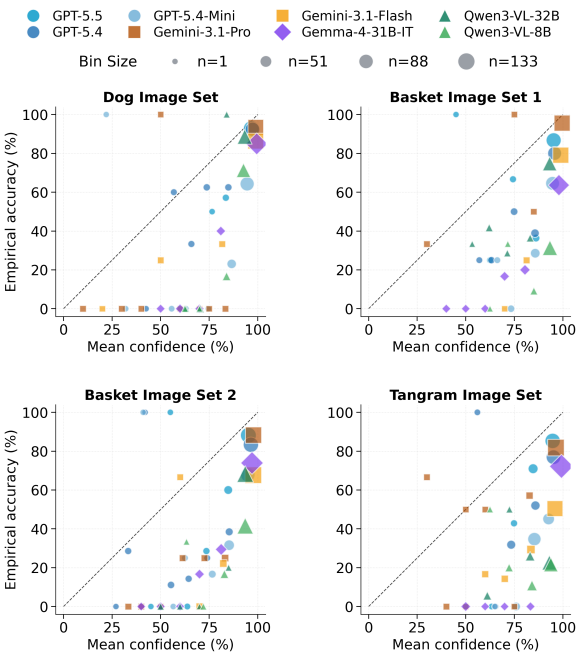


Figure 8: Protocol-level calibration from the follow-up experiments with AI-generated descriptions. Points show model-specific confidence bins from 10 evenly spaced bins between 0 and 100; point size is proportional to bin count. The dashed diagonal denotes perfect calibration.

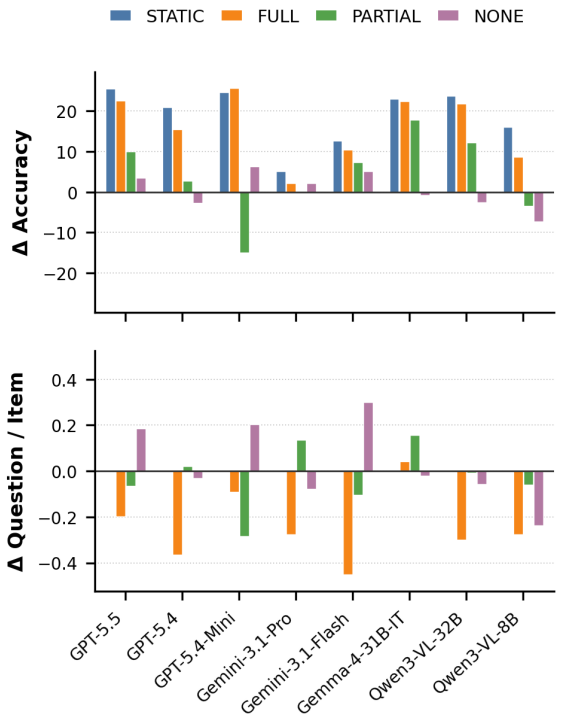


Figure 9: Differences in LVLM outcomes and question-asking behavior between AI- and human-generated descriptions, averaged across all four datasets. Bars show $\Delta$ = AI-generated − human-generated descriptions for accuracy and questions per item.

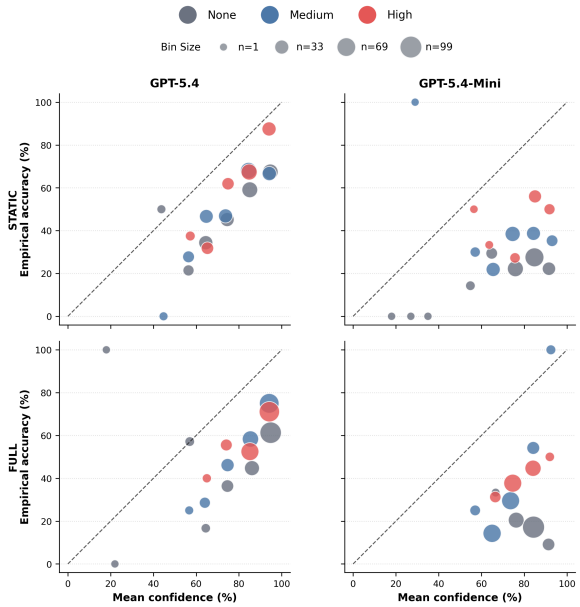


Figure 10: Calibration under different reasoning-effort settings for GPT-5.4 and GPT-5.4-Mini. Each point represents a confidence bin, with the $x$-axis showing mean reported confidence and the $y$-axis showing empirical accuracy. Point size indicates bin size, colors denote reasoning effort, and the dashed diagonal marks perfect calibration.

