Can Vision Language Models Infer Human Gaze Direction? A Controlled Study

Zory Zhang[†] Brown University
Pinyuan Feng[†] Columbia University
Bingyang Wang Emory University
Tianwei Zhao Johns Hopkins University
Suyang Yu University of Washington
Qingying Gao Johns Hopkins University
Hokin Deng[*] Carnegie Mellon University
Ziqiao Ma[*] University of Michigan
Yijiang Li[*] University of California, San Diego
Dezhi Luo[*] University of Michigan
GrowAI Team growing-ai-like-a-child.github.io **Project Page** | **Data** | **Code**

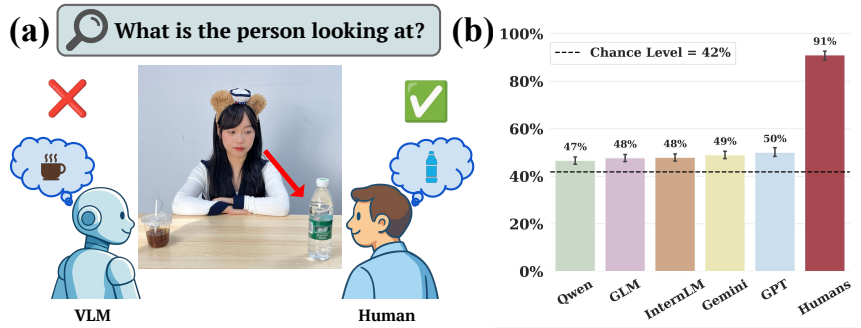


Figure 1: (a) The gaze referential inference task. (b) 99.9% confidence intervals of the accuracy means are depicted. A random-guessing machine achieves an accuracy of around 42%. A performance gap exists between top-tier Vision Language Models and humans.

Abstract

Gaze-referential inference—the ability to infer what others are looking at—is a critical component of a theory of mind that underpins natural human-AI interaction. In a controlled study, we evaluated this skill across 111 Vision Language Models (VLMs) using photos taken with manipulated difficulty and variability, comparing performance with that of human participants (N = 65), and analyzed behaviors using mixed-effects models. We found that 94 of the 111 VLMs failed to do better than random guessing, while humans achieved near-ceiling accuracy. VLMs even respond with each choice almost equally frequently. Are they randomly guessing? Although most VLMs struggle, when we zoom in on five of the top-tier VLMs with above-chance performance, we find that their performance declined with increasing task difficulty but varied only slightly across different prompts and scene objects. These behavioral patterns cannot be explained by considering them as random guessers. Instead, they likely use a combination of heuristics and guessing such that their performance is subject to the task difficulty but robust to perceptual variations. This suggests that VLMs, lacking gaze inference capability, have yet to become technologies that can naturally interact with humans, but the potential remains.

[†]Co-leading. [*]Co-advising. Correspondence to zory@brown.edu

1 Introduction

Theory of Mind (ToM) is our ability to attribute mental states (e.g., intention, desire, belief, and goals) to ourselves and others (Premack and Woodruff, 1978). It underpins our capacity to understand others, engage in natural interactions, and collaborate effectively. As recent advances in language modeling are being rapidly integrated into every aspect of human society, their ability to interact meaningfully with humans depends on possessing a machine form of ToM (Ma et al., 2023). While early exploration in text-based technologies has yielded intriguing findings (e.g., Kosinski 2024), recent advances in generative Vision-Language Models (VLMs, e.g., Gemini et al. 2023; Liu et al. 2023), which process both visual and textual inputs, offer promising hosts for a more capable ToM with natural interactions. A key challenge today is uncovering the extent to which VLMs have a ToM, and if not much, to determine whether they are on the right track of improving ToM as scaling continues. However, most evaluations are not adequate in furthering understandings of the inference behind the scores (e.g., Kosinski 2024, but see Pi et al. 2025), and VLMs' use of superficial shortcuts is hard to control without access to their training data, echoing Sap et al. (2022).

