Measuring How (Not Just Whether) VLMs Build Common Ground

Saki Imai **Mert İnan** **Anthony Sicilia** **Malihe Alikhani**
Northeastern University, Boston MA
{imai.s, m.alikhani}@northeastern.edu

Abstract

Large vision language models (VLMs) increasingly claim reasoning skills, yet current benchmarks evaluate them in single-turn or question answering settings. However, grounding is an interactive process in which people gradually develop shared understanding through ongoing communication. We introduce a four-metric suite (*grounding efficiency*, *content alignment*, *lexical adaptation*, and *human-likeness*) to systematically evaluate VLM performance in interactive grounding contexts. We deploy the suite on 150 self-play sessions of interactive referential games between three proprietary VLMs and compare them with human dyads. All three models diverge from human patterns on at least three metrics, while GPT4o-mini is the closest overall. We find that (i) task success scores do not indicate successful grounding and (ii) high image-utterance alignment does not necessarily predict task success. Our metric suite and findings offer a framework for future research on VLM grounding.

Figure 1: This figure shows our VLM evaluation suite using the Photobook Task to test the grounding capabilities and human-likeness of language models, in addition to two other metrics. Our benchmarking strategy clearly identifies the differences between VLM and human interactions.

1 Introduction

To build collaborative AI systems, it is not enough to produce locally correct answers; agents must *establish common ground* efficiently through interaction, as humans do in situated dialogue (Clark and Brennan, 1991).[1] Human partners achieve this via rapid lexical entrainment (Brennan and Clark, 1996; Krauss and Weinheimer, 1964; Brennan, 1996; Garrod and Anderson, 1987), and multi-level interactive alignment (Pickering and Garrod, 2004), that yields shorter, more precise utterances over time. These are shown to both increase task success (Reitter and Moore, 2007) and reduce cognitive load.

Despite growing support for multi-turn interaction, contemporary training and evaluation pipelines for large language and vision language models (VLMs) still prioritize single-turn answer quality, through supervised fine-tuning, RLHF (Ouyang et al., 2022), or DPO (Rafailov et al., 2023). As a result, they neither measure nor reward the interactive skills that underpin grounding, such as reusing a partner's words or pruning redundant detail once mutual understanding is achieved. Recent evidence also points to low communication efficiency and degraded multi-turn performance relative to single-turn settings (Hua and Artzi, 2024; Laban et al., 2025).

In this paper, we operationalize grounding for

[1]Following Clark (1996), we use *common ground* to mean the set of propositions that all interlocutors mutually believe, know that the others believe, and recognize as a basis for subsequent action.

Proceedings of Recent Advances in Natural Language Processing, pages 441–451
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-053

multimodal dialogue and evaluate it directly. We introduce a task-agnostic evaluation metric suite that captures *grounding efficiency*, *content alignment*, *lexical adaptation*, and *human-likeness*—and instantiate it on the PhotoBook referential game (Haber et al., 2019), which features five-round dialogues to identify shared images (Fig. 1; §4). [2] While the original Photobook corpus studied the common ground formulation and referring expression generation using LSTM models, they did not (i) test with contemporary billion-scale VLMs, (ii) allow model-model self-play, or (iii) quantify distributional human-likeliness. We benchmark contemporary VLMs in model-model self-play and compare against human transcripts to ask:

1. How efficiently do VLM pairs reach common ground compared to humans? (grounding efficiency)
2. Do VLMs describe the exact visual cues and is that predictive of task success? (content alignment)
3. Do VLM pairs form human-like conceptual pacts, reusing each other's terms and pruning redundant detail over rounds? (lexical adaptation)
4. To what extent do the grounding behaviors of VLM pairs resemble human dialogue patterns at the distributional level? (human-likeliness)

We show that VLM diverges from human baselines on ≥ 3 metrics, where GPT4o-mini is closest overall (§5). Notably, high image-utterance alignment does not guarantee task success—there is no correlation between CLIPScore and task outcomes (§5.2). Finally, we show that task success does not imply grounding. GPT4.1 often inflates its score by mirroring partner's preferences when ground-truth labels coincide (§6).

2 Related Work

Common ground and lexical entrainment Many lines of research in cognitive science and linguistics have focused on modeling common ground establishment in human-human interactions. Studies demonstrated that conversational partners converge on concise, mutually understood referring expressions across successive turns (Krauss and Weinheimer, 1964; Clark and Wilkes-Gibbs, 1986; Brennan, 1996; Garrod and Anderson, 1987). The interactive alignment model (Pickering and Garrod, 2004) proposed that lexical, syntactic and discourse level alignments emerge and support higher level coordination and mutual understanding.