## E.2 Reasoning Effort

**Reasoning-effort setup.** Default reasoning effort is model-specific rather than uniform across GPT models. GPT-5.5 uses medium reasoning effort by default, whereas GPT-5.4 defaults to no reasoning. We therefore vary the reasoning effort of GPT-5.4 and GPT-5.4-Mini through the official Responses API to isolate the effect of test-time reasoning.

**Reasoning effort changes question asking modestly.** Table 13 shows that higher reasoning effort slightly increases question asking for GPT-5.4 under FULL, while GPT-5.4-Mini asks very few questions across settings. Together with Table 4, this suggests that reasoning improves interactive performance partly by supporting information seeking, but not simply by increasing question volume.

**Reasoning effort slightly improves calibration, but gains plateau from medium to high.** Figure 10 shows that higher reasoning effort generally brings more calibration bins closer to the diagonal, especially when accounting for bin sizes, indicating slightly better alignment between confidence and accuracy. However, the shift from medium to high reasoning is small, suggesting that scaling reasoning effort does not reliably improve calibra-

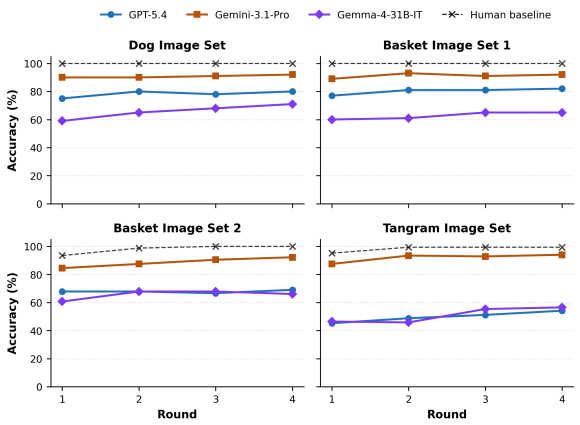


Figure 11: Accuracy (%) across rounds on the four image sets under STATIC.

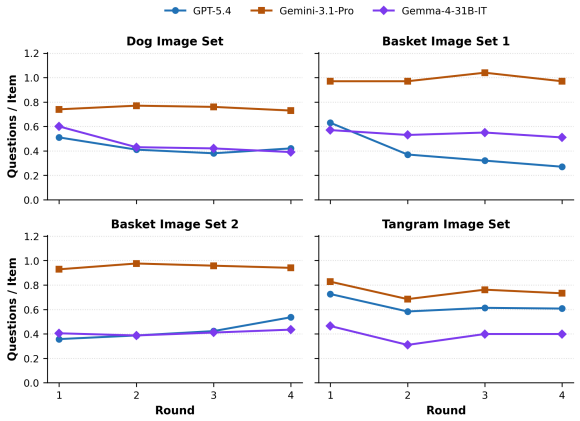


Figure 12: Average number of questions per item across rounds on the tangrams under FULL.

tion. Most bins still fall below the diagonal, so both models remain overconfident.

### E.3 Multi-Round Repeated References

**Per-dataset accuracy results under STATIC.** Figure 11 shows model performance on tangrams across rounds under STATIC. Models improve matching performance even without the ability to ask questions, confirming that they can comprehend the increasingly efficient referring expressions generated by humans over repeated interaction.

**Question asking under FULL.** Figure 12 shows the average number of questions per item across rounds under FULL. Question rates are generally stable or decrease slightly over rounds for most models and datasets.

**Model calibration across rounds.** Figure 13 shows round-level calibration across datasets, aggregating results from STATIC and FULL because the two protocols exhibit broadly similar patterns. Models remain imperfectly calibrated across rounds: many confidence bins fall below the diagonal, indicating overconfidence. Calibration does not consistently improve with repeated exposure, even when accuracy increases over rounds. This suggests that LVLMs can sometimes benefit from repeated references without reliably improving their uncertainty estimates.

### E.4 Alternative Director Robustness

**Experimental procedure.** To test whether Matcher results depend on one simulated information source, we replace GPT-5.4 with Gemini-3.1-Flash as the Director and rerun FULL and PARTIAL on Basket Image Set 2 and Tangrams. We evaluate both human- and GPT-5.4-generated initial descriptions and otherwise preserve the original prompts, candidate views, and turn budget.