Interpreting gaze, a skill humans are uncannily good at, can illuminate this challenge, as it serves as one of the basic elements that more complex ToM builds on. Gaze is a window into other people's unconscious knowledge: their visual focus (Posner and Petersen, 1990), conversational referents (Prasov and Chai, 2008; Qian et al., 2023), linguistic knowledge (Golinkoff et al., 1987), intention to speak during conversation turn takings (Kendrick et al., 2023), intended motor actions (Land, 2006), and immediate intentions and desire in general (Mikulincer et al., 2014). Gaze is also one of the starting points where infants begin to generalize their experience (with open and closed eyes) to other social agents, which allows them to begin understanding others who are "like them" (Mikulincer et al., 2014). All of these are critical ingredients of ToM. More broadly, gaze inference likely bootstraps our later social learning, cognitive development (Csibra and Gergely, 2009), and possibly language acquisition (Brooks and Meltzoff 2005, but also see Sander et al. 2024), all of which furthers our grasp of ToM. Considering that gaze inference constitutes the content and helps the development of a ToM, the pressing question is whether VLMs can pick up gaze cues like humans do. If they cannot, they are equipped with a weaker ToM that will leave them a harder time understanding humans and interacting and collaborating with humans. Investigating this will help build a more nuanced understanding of machine ToM in VLMs.

Instead of building a general benchmark of gaze understanding, we aim to characterize VLMs' behavioral patterns through a controlled study to constrain hypotheses of their underlying inference. We formulate the basic task as inferring the object that falls along gaze direction ("line of sight," as shown in Fig. 1), one that more sophisticated gaze understanding builds on. We controlled variables that might affect the task difficulty and potential sources of variability in performance. We took 900 photos of a scene with an actor (as the gazer) by a table and objects on the table, covering multiple photo-shooting views, gazers, proximities *between* objects, and the combinations of objects with different visual contrast in the scene. See Fig. 2 for examples of the pictures. Then, we asked VLMs and humans (N=65) to choose which object the person in the photo is looking at from a set of options and inferred behavioral patterns from their performance.

Out of 111 VLMs, 94 performed about as well as if they had guessed randomly without looking at the images. Even worse, they responded with every possible option almost equally often, which is consistent with the account that they chose randomly from the given options. On the other hand, human participants showed near-ceiling accuracy, as depicted in Fig. 1. Still, as we took a closer look at five chosen top-level VLMs, we found patterns that nevertheless cannot be explained by labeling them as mere "random guessers" or "approximate retrievers" (Kambhampati, 2024), which will predict either a chance-level performance, or their performance only varies with the superficial familiarity of the context relative to their "database." Instead, they performed well above chance (all $p < .002$), and their performance declined with increasing task difficulty but varied only slightly with the specific prompt phrasing and the specific objects present in the scene. These findings suggest that the underlying inference of the top-tier models is imperfect but approaches the gaze inference task in a way that is affected by the difficulty (rather than familiarity), and the inference is robust to perceptual variations, such as object visual features, sizes, and background distractors. We hypothesized that VLMs may rely solely on head direction without eye direction to perform this task—an observation that opens avenues for further investigation.

2 Related Work

Gaze Estimation and Following It is helpful to distinguish gaze direction estimation and gaze following. Estimation involves inferring the direction of gaze, while gaze following is the phenomenon that infants look where someone else is looking (Mikulincer et al., 2014). This work is concerned with gaze estimation, as gaze following builds on estimation and additionally requires registering gaze as communicative (introducing a third party; Grossmann et al. 2008; Senju and Csibra 2008) and referential (object-directed). 2- to 5-day-old newborns can detect gaze direction (or motion direction) in the sense of anticipating object appearance on the left or right side of a face cued by eye movements (Farroni et al., 2004). On the other hand, gaze understanding (the true subject of interest behind gaze following) is not mature at birth: infants only begin to understand that eye gaze reflects visual experience around ten months old (Brooks and Meltzoff, 2005) and know head turns and eye movements are object-directed after fourteen months old (Caron et al., 2002). Therefore, if VLMs can estimate gaze direction, a further question will be what roles do VLMs attach to gaze (e.g., the referential nature of gaze, using gaze to resolve conversational referent ambiguity).

Gaze Estimation in Computer Vision and VLMs There is evidence that explicitly teaching VLMs to infer gaze referents can lead to more effective natural language and motor interactions between robots and humans (e.g., Qian et al. 2023; Prasov and Chai 2008), suggesting the importance of gaze estimation for human-centered technologies. Computer vision methods specialized for gaze estimation might achieve performance comparable to humans' (Han et al., 2021). However, this result is based on an evaluation that mainly requires head direction inference but less about eye direction, which turns out to be an important factor, as revealed in our Analysis B in Section 5. Furthermore, these methods require specialized designs for gaze, but it is unclear how they can be seamlessly integrated into the general-purpose VLMs that are more widely applied. Bridging the gap from specialized methods to VLMs, Gupta et al. (2024) evaluated VLMs on gaze estimation benchmarks. In this work, we conducted a controlled study to build on benchmark scores and characterize the underlying inference.