Computational models have attempted to replicate these behaviors in task oriented dialogue systems (Stoyanchev and Stent, 2009; DeVault et al., 2011; Visser et al., 2014; Ohashi and Higashinaka, 2022). However, much of this work has focused on text or spoken dialogue systems. Large VLMs have only recently been examined for lexical adaptation (Hua and Artzi, 2024). Our work builds on these insights by operationalizing common ground into four concrete metrics.

Visual reference games Reference games provide a controlled environment to study grounding processes. In these tasks, participants must identify shared referents, through dialogue (Krauss and Weinheimer, 1964; Clark and Wilkes-Gibbs, 1986; Hawkins et al., 2017; Monroe et al., 2017). These tasks have been adapted for computational modeling to evaluate alignment, reasoning, and visual understanding in interactive contexts (He et al., 2017; Hawkins et al., 2017, 2020). The PhotoBook dataset (Haber et al., 2019), used in our study, extends this paradigm to three-round dialogues. This offers an environment to study common ground formation across time.

More recent work has explored reference games for model evaluation in both abstract and grounded domains (Ji et al., 2022; Chalamalasetti et al., 2023; Hakimov et al., 2025). However, these approaches typically treat interaction as a means to an end, without probing how communicative strategies evolve. In contrast, we introduce a new set of metrics and visualizations to trace the evolution of grounding behaviors and lexical strategies across dialogue rounds.

VLM evaluation in multimodal interactions While large-scale VLMs have achieved impressive zero-shot accuracy on static benchmarks (Liu et al., 2023; Achiam et al., 2023; Liu et al., 2024), we still lack an understanding of how they behave in extended, collaborative interaction (Sicilia et al., 2022). Recent work suggests that VLMs fall short of interactive human behaviors. Hua and Artzi (2024), for example, show that even the most capable multimodal models struggle to adapt to their partner's word choices in-context, and tend to produce verbose but low-efficiency utterances. Our study simulates VLM-VLM dialogues in the Photo-

[2] https://github.com/sakimai/vlm-grounding-benchmark

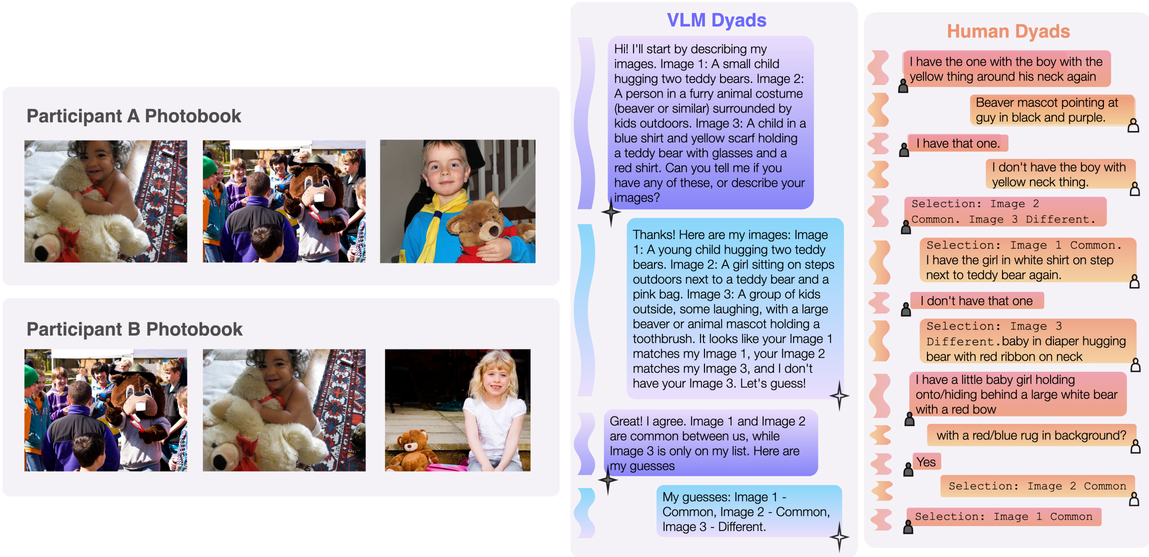

Figure 2: This figure shows a comparative example of human-human and VLM dyadic conversations on the same Photobook game. Human dialogues are more incremental, contain shorter utterances, and have more turns, while VLMs engage in long explanations for each image, with fewer turns, and with sycophancy. The images are from the MSCOCO dataset (Lin et al., 2014).