**Results.** Table 14 shows that individual model–protocol combinations vary with the Director, as expected in a collaborative task, but Matcher rankings remain highly stable. Rank correlations between results with GPT-5.4 and Gemini-3.1-Flash Directors are $\rho = .929$ under FULL and $\rho = .976$ under PARTIAL. GPT-5.4 is generally the stronger Director, yet the main protocol-level patterns and Matcher hierarchy do not depend on using GPT-5.4 alone. The correlations are computed over Matcher-level accuracies aggregated across the two datasets and description sources.

### E.5 Face-Domain Robustness

**Experimental procedure.** We evaluate five representative Matchers on six image sets from the public FACES database (Holland et al., 2018). Each set contains 12 images of one individual displaying varied facial expressions. Because these image sets do not include human interaction data, we generate initial descriptions with GPT-5.4 and evaluate the same four protocols. Director and Matcher indices are randomized independently as in the main study. This experiment contains no protocol-matched human baseline and is intended as a robustness check in another fine-grained within-category context.

**Results and scope.** Table 15 reproduces the main aggregate pattern: STATIC and FULL are comparable (48.7% and 47.8%), PARTIAL is lower (41.9%), and NONE is most difficult (30.7%). The

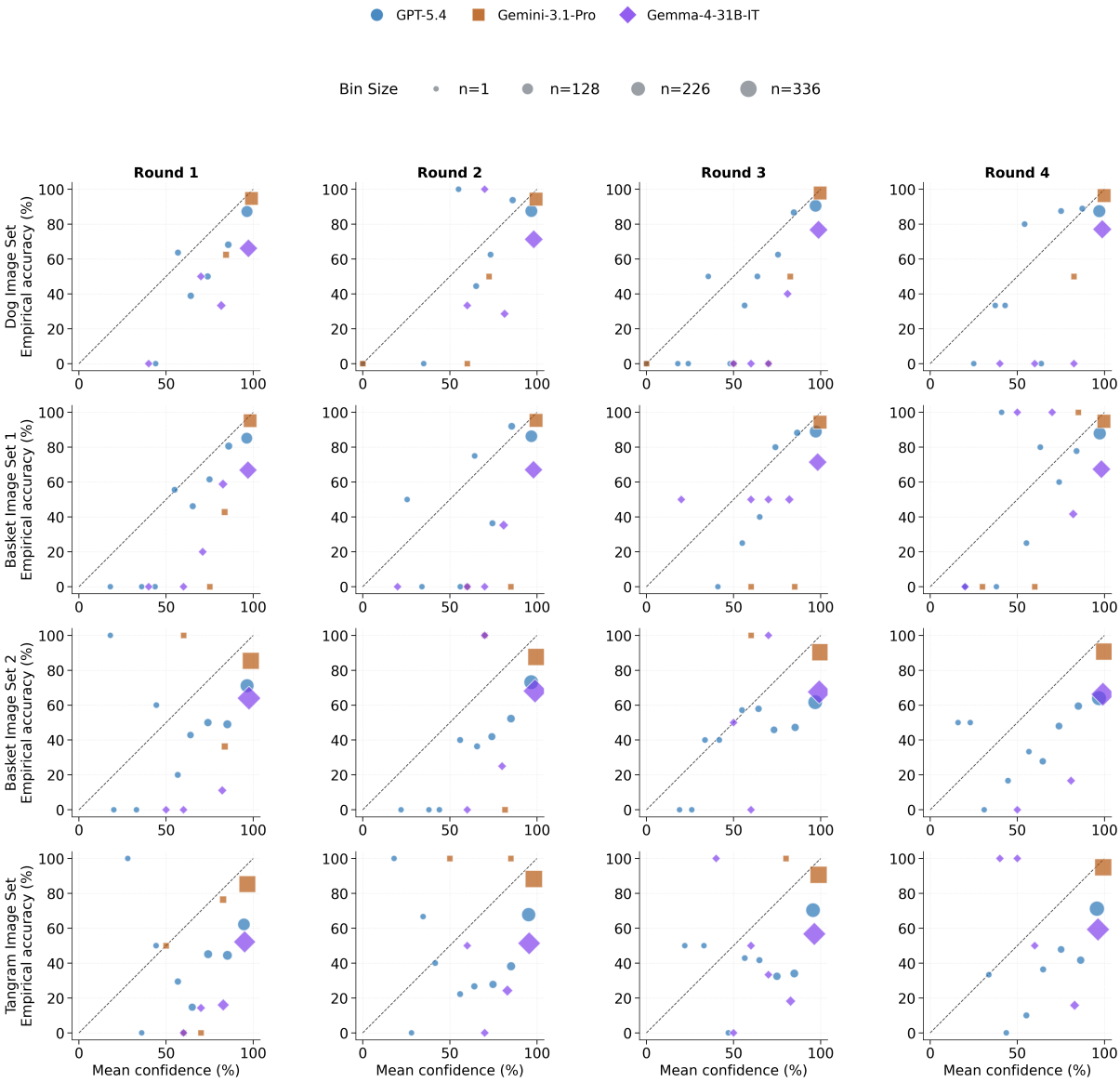


Figure 13: Round-level calibration across datasets in the four-round follow-up experiments. Points show model-specific confidence bins from 10 evenly spaced bins between 0 and 100; point size is proportional to bin count. The dashed diagonal denotes perfect calibration.

| Matcher | GPT-5.4 Director | | Gemini-3.1-Flash Director | |
|---|---|---|---|---|
| | FULL | PARTIAL | FULL | PARTIAL |
| GPT-5.5 | 59.5 | 61.3 | 59.5 | 55.7 |
| GPT-5.4 | 62.8 | 58.3 | 57.4 | 51.2 |
| GPT-5.4-Mini | 39.3 | 28.6 | 38.7 | 28.0 |
| Gemini-3.1-Pro | 81.5 | 83.6 | 86.3 | 85.1 |
| Gemini-3.1-Flash | 58.0 | 43.8 | 58.3 | 37.5 |
| Gemma-4-31B-IT | 54.5 | 54.8 | 53.9 | 46.4 |
| Qwen3-VL-32B | 33.3 | 28.3 | 33.0 | 33.0 |
| Qwen3-VL-8B | 23.2 | 27.4 | 16.1 | 21.1 |