3 Experiment Settings

3.1 Task Formulation

As depicted in Fig. 1, we study gaze direction inference in a minimal setting. An image depicts a single gazer and a set of objects on the table in front of the gazer, which are the main focus of the image. We present the images with multiple-choice questions where the question is about which object the gazer looks at, and the choices are names of objects on the table in a randomized order. We essentially use referent inference as a proxy for direction inference. Such a simple setting allows us to reduce confounders (e.g., VLMs' attentional skills for cropping, object naming skills for open-ended categorization, and meta-cognitive skills for reporting direction numerically).

3.2 Stimuli Curation

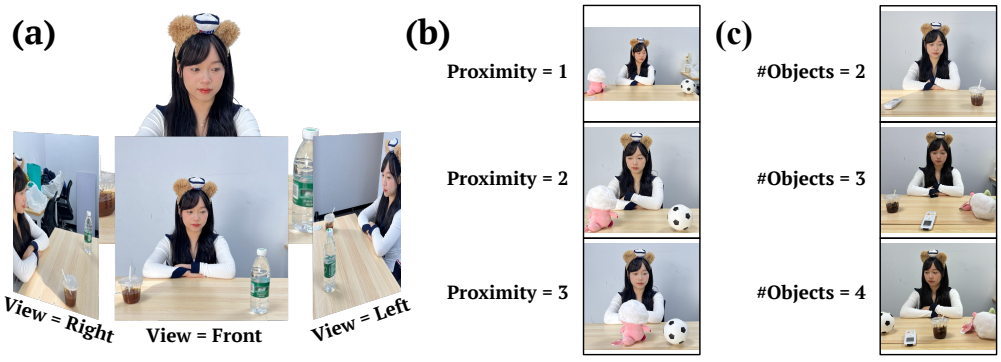


Figure 2: Systematic manipulation of `View` (left/right/front), `Proximity` (1-3 scale), `#Objects` (2-4), `Objects` (18 combinations of 9 distinct items), and `Gazer` (2 actors) across 900 test stimuli. Stimuli in subfigure (c) have a `Proximity` value of 2. Here `Gazer=ActorX`.

As illustrated in Fig. 2, when constructing the scenes for photo taking, we systematically controlled the following variables, which allow us to go beyond the benchmark scores we obtained in Section 4 and characterize behavioral patterns of VLMs and humans in Section 5:

- `Objects`: The specific combination of objects on the table, among 18 combinations of 9 distinct objects. See Appendix A.1 for all 9 objects and the 18 combinations of them.
- `#Objects`: The number of `Objects` on the table, from 2 to 4.
- `Gazer`: Either Actor X or Actor Y. The actor in Fig. 2 is X, and the one in Appendix A.1 is Y.
- `View`: 3 possible values. Two views show the gazer's left and right profile (`View=left/right`), and one shows a frontal view directly facing the gazer (`View=front`).
- `Proximity`: On a scale from 1 to 3, where 3 represents the highest relative proximity (i.e., smallest distance) between `Objects` (between each other, not to the gazer).

After stimulus cleansing (see Appendix A.1 for details), there are 900 test stimuli for both VLMs and humans, and additionally, 7 attention-check stimuli for the human study. [1]

3.3 VLM Evaluation Procedure

We follow the reasoning-model-friendly pipeline developed by Duan et al. (2024). We present each VLM with every stimulus once for Analysis A (in Section 4) and additionally ten times for Analysis B (in Section 5), meaning 900 and 9000 trials for each VLM, respectively. In each presentation of stimuli, we construct the prompt based on a uniformly sampled template with replacement from a pool of 12 templates (See Appendix A.2 for a complete list). The response decoding temperature parameters are set to the ones recommended by the corresponding VLM provider. Then, VLM responses are first matched to options (A, B, C, or D) using manually defined templates, followed by summarization and matching using Meta-Llama-3.1-70B-Instruct if template matching fails, and then followed by manual review if still not resolvable. Infrequently (5741 such trials among 900×111 in Analysis A plus 9000×5 trials in Analysis B), responses that are indeed not manually categorizable are counted as incorrect responses, since they are often nonsensical responses when the problem is too difficult for the VLMs.