Book game, comparing them to human interactions. This approach, unlike prior work using synthetic or asymmetrical dialogue, quantifies how VLMs build common ground, not just if they do.

3 Grounding Evaluation Suite

This section details the controlled setting we use to examine how VLMs build common ground in dialogue (see Figure 2 for an example scenario).

3.1 The PhotoBook Task

Task. PhotoBook is a five round referential game in which two conversational partners must discover which of three images they share, and which are unique to each speaker (Haber et al., 2019). At each round, they privately annotate every image as *common* or *different*. The images in the game extracted from the MS COCO Dataset (Lin et al., 2014) are deliberately structured to be visually similar to elicit non-trivial referring expressions. Moreover, the images appear exactly five times throughout a game, which allow us to study the collaborative referring expression generation and resolution (Clark and Wilkes-Gibbs, 1986).

Human corpus. The released dataset contains 2,506 human–human dialogues (164,615 utterances, 130,322 actions and spans a vocabulary of 11,805 unique tokens). This serves as an empirical upper bound baseline for grounding efficiency. Additional work on referring expression extracted 41,340 referring utterances and 16,525 chains from this dataset (Takmaz et al., 2020).

3.2 VLM Self-Play Protocol

Models. We study three recent proprietary VLMs that differ in size and architecture: (i) GPT4.1 (ii) GPT4o-mini (iii) Claude3.5-Haiku. Each VLM dyad initialized with default parameters plays the same set of 50 games, resulting in a total of 150 games. This setup enables a comparison of lexical strategies, task performance, and communicative behavior between VLMs and human.

Prompting and turn scheduling. We instantiate two agents that alternate turns until both submit non-null guesses. Each turn must be a valid JSON object:

- `"message"` the natural-language utterance,
- `"reference"` either ”Image k” ($k \in \{1, 2, 3\}$) or `null`,
- `"guesses"` `null` until a player is ready, otherwise a three-letter array such as `"C","D","C"` where `"C"` $\equiv$ common, `"D"` $\equiv$ different.

A prompt engineered variant designed to prevent three recurrent failure cases, (1) prematurely revealing guesses, (2) comparing images one by one rather than as a set, (3) generating fillers. Unless stated otherwise, all results employ the original prompt used for the human data, to mirror the setup.

Data summary. Human speakers show clear lexical convergence by round 2, as observed in Figure 2 of Haber et al. (2019). To capture this effect without excessive context length, we stopped the VLM self-play at round 3. From 150 simulated games, we collect 2662 utterances, 101701 tokens. These dialogues constitute the VLM generated corpus used in all subsequent analyses.

3.3 Extracting Referring Expressions

Our downstream metrics (§4.2, §4.3) should operate only on referring expressions, not on meta dialogue (“Ready?”, “Let’s guess”). We therefore processed each utterance to isolate referring expressions. While embedding based filters (BERTScore (Zhang et al., 2019), and CLIPScore (Hessel et al., 2021)) were considered, they are less interpretable. Moreover, we observed that VLMs follow highly regular patterns (*“Image 1 is. . . ”*, *“In my first image. . . ”*), which made rule-based extraction viable.

Linking utterances to images. In the human corpus, the alignment between each utterance and the image being discussed is obtained from the annotation click logs. Participants indicated which images they considered common or different, which served as a proxy for determining the referent of each referring expression. For the VLM self-play, we explicitly prompted the models to include a `"reference"` field in their JSON responses (§ 3.2) to indicate which image each utterance pertained to. However, we observed that a substantial proportion of turns included a `"reference"` field set to `null`. Nonetheless, most referring expressions were explicit within the utterances themselves which allowed us to leverage the textual content to determine the referent image.

Human validation. To validate the pipeline, we randomly sampled 50 rounds ($\approx 10\%$ of the VLM corpus) spanning all three models. Manual annotation shows 0.99 precision, recall of 0.55, yielding $F1$ score of 0.66. We intentionally prioritized precision in our extraction approach to minimize false positives. This design ensures the integrity of analyses we depend on extracted referring expression, even at the cost of recall.

4 Metrics

We formalize four families of metrics: *grounding efficiency*, *content alignment*, *lexical adaptation*, and *human-likeness* each grounded in psycholinguistic theory.