Table 14: Accuracy (%) with GPT-5.4 and Gemini-3.1-Flash Directors, aggregated across Basket Image Set 2 and Tangrams and across human- and AI-generated descriptions. Matcher rankings across Directors remain highly correlated ($\rho = .929$ for FULL and .976 for PARTIAL).

| Matcher | STATIC | FULL | PARTIAL | NONE |
|---|---|---|---|---|
| GPT-5.4 | 53.7 | 55.1 | 56.9 | 52.3 |
| GPT-5.4-Mini | 44.9 | 47.7 | 26.9 | 16.2 |
| Gemini-3.1-Flash | 55.6 | 50.0 | 38.0 | 17.6 |
| Gemma-4-31B-IT | 55.1 | 54.6 | 54.2 | 49.4 |
| Qwen3-VL-8B | 34.3 | 31.5 | 33.3 | 20.8 |
| Overall | 48.7 | 47.8 | 41.9 | 30.7 |

Table 15: Accuracy (%) on six face image sets with AI-generated descriptions. Each image set contains 12 images. There is no protocol-matched human baseline for this robustness experiment.

result extends the protocol-level observation beyond dogs, baskets, and tangrams, but faces remain a controlled within-category matching context; this study does not establish generality to cluttered scenes, spatial relations, or open-world visual search.

## F Prompts for Visual Grounding Experiments

This section reports the prompts used in the study for the visual grounding experiments.

### F.1 Task Overview

We reuse the same task overview prompt for both the Director and Matcher across all interaction protocols to maximize the comparability across the four interaction protocols.

```
### Task Overview

You will participate in a visual reference task involving two roles: Director and Matcher. The two
     participants may not share the same visual context.

Both participants are shown randomized sets of indexed visual items. These visual items may be
    objects, scenes, or other image-based stimuli. The participants share a subset of target
    items, but the Matcher may also see additional distractor items that are not visible to the
    Director.

The goal is for the Director and Matcher to collaborate so that the Matcher can identify each
    target item as accurately as possible.

The task may consist of multiple rounds. Please strictly follow the instructions for your assigned
     role across all rounds.
```

### F.2 Matcher Output Format Instructions

Across all interaction protocols, Matcher is prompted to produce a JSON object containing the target item index and its confidence level (%) rated from 0 to 100.