3.4 Human Response Collection

We used Prolific to recruit 79 participants around the globe who are fluent in English and use Desktop browsers to access our survey (created using JsPsych; developed by de Leeuw et al. 2023). The test stimuli are split into 20 predetermined stimulus lists with 45 test stimuli per list. There are additionally 7 attention-check stimuli with `#Objects=2` and `Proximity=1` while covering a range of `Objects`, all three `Views`, and both `Gazers`. Each participant was assigned to one of the stimulus lists and received a mix of 45 test stimuli and 7 attention checks with random presentation order. 13 participants failed at least one attention check and were excluded. After that, one participant was also excluded due to their performance being far from the mean accuracy for more than three standard deviations, resulting in 65 valid participants for analysis (65×45 valid trials). More details are in Appendix A.3.

4 Analysis A: Can VLMs Infer Human Gaze Referent?

In this section, we aim to uncover the simplest question: to what extent can VLMs infer the referent of human gaze? Based on 900 trials per VLM from 111 VLMs and 45 trials per person from 65 participants, we saw a substantial performance gap between VLMs and humans and speculated that VLMs just randomly guessed the answers, but it turned out to be more complicated than that.

Performance Gap Fig. 1 depicts a comparison between human and top-tier VLM performance. The performance of all 111 VLMs is available at Appendix A.4. On one hand, human participants can successfully infer the gaze referent in ninety-one percent of the questions. On the other hand, the accuracies of most VLMs are very close to the expected accuracy of a machine that randomly selects a valid option in the multiple-choice question ($\mathbb{E}(\text{Accuracy}) \approx 42\%$). The five VLMs shown in Fig. 1 are top-level VLMs, and they did perform well above chance-level (all $p < .002$). Still, they can only solve about half of the questions, suggesting a substantial gap from human-level performance.

[1] The stimulus set is available for download at https://osf.io/kyaeu.

VLMs Might be Randomly Guessing. We performed a 2-sided test for proportions based on normal (i.e, a z-test) and found that 94 of the 111 VLMs fail to perform significantly better than chance ($\alpha = .05$). Fig. 3 shows a subset of confusion matrices comparing humans and VLMs across different combinations of `Objects` (more in Appendix A.5). A random guesser will produce a uniform confusion matrix, while the ones of VLMs are not substantially different than that. They did not achieve that by always outputting option A (See Appendix A.7).

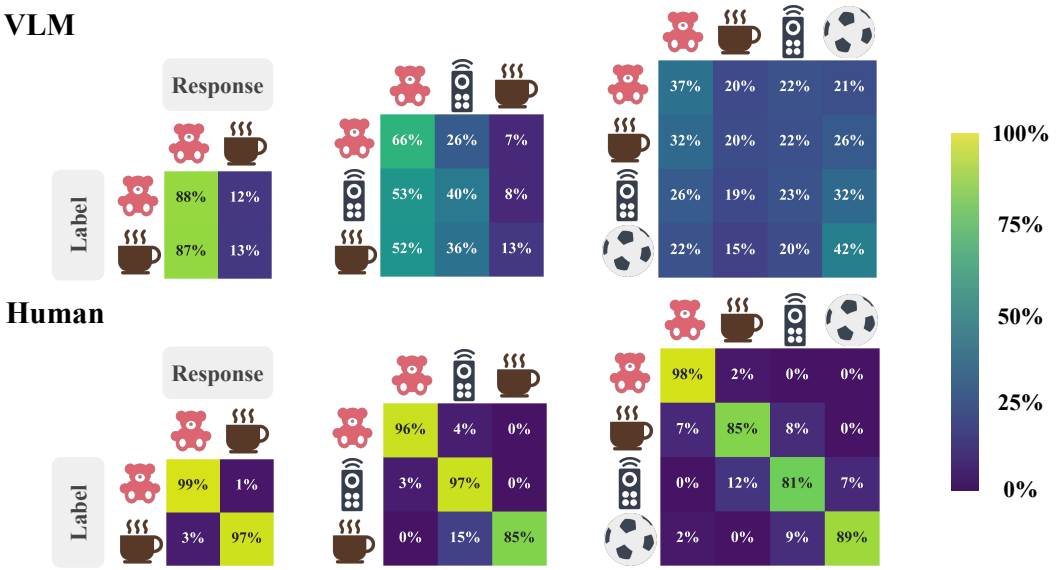


Figure 3: A row in a confusion matrix indicates the proportion of trials across all 111 VLMs (or human participants) that were responded with the *column object* (e.g., the coffee for *columns* with a coffee emoji) among trials in which the correct answer is the *row object* (e.g., the doll for *rows* with a doll emoji). Overall, humans occasionally choose the object adjacent to the correct one, which the gazer looks at, while VLMs show a combination of a slight tendency towards certain items (e.g., the doll) and near-uniform sampling across other options (alternatively speaking, probability-matching to their priors).