4.1 Grounding efficiency

Psycholinguistic theory suggests that efficient grounding involves refining referring expressions and minimizing unnecessary dialogue over time (Clark and Brennan, 1991; Brennan and Clark, 1996). In human interactions, this is reflected in both reduced lexical effort (fewer words) and more streamlined turn-taking (fewer turns) while maintaining or improving task performance (Hawkins et al., 2020). To evaluate this, we compute:

- **Task success**: Total number of correctly identified common and different images in each round (maximum of 18 points for 3 rounds and 3 images).
- **Word count**: Total number of words produced in each round.
- **Turn count**: Total number of conversational turns in each round.

We report grounding efficiency at 1) game level to capture overall communicative cost and task success, and 2) round level dynamics to evaluate how grounding efficiency evolves as interlocutors accumulate shared knowledge over time.

4.2 Content alignment

Assessing RQ2, we measure how closely utterances align with the visual referents.

Absolute CLIPScore. We compute CLIPScore of the utterance u and the image embedding of the target img_t: $\text{CLIPS}\textsc{core}(u, img_t)$.

Contrastive CLIPScore. Psycholinguistic studies show that humans emphasize diagnostic features, which are properties that uniquely identify the target among distractors (Dale, 1989; Dale and Haddock, 1991; Sedivy, 2003). We capture this with the contrastive score defined as

$$\begin{aligned} \textsc{ClipCon} = {} & \text{CLIPScore}(u, \text{img}_t) \\ & - \tfrac{1}{|D|} \sum_{d \in D} \text{CLIPScore}(u, d) \end{aligned} \tag{1}$$

where u is the utterance (i.e., referring expression), the divisor $|D|$ converts the raw sum into a mean and makes the score invariant to the number of distractors.

4.3 Lexical adaptation

To assess RQ3, we measure whether VLM pairs form human-like conceptual pacts, reusing each other’s terms and pruning redundant detail.

System	Total Score (max 18)	# Words	# Turns
Claude3.5	12.62 ± 2.07	805.72 ± 123.85	15.48 ± 2.52
GPT4.1	15.02 ± 1.81	800.08 ± 116.87	14.68 ± 2.51
GPT-4o-mini	13.52 ± 2.34	428.22 ± 80.74	23.08 ± 2.46
Human	**16.62** ± 1.14	338.10 ± 109.37	74.08 ± 12.08

Table 1: Mean $\pm$ standard deviation for total score, number of words, and number of turns per game across systems. Humans achieve the highest task success while using fewer words but significantly more turns. In contrast, VLMs achieve lower task success with longer word counts and fewer turns.

Word Novelty Rate (WNR). To quantify how speakers adjust their vocabulary as common ground builds, we adopt the Word Novelty Rate proposed by Hua and Artzi (2024). WNR is a variant of word error rate that counts only insertions and substitutions, and ignores deletions. Past work shows that interlocutors progressively drop previously established material once it is mutually known (Hawkins et al., 2020). By focusing on insertions and substitutions, WNR captures the moments where a speaker adds or changes wording, i.e. where lexical innovation or repair occurs. A declining WNR across rounds indicates that fewer novel words are being introduced, consistent with successful adaptation.

4.4 Human-likeness

Discrete energy distance. To gauge how *human-like* VLM utterances are at the distributional level, we adopt the Discrete-Energy Distance of Sicilia and Alikhani (2022). While our previous metrics target specific grounding mechanisms, this distributional measure captures whether the overall distribution of VLM dialogues resemble human interactions. We first embed each game dialogue with `all-MiniLM-L6-v2` Sentence Transformer.[3] This metric compares the average cross group distance (human-VLM) with the average within group distances (human-human, VLM-VLM). Lower energy distance values indicate that the VLM distribution is closer to the human distribution.

5 Results

5.1 Grounding efficiency

Addressing RQ1, we assess how efficiently VLM pairs establish common ground compared to human speakers. We operationalise grounding efficiency as the balance between communicative cost (measured by total words and turns) and task performance (total score).

Game-level performance. Table 1 summarizes performance across the entire game (capped at 3 rounds) for each system. Humans achieve the highest mean score (16.62), with fewer words (338.1) but more turns (74.08) than VLMs. GPT-4.1 closely approaches human performance in score (15.02) while requiring nearly double the word count and markedly fewer turns. Claude-3.5 shows lower task success (12.62) despite the highest word count (805.72) and reduced turn count.

Round-level performance. Figure 3 analyzes grounding efficiency across rounds. We present the average total score per round, percentage change in word count from round 1, and percentage change in turn count from round 1.