The output format instruction for the STATIC protocol is as follows.

```
### Output Format Instruction

Output your response in JSON format using the following structure:

{
    "target_item_index": 5,
    "confidence": 75
}

**Field Definitions**

- "target_item_index": the index of the item you are selecting as the match for the identified
    target item.
- "confidence": your confidence level (%) in the selection, rated from 0 (not confident at all) to
     100 (most confident).
```

The output format instruction for the other three interactive protocols (FULL, PARTIAL, and NONE) is as follows.

```
### Output Format Instruction

Output your response in JSON format using the following structure:

{
  "utterance": "Your natural, concise utterance to the Director.",
  "selections": [
    {
      "nth_item": 2,
      "target_item_index": 5,
      "confidence": 75
    },
```

```
    ...
  ]
}

**Field Definitions**

- "utterance" should contain your natural language utterance you say to the Director. Do not
    mention indices or positional cues here.
- "selections" should be a list of items you are selecting based on the Director's responses. Each
     item selection should include:
  - "nth_item": the order of the target item being identified (starting from 1 for the first item
    identified in the current round).
  - "target_item_index": the index of the item you are selecting as the match for the identified
    target item.
  - "confidence": your confidence level (%) in the selection, rated from 0 (not confident at all)
    to 100 (most confident).

**Output Rules**

- If you do not have enough information to make a selection, set "selections" to an empty list.
- If you have enough information to make a selection for the current item, include that item in "
    selections".
- If you are revising one or more earlier selections, include each revised item as a separate
    entry in "selections".
- If you are both revising earlier selections and selecting the current item, include all of those
     updates in "selections" in the same response.
- Unless you are revising an earlier choice, "selections" should just contain only one entry for
    the most recently identified item.
- Do not include duplicate entries for the same nth_item in a single response.
- If you revise a previous selection, your "utterance" may briefly acknowledge the correction, but
     it must not disclose any item index or positional information.
```

### F.3 STATIC Protocol

Since there is no interaction between the Director and Matcher under this protocol, we simulate the Director model by programmatically sending the Matcher a full item description at a time without actually prompting a Director LVLM. The Matcher system prompt is given below.

```
### Task Overview

{Task Overview}

### Role Instructions

You are the Matcher. In this role, you must carefully analyze the Directors descriptions and
    choose the most likely match for each target item based on the visual items visible to you.

### Output Format Instruction

{Matcher Output Format Instruction}
```

### F.4 FULL Protocol

**Director system prompt.**

```
### Task Overview

{Task Overview}

### Role Instructions

You are the Director. In this role, you must describe each visual item visible to you from the
    lowest index to the highest index one at a time and answer any questions the Matcher may have.
```

```
Analyze the Matchers question carefully based on your visual context and answer the question to
    the best of your ability.

### Communication Guidelines

- Describe exactly one target item at a time.
- Do not move on to the next item until the Matcher signals to proceed. Wait for them to signal
    when they are ready.
- Use natural, human-like, simple, and clear language that supports efficient communication.
- Try your best to make each description unambiguous and sufficiently informative for the Matcher
    to identify the related target item.
- Keep each description within 50 words, favoring shorter descriptions when clarity is maintained.
- Do not reference display position, display order, neighboring-item relations, or explicit item
    indices. You may describe within-item spatial structure when it is part of the item's
    appearance.
```

**Matcher system prompt.**

```
### Task Overview

{Task Overview}

### Role Instructions

You are the Matcher. In this role, you must carefully analyze the Directors descriptions and ask
    any clarification questions if needed to identify each target item based on the visual items
    visible to you.

### Communication Guidelines

- In each turn, you are allowed to ask a question or signal to proceed to the next target item,
    but keep it no longer than 50 words.
- If you ask a question or a confirmation, you must not internally select any target item until
    the Director provides an answer to your question.
- If you signal to proceed, you must have internally selected a target item in your output.
- Use concise, natural, human-like language and never mention item indices or positional cues in
    your natural-language utterance.
- You may ask as many questions as needed to clarify your understanding of the target items, but
    avoid asking redundant or irrelevant questions.
- Once you have identified a target item, you can move on to the next one until all the target
    items have been identified.
- You can ask questions about your previous selection(s), in case you want to double check or
    revise them.
- Do not reference display position, display order, neighboring-item relations, or explicit item
    indices. You may describe within-item spatial structure when it is part of the item's
    appearance.

### Output Format Instruction

{Matcher Output Format Instructions}
```

