Can Vision Language Models Follow Human Gaze?

Zory Zhang[†] **(hereiszory@gmail.com)**
The University of Illinois, Urbana-Champaign

Pinyuan Feng[†] **(pf2477@columbia.edu)**
Columbia University in the City of New York

Bingyang Wang (icy.bingyang.wang@alumni.emory.edu)
Emory University

Tianwei Zhao (tzhao27@jh.edu)
Johns Hopkins University

Qingying Gao (qgao14@jh.edu)
Johns Hopkins University

Suyang Yu (yu000434@uw.edu)
University of Washington

Ziqiao Ma (marstin@umich.edu)
University of Michigan

Hokin Deng[*] **(hokind@andrew.cmu.edu)**
Carnegie Mellon University

Yijiang Li[*] **(yijiangli@ucsd.edu)**
University of California, San Diego

Dezhi Luo[*] **(ihzedoul@umich.edu)**
University of Michigan

[†]Equal Contribution. Correspondence: hereiszory@gmail.com
[*]Equal Advising.

Abstract

Gaze understanding is suggested as a precursor to inferring intentions and engaging in joint attention, core capacities for a theory of mind, social learning, and language acquisition. As Vision Language Models (VLMs) become increasingly promising in interactive applications, assessing whether they master this foundational socio-cognitive skill becomes vital. Rather than creating a benchmark, we aim to probe the behavioral features of the underlying gaze understanding. We curated a set of images with systematically controlled difficulty and variability, evaluated 111 VLMs' abilities to infer gaze referents, and analyzed their performance using mixed-effect models. Only 20 VLMs performed above chance, with still low overall accuracy. We further analyzed 4 of these top-tier VLMs and found that their performance declined with increasing task difficulty but varied only slightly with the specific prompt and gazer. While their gaze understanding remains far from mature, the patterns suggest that their inferences are far different than merely stochastic parroting. This early progress highlights the need for mechanistic investigations of their underlying emergent inference.

Introduction

To function effectively in human social environments, artificial agents should be able to understand gaze. This foundational socio-cognitive skill likely bootstraps later social learning and cognitive development (Csibra & Gergely, 2009) and is necessary for natural human-artificial intelligence interactions. This demand for gaze understanding motivates us to look at how well Vision Language Models (VLMs; OpenAI, 2024; Gemini et al., 2024, *inter alia*) can infer gaze referents from static visual cues, a basis of more advanced gaze understanding. VLMs are subjects of particular interest for this study because they possess emergent capacities that are poorly understood compared with other technologies, like computer vision expert models. By systematically manipulating variables in the evaluation stimuli, we aim to test four hypotheses:

- *Angle effect*: VLMs find it harder to follow the gaze in images taken from a side view of the person than a front view.
- *Proximity effect*: As referent candidates get closer, VLMs find it harder to infer gaze direction.
- *Choice effect*: As the number of objects increases (from 2 to 4), the performance drops at a rate greater than the baseline of choosing answers randomly from choices (which drops from 50% to 25%).
- *Bias*: VLMs exhibit biases such as a preference for a particular viewing angle (left vs. right), specific referent objects, or individual actors.

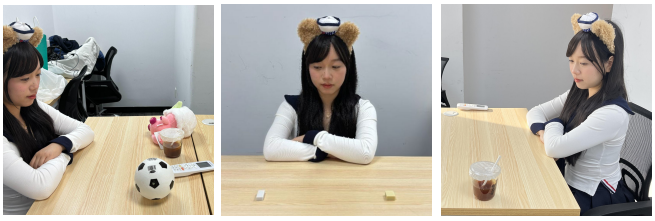

Figure 1: Three examples from stimulus set with different **Angle**. The task is to choose the object the actor is looking at among the given options.