Total score (Fig 3 left) improves with additional rounds for humans and GPT4o-mini. In contrast, GPT4.1 and Claude3.5 exhibit declining scores, possibly due to challenges of managing longer context lengths. Specifically, GPT4.1 and Claude3.5 generate nearly double the total word count per game compared to GPT4o-mini (800 and 806 words vs. 428 words; see Table 1), which may contribute to degraded performance. This observation aligns with prior work demonstrating that LLM performance tends to degrade with longer context windows (An et al., 2024).

Word count (Fig 3 middle) consistently decreases across rounds, and declines significantly for humans in round 3. While all models show reduced word count by round 3, GPT4o-mini and Claude3.5 initially increase their word count in round 2, and this pattern contrasts with humans. This is consistent with psycholinguistic theories of lexical entrainment and collaborative efficiency (Holler and Wilkin, 2020), that the decrease in word count indicates common ground. Further, turn count (Fig 3

[3] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

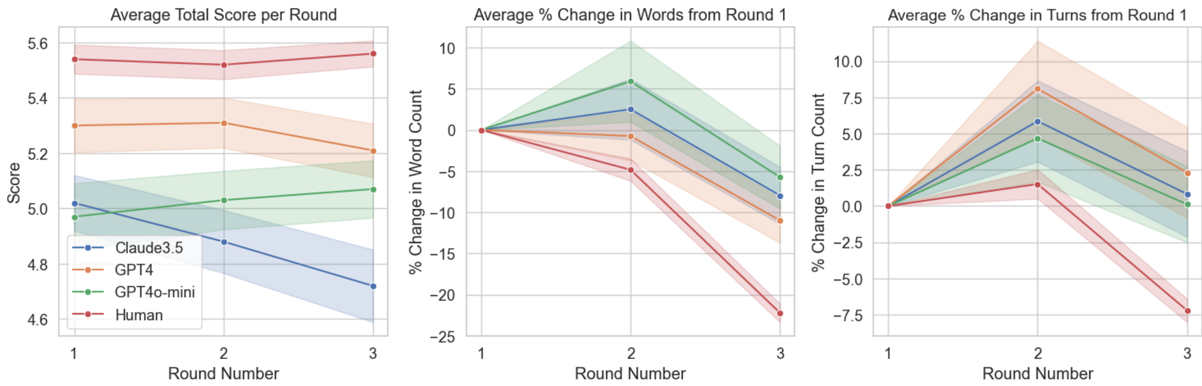

Figure 3: (Left) Average total score per round; (Middle) Average percent change in word count from round 1; (Right) Average percent change in turn count from round 1. Shaded region indicate standard deviation. **Key takeaways:** (i) Humans and GPT4o-mini improve task performance over rounds; (ii) Humans sharply reduce word count and turn count in later rounds; (iii) VLMs show inconsistent cost reduction, with some increasing word and turn counts.

right) follows a similar trend that humans generally reduce their turns across rounds, while VLMs tend to increase turn count from round 1 to 2, before exhibiting a slight reduction in round 3. Overall, these patterns suggest that GPT4o-mini demonstrates a grounding trajectory resembling human efficiency with improved task scores with moderate communicative cost, while GPT4.1 and Claude3.5 struggle with verbosity and performance as rounds progress.

Does prompt tuning increase grounding efficiency? Motivated by these results, when we specifically ask models to be more human-like in the prompt, we observe that their performance, in fact, becomes closer to humans in various metrics. Specifically, we crafted a revised prompt that preemptively mitigates three recurrent failure modes (§ 3.2), to guide VLMs toward more concise and targeted communication. Results suggest that tailored prompting can improve efficiency metrics and promote more adaptive behavior similar to humans, though inherent reasoning limitations persist.

5.2 Content alignment

Figure 4 (left) shows declining absolute CLIPScore for humans. This aligns with prior findings that once common ground is established, speakers economise on explicit visual detail and drop redundant description. GPT4o-mini had a consistent description strategy with absolute CLIPScore of around 31.5 across rounds. In contrast, CLIPScore for Claude3.5 increases across rounds, indicating longer context length might lead the model to provide additional descriptions than prune detail.

The contrastive variant in Fig 4 (right) separates the systems further. GPT4.1 and Claude3.5 briefly spike in diagnostic detail during round 2, whereas GPT-4o-mini and humans follow a flatter, lower trajectory with lower score in round 2.