### F.5 PARTIAL Protocol

**Director system prompt.**

```
### Task Overview

{Task Overview}

### Role Instructions

You are the Director. In this role, you must describe each visual item visible to you from the
    lowest index to the highest index one at a time and answer any questions the Matcher may have.
```

```
Analyze the Matchers question carefully based on your visual context and answer the question to
     the best of your ability.

### Communication Guidelines

- Describe exactly one target item at a time.
- Do not move on to the next item until the Matcher signals to proceed. Wait for them to signal
     when they are ready.
- Use natural, human-like, simple, and clear language that supports efficient communication.
- For each new target, begin with an underspecified description to encourage the Matcher to ask
     clarifying questions.
- When responding to the Matcher's questions, be helpful and informative, but also be incremental
     and partial descriptions at a time, instead of over-sharing.
- Do not reference display position, display order, neighboring-item relations, or explicit item
     indices. You may describe within-item spatial structure when it is part of the item's
     appearance.
```

**Matcher system prompt.**

```
### Task Overview

{Task Overview}

### Role Instructions

You are the Matcher. In this role, you must carefully analyze the Directors descriptions and ask
     any clarification questions if needed to identify each target item based on the visual items
     visible to you.

### Communication Guidelines

- In each turn, you are allowed to ask a question or signal to proceed to the next target item,
     but keep it no longer than 50 words.
- If you ask a question or a confirmation, you must not internally select any target item until
     the Director provides an answer to your question.
- If you signal to proceed, you must have internally selected a target item in your output.
- Use concise, natural, human-like language and never mention item indices or positional cues in
     your natural-language utterance.
- You may ask as many questions as needed to clarify your understanding of the target items, but
     avoid asking redundant or irrelevant questions.
- Once you have identified a target item, you can move on to the next one until all the target
     items have been identified.
- You can ask questions about your previous selection(s), in case you want to double check or
     revise them.
- Do not reference display position, display order, neighboring-item relations, or explicit item
     indices. You may describe within-item spatial structure when it is part of the item's
     appearance.

### Output Format Instruction

{Matcher Output Format Instruction}
```

### F.6 NONE Protocol

**Director system prompt.**

```
### Task Overview

{Task Overview}

### Role Instructions
```

```
You are the Director. In this role, you must answer the Matchers questions using only one of the
     following responses: Yes, No, or Unclear.

Analyze the Matchers question carefully based on your visual context and answer the question to
     the best of your ability.

### Communication Guidelines

- Just output Yes, No, or Unclear without any additional explanation or information in each turn.
- If the Matcher's response indicates the target item index they selected, just respond with "
     Unclear."
```

**Matcher system prompt.**

```
### Task Overview

{Task Overview}

### Role Instructions

You are the Matcher. In this role, you must ask only yes or no questions to the Director to
     identify each target item based on the visual items visible to you. There are a total of \
     $num_targets target items for you to identify in each round.

### Communication Guidelines

- In each turn, you are only allowed to ask a clear and concise question at a time, no longer than
      50 words.
- Avoid asking questions that require the Director to provide information beyond "Yes," "No," or "
     Unclear."
- Use concise, natural, human-like language and never mention item indices or positional cues in
     your natural-language utterance.
- You may ask up to \$max_questions questions as needed to clarify your understanding of the
     target items, and avoid asking redundant or irrelevant questions.
- Once you have identified a target item, you can move on to the next one until all the \
     $num_targets target items have been identified.
- You can ask questions about your previous selection(s), in case you want to double check or
     revise them.
- For each question, make clear which target item it refers to, such as "the first target item," "
     the fifth target item," and so on.

### Output Format Instruction

{Matcher Output Format Instruction}
```

# G Prompts for Referring Expression Processing

## G.1 Referring Expression Extraction

Prompt used for the automatic extraction of referring expressions for each target visual referent produced in a referential conversation. Words within "<>" denote placeholders.

This prompt is from Zeng et al. (2026) and we only use it for the dog and basket datasets. The tangrams dataset already comes with manually curated object descriptions for each target object.

```
This is an extractive task.

You will be given a transcript of a conversation between two participants engaged in a
     collaborative object-matching task. There are exactly <num_objects> target objects. One
     participant (the describer) describes each target object, and the other participant (the
     matcher) attempts to identify them.

Your task is to extract the descriptive phrases used by the describer for each target object.
- Extract phrases verbatim from the transcript.
```

```
- Do not extract the whole utterance, only the descriptive phrases.
- Exclude disfluencies, fillers, and false starts (e.g., "um", "uh", "like").
- Do not paraphrase or infer missing information.
- Each object may have one or multiple descriptive phrases.

Return the results in the following JSON format:

{
    "object_#1": "descriptive phrases for object 1",
    "object_#2": "descriptive phrases for object 2",
    ...
    "object_#<num_objects>": "descriptive phrases for object <num_objects>"
}

Example description phrases:

- doesn't have handle, tip of it is thicker than rest of body, brownish color, weaves are in
    squares if you look at it directly
- half circle, no handles, top tip of it is a little bit thicker than rest of body
- tip which is a little bit thicker than rest of body
- tip that is a little bit larger than body, looks a little bit thicker

Transcript:

<transcript>

Output only the JSON object. Do not include any additional text or explanations.
```