Experiment Settings

Materials

The stimulus pool has 900 stimuli. Each stimulus is a photo of an actor seated at a clean table with 2 to 4 objects ("referent candidates") placed on it, as in Fig. 1. The following variables are controlled:

- **Candidates**: The specific set of objects on the table as referent candidates, among 18 combinations of 9 objects.
- **Actor**: 2 actors, either actor X or actor Y.
- **Angle**: 3 camera angles, two showing the actor's left and right profile, and one facing the actor front-on.
- **Proximity**: On a scale from 1 to 3, where 3 represents the highest proximity (i.e., smallest distance) between **Candidates**.
- **#Candidates**: The number of **Candidates** on the table, from 2 to 4.

VLM Representative Screening

To find the top-tier VLMs for larger-scale evaluation and finer-grain analysis, we first evaluated 111 VLMs by presenting each VLM with every stimulus once. Among the most performant ones, four are chosen for further analysis: GPT-4o-2024-08-06 (OpenAI, 2024), Gemini 1.5 Pro 002 (Gemini et al., 2024), Qwen2.5-VL-72B-Instruct (Qwen et al., 2025), and GLM-4V-9B (GLM et al., 2024).

Method

We repeat each test stimulus 11 times and construct the prompt based on a uniformly sampled template from a pool of 12 templates, allowing the same stimulus-prompt pair to appear multiple times, as the same VLM could produce varying responses. The prompt includes a multiple-choice question with choices being **Candidates** in a randomized order. Each of the four representative VLMs is presented with this set of evaluation cases, resulting in a total of $4 \times 900 \times 11$ trials. Using the reasoning-model-friendly pipeline developed in Duan et al. (2024), responses are first matched to options (A, B, C, or D) using manually defined templates, followed by a finer matching using a language model if template

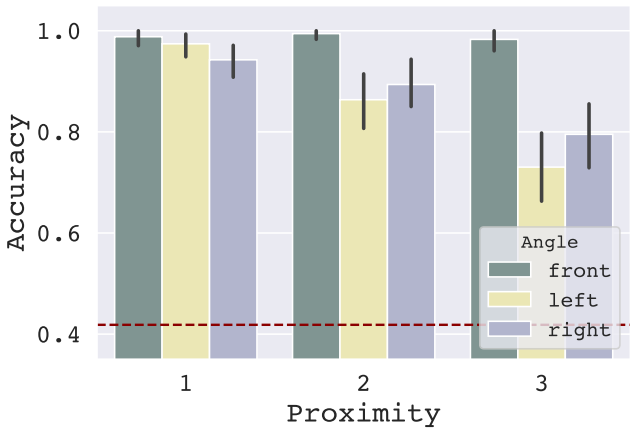

Figure 2: Summary of accuracy vs. proximity between referent candidates on the table and the perspective angle, aggregated across 4 chosen VLMs. The dark red dashed line is the expected accuracy of 0.42 if choosing uniform-randomly from the choices. 95% confidence intervals are depicted in black.

matching fails. Trials with unresolvable outputs are manually reviewed and excluded if they cannot be categorized. The evaluation procedure used in representative screening is almost the same, except that each stimulus is only presented to a VLM once.

Analysis

With all the evaluation results collected, we performed the one-portion z-test and found that only 20 VLMs performed significantly better than chance ($\mathbb{E}(\text{Accuracy}) = 42\%, \alpha = .05$). All four representative VLMs (GPT-4o, Gemini, Qwen, and GLM) perform well above chance (all $z > 9$, $p < 10^{-16}$) but still below an accuracy of 55%. Their aggregated results are shown in Fig. 2. Preliminary trends suggest the *Proximity effect* and *Angle effect* (especially in high-proximity trials), motivating fine-grained analyses.

We mean-centered **Proximity** and **#Candidates** and fitted separate mixed-effect logistic regression models for each representative VLM ($n = 9854, 9898, 9900, 9900$ trials, respectively, after the drop of invalid trials), as the statistical model with the best fit may not be the same. We use **StimulusID** to denote the index of stimulus from 0 to 899, **PromptID** for the index of prompt template from 0 to 11, and **Accuracy** for the binary correctness of a trial.