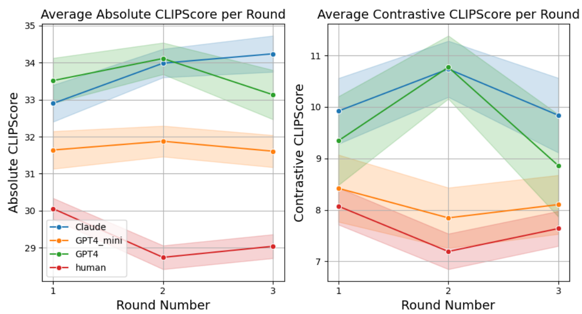

Figure 4: (Left) Absolute CLIPScore. (Right) Contrastive CLIPScore (↑ means the utterance is more diagnostic of the target versus distractors). Shaded region indicate standard deviation. **Takeaway:** Humans steadily lower their CLIPScore while still completing the task, suggesting that they simplify their descriptions as mutual knowledge accrues. LLMs diverge in how they adapt across rounds.

Does alignment drive success? To test whether raw image–utterance similarity translates into better coordination, we relate Absolute CLIPScore to the score actually obtained in the same round (Fig 5). This plot makes the disconnect between alignment and task success explicit, as high and low CLIPScores are scattered across all outcome bins. Results from humans illustrate the point most clearly, as they achieve near-perfect task scores (Table 1) despite the lowest alignment scores (Fig 4). These observations show that CLIP-based alignment metrics capture a surface-level resemblance

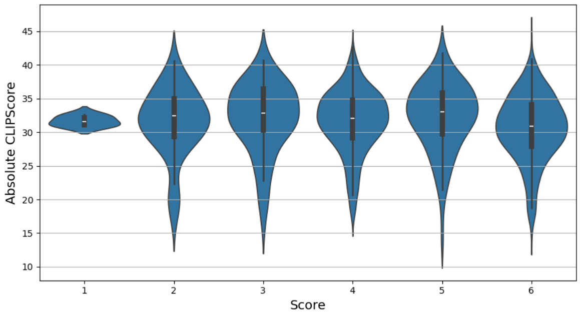

Figure 5: Violin plots show the distribution of Absolute CLIPScore conditioned on the total game score (0–6) in the same round; white horizontal bars mark medians, boxes the inter-quartile range. **Takeaway:** High and low alignment scores scattered across all outcome bins confirm that CLIP-based metrics alone do not predict task success.

between words and pixels, but miss the pragmatic reasoning that enables interlocutors to establish common ground.

5.3 Lexical adaptation

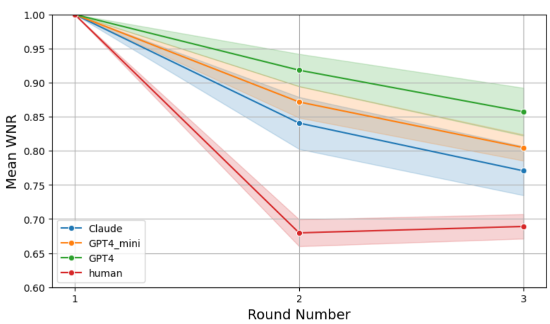

Figure 6: WNR drops over rounds for all systems, but the steepest decline occurs in human referring expressions. Claude and GPT-4o-mini show moderate adaptation; GPT-4 lags behind.

Word Novelty Rate. Fig 6 shows the mean WNR for referring expressions across rounds. In referring expressions, humans achieve the steepest decline in WNR. This pattern reflects the formation of conceptual pacts and lexical stabilization as common ground builds. Other VLMs show moderate and slower adaptation compared to humans.

These metrics show that VLM pairs do not fully replicate human strategies of lexical adaptation. While some VLMs, such as Claude, exhibit partial human-like adaptation in lexical choices, GPT-4 models struggle to stabilize and reuse previously grounded referring expressions.

Model	Energy Distance ↓
GPT4.1	62%
Claude3.5	63%
GPT4-mini	**39%**

Table 2: Distributional human-likeness measured by discrete energy distance. Lower values indicate that a model's utterance distribution is closer to human dialogue. **Takeaway:** GPT4o-mini is the most human-like overall, while Claude 3.5 and GPT-4.1 remain stylistically farther from human discourse.