## G.2 Partial Description Extraction

```
Given a description map in JSON format, convert it into a new JSON object with the same top-level
    keys.

1. "full": the original full description as a string, unchanged.
2. "attributes": a list of atomic attributes extracted from the full description.
3. "least": the least informative attribute for this visual item, selected from its "attributes"
    list.

Attribute extraction rules:
- Each attribute should be expressed in natural language.
- Each attribute must describe only one visual feature.
- Keep attributes as close as possible to the original wording; verbatim if possible.
- Do not infer attributes that are not explicitly stated in the description.
- Do not treat missing attributes as negative evidence.
- Remove duplicate attributes within the same item.
- Split compound descriptions into separate atomic attributes.
- Preserve relative/comparative attributes when explicitly stated, such as "tallest", "shortest",
    "second shortest", or "shorter than 9".
- Resolve coreferences when extracting attributes, so that each attribute is self-contained and
    does not rely on pronouns or other references to other attributes.
- If there is only one attribute in the description, simply reuse the full description as the
    single attribute.

Least attribute selection rules:
- For each item, choose exactly one attribute from that item's "attributes" list.
- Choose the attribute that is least informative for identifying the item, taking into account all
     attributes across all items in the input map.
- Do not choose an attribute that gives the gestalt impression of the whole item; instead, choose
    an attribute that describes a specific visual feature.
- Prefer a generic, common, or weakly distinguishing attribute over an attribute that is
    distinctive or rare across items.
- If several attributes are similarly generic, choose the one that would be least helpful on its
    own for distinguishing the item from the other items.
- Do not invent, rewrite, or combine attributes for "least"; it must exactly match one string in
    that item's "attributes" list.

Return only valid JSON. Do not include explanations, markdown, or comments.
```

```
Input format:

{
  "image_id": "full natural-language description"
}

Output format:
{
  "image_id": {
    "full": "original full natural-language description",
    "attributes": [
      "attribute 1",
      "attribute 2"
    ],
    "least": "attribute 2"
  }
}
```

### G.3 Object Description Generation

We use the following prompt to generate full descriptions for each object across the four image sets with GPT-5.4. To mirror the human Director setting, the model is shown only the target images. We prompt it to produce a description map that can be used directly in our experimental pipeline.

```
You are creating a full description map for a visual reference task.

The task involves two roles: a Director and a Matcher. The Director sees a set of target visual
     items and must describe them so the Matcher can identify the same items among possible
     distractors. Your job is to write the Director-style initial descriptions for one round.

You will receive all director images for the round in a single message. Each image is labeled with
      its filename only so the output can be keyed correctly.

Return only valid JSON: an object mapping each exact filename to one short natural-language
     description.

Description rules:
- Describe every provided image exactly once.
- Use the exact filenames as JSON keys.
- Each description must be fewer than 50 tokens.
- Use natural, human-like, simple, and clear language.
- Make each description unambiguous and sufficiently informative for a Matcher to identify the
     item among similar items.
- Describe the item itself: shape, pose, orientation, silhouette, distinctive parts, and within-
     item spatial structure when useful.
- Do not reference display position, display order, neighboring-item relations, explicit item
     indices, file paths, or filenames in the descriptions.
- Do not say "image", "picture", "item number", or similar display-oriented labels in the
     descriptions.
- Do not include markdown, comments, explanations, or extra keys.
```

## H Dataset Details

Figures 14, 15, 16, and 17 show the four object image sets used in our experiments. These sets span real-world objects (dogs and baskets) and abstract tangrams, and differ in the number of target and distractor items. For the basket and dog sets, the Director sees only the target items, while the Matcher sees the full candidate set. For the tangram set, both players see the full set. In all experiments, item orders for Director follow the item-description order in the original dataset, while Matcher's item orders are randomized to prevent trivial index matching or data contamination.

We obtain the image sets and corresponding human–human interaction data from the GitHub repositories associated with the cited papers.

Figure 14: Basket image set from Wang et al. (2025a). The first 10 baskets are target items visible to the Director; the Matcher sees the full set.