Gemini The model for Gemini is $\text{Accuracy} \sim \text{Angle} + \text{Proximity} + \text{\#Candidates} + (1|\text{StimulusID}) + (1|\text{PromptID})$. This model includes the maximal random effect structure among those that successfully converge. ANOVA comparison shows that removing either random intercept significantly worsens the fit ($p = 2.2 \times 10^{-16}$ and $p = 9 \times 10^{-3}$, respectively). Although adding potential random effects like **Actor** and **Candidates** results in divergence, we observed slight performance variation across **PromptID**, with a small variance of $.024$. This contrasts with Gupta et al. (2024)'s previous finding that VLMs' gaze inference performance depends largely on the prompt.

As expected, the model reveals a *Proximity effect* ($p = .0002$). In contrast, the *Angle effect* is not significant: performance does not differ when the actor is viewed from the left or right compared to front-facing ($p = .22$ and $.46$, respectively). This pattern holds even when the model is restricted to trials with higher proximity levels (**Proximity** $= 2$ or 3). We also observe a *Choice effect*: as **#Candidates** increases, performance not only declines ($p < 10^{-10}$) but also declines at a steeper rate ($z = 4.5, p < .001$) than what it would have been if a baseline machine chooses randomly from the options, derived from the slope to **#Candidates** in an analytic logistic regression of expected accuracy.

GPT-4o, Qwen, and GLM The model for Gemini is also the most appropriate model for Qwen. In contrast, the best random effect structure for GPT-4o and GLM includes both **StimulusID** and **Actor** as random intercepts, indicating performance varies with the specific gazer. Notably, Qwen does not exhibit a significant *Proximity effect* ($p = .16$), while both GPT-4o and GLM do ($p < .001$ and $p = .004$, respectively). All three models show a robust *Choice effect* ($p < .002$ for all), and none show a *Angle effect* ($p > .18$ for all).

Discussion and Conclusion

We found that 91 out of 111 VLMs failed to perform better than chance at identifying the object a person is looking at in an image, even when the options are named. We thus focused our analysis on four VLMs that outperformed chance. By systematically manipulating variables in the evaluation stimuli, we observed that changes in camera angle from front to side view did not affect performance. However, most VLMs showed clear declines in accuracy as the proximity between candidate objects or the number of choices increased. Contrary to expectations of brittle behavior, performance varied only slightly across prompts and gazers.

These characteristics of their behavior match with what an imperfect but meaningful gaze inference algorithm might show, suggesting that the emergent computation VLMs perform is moving toward meaningful inference, as opposed to just "stochastic parrots" or approximate retrieval (Kambhampati, 2024), calling for mechanistic investigations into their underlying inference.

Acknowledgement

We thank Zillion Network Inc. for providing the computation resources used in this work. Their optimized peak/off-peak scheduling, high-throughput storage infrastructure, and automated environment management enabled the cost-efficient and reliable execution of our experiments.

References

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148-153. doi: https://doi.org/10.1016/j.tics.2009.01.005

Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., ... others (2024). Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd acm international conference on multimedia* (pp. 11198–11201).

Gemini, T., et al. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.* Retrieved from https://arxiv.org/abs/2403.05530

GLM, T., et al. (2024). *Chatglm: A family of large language models from glm-130b to glm-4 all tools.*

Gupta, A., Vuillecard, P., Farkhondeh, A., & Odobez, J.-M. (2024). Exploring the zero-shot capabilities of vision-language models for improving gaze following. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 615–624).

Kambhampati, S. (2024, March). Can large language models reason and plan? *Annals of the New York Academy of Sciences*, *1534*(1), 15–18. doi: 10.1111/nyas.15125

OpenAI. (2024). *Gpt-4o system card.* Retrieved from https://openai.com/index/gpt-4o-system-card/

Qwen, T., et al. (2025, January). *Qwen2.5-vl.* Retrieved from https://qwenlm.github.io/blog/qwen2.5-vl/