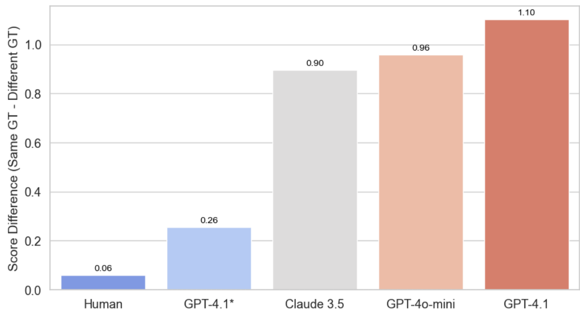

Figure 7: Bar chart showing the performance boost (score difference Δ = Same GT − Different GT) for humans and VLMs. Humans exhibit minimal susceptibility ($\Delta = 0.06$), while GPT4.1 shows the highest score inflation ($\Delta = 1.10$). The prompt-tuned variant GPT4.1* reduces susceptibility ($\Delta = 0.26$), suggesting that targeted instructions can mitigate imitation effects.

5.4 Human-likeliness

Table 2 lists the energy distance between human dialogues and each VLM pair. GPT4o-mini attains the smallest distance (39%), which indicates that utterance distribution is the closest to human data. Claude3.5 and GPT4.1 yield substantially higher distances, suggesting that their dialogue style diverges more from human patterns. Although Claude3.5 matched humans in lexical adaptation (§ 5.3), its higher energy distance reveals that its overall dialogue style still diverges from human discourse.

6 Case Study: Sycophantic VLM Guesses

Even though we have used total task score as a proxy for grounding success, we observe that high scores in VLM dialogues can stem from influences from the other interlocutor rather than grounded mutual understanding. Unlike human players, VLMs often exhibit sycophantic behavior, where they adapt their guesses based on their partner's revealed responses (as illustrated in Figure 2).

Each round requires annotating three images with binary labels (Common or Different), yielding only $2^3 = 8$ possible label combinations. It is possible for dyads to coincidentally receive identical ground-truth labels, even if their visual inputs differ. In our sample of 150 rounds, 56 rounds (˜37%) involved dyads with matching ground-truth labels. If VLMs influence each other's guesses, such cases can inflate scores and create the illusion of successful grounding.

Score inflation analysis. To test whether this was the case, we grouped rounds based on whether the dyad's ground-truth labels were identical or different. We then computed the score difference between these conditions to see the score boost that is attributable to shared ground-truth labels. As shown in Figure 7, humans were robust to shared ground-truth conditions, with only a minor score difference ($\Delta = 0.08$). In contrast, despite achieving the highest average task score among VLMs, GPT4.1 exhibited the greatest susceptibility to this effect, with a score difference exceeding one point ($\Delta = 1.10$). Claude3.5 and GPT4o-mini also exhibited significant susceptibility.

Mitigation experiment. To mitigate this effect, we used a prompt-tuned variant from above (GPT4.1*) that explicitly warned against sharing guesses during the dialogue. This intervention significantly reduced GPT4.1's susceptibility, to achieve the lowest susceptibility to ground-truth alignment among VLMs. This demonstrates that tailored prompts can mitigate such effects.

7 Discussion

Our metric suite shows that the current VLMs nearly reproduce the outcomes of human dialogue (as measured with task success), without reproducing the process by which humans achieve those outcomes. Across efficiency (§5.1), alignment (§5.2), and human-likeliness (§5.4), GPT4o-mini consistently approximates human dialogue most closely. In contrast, while GPT4.1 and Claude3.5 exhibit strengths in task score (§5.1) and lexical adaptation (§5.3), respectively, both models exhibit limitations in other metrics.

Why do the models diverge? We identify three factors. (i) Training data mismatch. Pre-training corpora contain millions of single image captions but almost no multi-round collaborative dialogues. Consequently, models optimize for listing visual details rather than incremental efforts that characterize human conversation. (ii) Reward alignment bias. RLHF typically rewards ”agreeable” or ”helpful” completions. When two VLMs converse, this can over penalise informative disagreement and over reward mirroring. Our case study revealed this pattern with inflated task score whenever two VLMs share ground-truth labels (§6). (iii) Effortless token generation. For VLMs, generating additional tokens is virtually costless, unlike for humans who face cognitive and temporal constraints. In the absence of incentives for brevity, VLMs tend to produce unnecessarily long utterances and rarely reuse previously established shorthand. This helps explain their limited improvement in grounding efficiency (§4.1).

8 Conclusion

In this paper, we introduced a novel benchmarking approach to assess how effectively VLMs establish common ground through interactive dialogue. Unlike previous evaluations that focus solely on task success, our four-metric suite, grounding efficiency, content alignment, lexical adaptation, and human-likeness, enables a more nuanced examination of VLM performance. Our experiments revealed significant differences between human interactions and VLM self-play, highlighting that achieving high accuracy alone does not imply successful grounding or human-like communicative patterns.