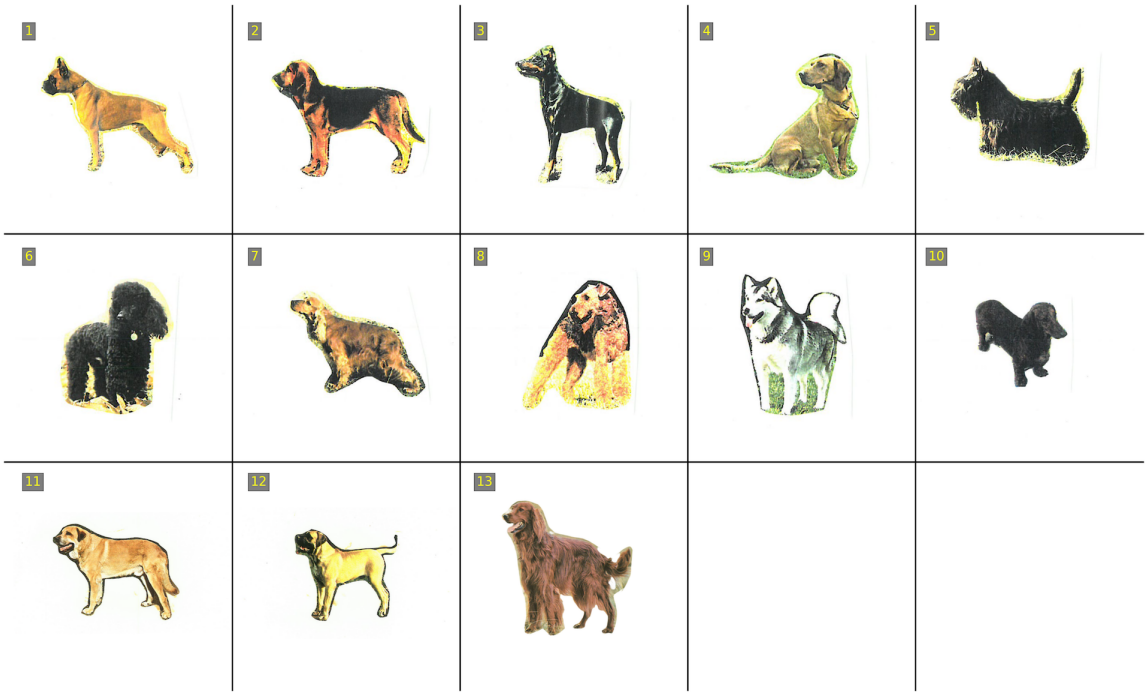


Figure 15: Dog image set from Wang et al. (2025a). The first 10 dogs are target items visible to the Director; the Matcher sees the full set.

Figure 16: Basket image set from Zeng et al. (2026). The first 12 baskets are target items visible to the Director; the Matcher sees the full set.

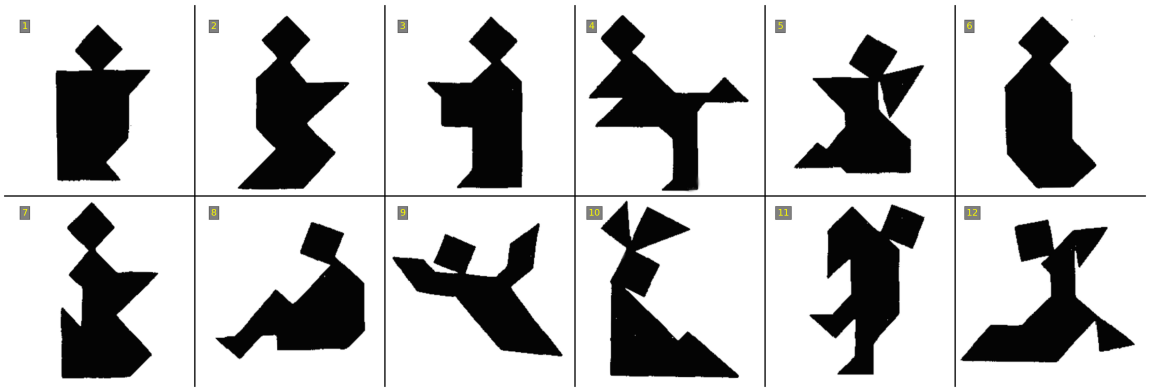


Figure 17: Tangram image set from Hawkins et al. (2020). Both the Director and Matcher see the full set; item orders are randomized independently for the two views in our experiments.