These findings underscore critical areas for future development, particularly the need for training methods that encourage incremental, collaborative dialogue rather than isolated, verbose responses. Addressing the biases inherent in reinforcement learning alignment methods and incentivizing conciseness could bring VLM interactions closer to human efficiency. By emphasizing the process rather than merely the outcome, our work provides settings for future research aimed at collaborative, human-like AI communication.

Limitations

Although our framework broadens the evaluation of VLMs beyond single turn accuracy, several limitations should be noted.

We analyze VLM–VLM dialogues to isolate model capabilities without human guidance. In deployment, systems will converse with humans who provide richer pragmatic cues, error corrections, and social feedback. Whether models adapt differently when paired with human partners remains an open question.

Moreover, because we tested on proprietary VLMs, the underlying architectures, training data, and alignment objectives are opaque. This makes it infeasible to determine whether the observed behaviors arise from model scale, fine-tuning or other design choices. Open source replications with transparent training methods are needed to evaluate the generality of our findings.

Ethics Statement

We use the publicly released PhotoBook dataset (Haber et al., 2019), which contains crowd-sourced dialogues and MS-COCO images licensed for research. The dataset does not include personally identifiable information.

All data splits, metric implementations, and analysis scripts will be made publicly available, to enable independent replication and extension.

Acknowledgments

This research was supported in part by the U.S. National Science Foundation under Award No. 2418664. We thank Asteria Kaeberlein for her helpful feedback.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. 2024. Why does the effective context length of llms fall short? *arXiv preprint arXiv:2410.18745*.

Susan Brennan and Herbert Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493.

Susan E Brennan. 1996. Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, 96:41–44.

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11174–11219, Singapore. Association for Computational Linguistics.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

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

Robert Dale. 1989. Cooking up referring expressions. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 68–75, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Robert Dale and Nicholas Haddock. 1991. Content determination in the generation of referring expressions. *Computational Intelligence*, 7(4):252–265.

David DeVault, Kenji Sagae, and David Traum. 2011. Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue & Discourse*, 2(1):143–170.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Sherzod Hakimov, Yerkezhan Abdullayeva, Kushal Koshti, Antonia Schmidt, Yan Weiser, Anne Beyer, and David Schlangen. 2025. Using game play to investigate multimodal and conversational grounding in large multimodal models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5686–5718, Abu Dhabi, UAE. Association for Computational Linguistics.

Robert Hawkins, Minae Kwon, Dorsa Sadigh, and Noah Goodman. 2020. Continual adaptation for efficient machine communication. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 408–419, Online. Association for Computational Linguistics.

Robert XD Hawkins, Michael C Frank, and Noah D Goodman. 2017. Convention-formation in iterated reference games. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 39.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of* *the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776, Vancouver, Canada. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Judith Holler and Katie Wilkin. 2020. Communicating common ground: How mutually shared knowledge influences speech and gesture in a narrative task. In *Speech Accompanying-Gesture*, pages 267–289. Psychology Press.

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

Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert Hawkins, and Yoav Artzi. 2022. Abstract visual reasoning with tangram shapes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 582–601, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.

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. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Atsumoto Ohashi and Ryuichiro Higashinaka. 2022. Adaptive natural language generation for task-oriented dialogue via reinforcement learning. *arXiv preprint arXiv:2209.07873*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Martin J. Pickering and Simon Garrod. 2004. The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, 27(2):212–225.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

David Reitter and Johanna D. Moore. 2007. Predicting success in dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 808–815, Prague, Czech Republic. Association for Computational Linguistics.

Julie C Sedivy. 2003. Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of psycholinguistic research*, 32:3–23.

Anthony Sicilia and Malihe Alikhani. 2022. LEATHER: A framework for learning to generate human-like text in dialogue. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 30–53, Online only. Association for Computational Linguistics.

Anthony Sicilia, Tristan Maidment, Pat Healy, and Malihe Alikhani. 2022. Modeling non-cooperative dialogue: Theoretical and empirical insights. *Transactions of the Association for Computational Linguistics*, 10:1084–1102.

Svetlana Stoyanchev and Amanda Stent. 2009. Lexical and syntactic adaptation and their impact in deployed spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 189–192, Boulder, Colorado. Association for Computational Linguistics.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

Thomas Visser, David Traum, David DeVault, and Rieks op den Akker. 2014. A model for incremental grounding in spoken dialogue systems. *Journal on Multimodal User Interfaces*, 8:61–73.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.