Scaling Does Not Help. We collected VLM release date information for 75 of the 111 VLMs (excluding smaller-size versions when the full-size versions are present) and size estimates for 106 VLMs (excluding outliers with estimated size larger than 100 billion parameters). They cannot linearly predict accuracy ($R^2 < 0.03, 0.01$ respectively), as shown in Fig. 4 and Fig. 5a, respectively, suggesting fundamental architectural limitations rather than scale issues.

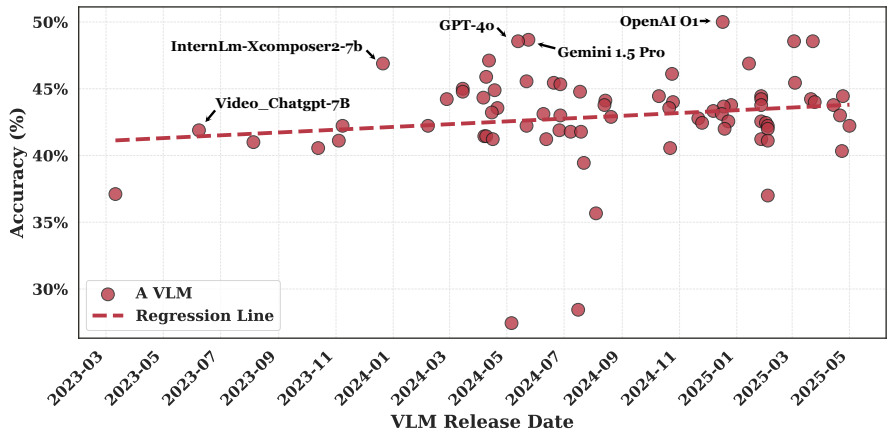


Figure 4: No strong linear relation between VLM accuracy and release date was found.

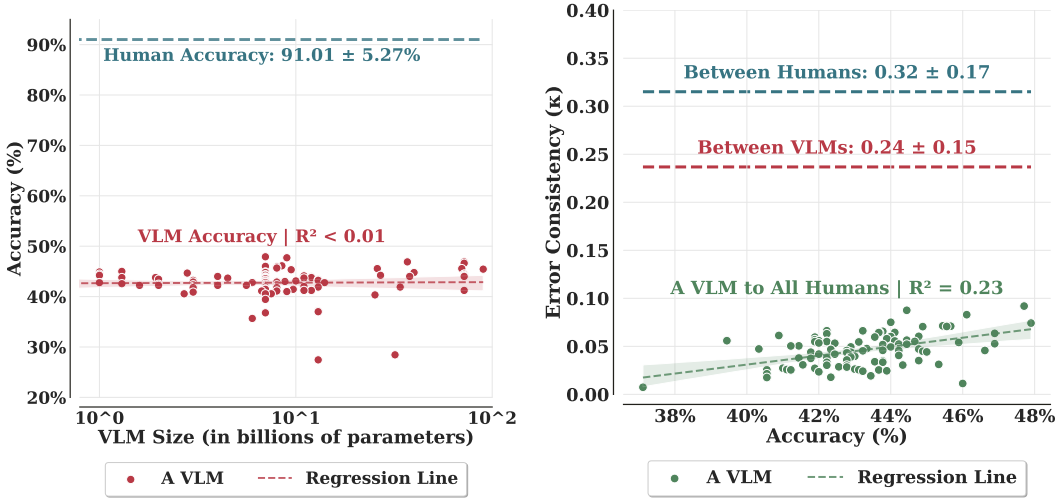
Substantially Different Error Patterns between VlMs and Humans. To go beyond accuracy scores, we evaluate agreement between humans and VLMs by comparing the mean accuracy of VLMs and humans for every stimulus in Appendix A.6.

Additionally, we also calculated a metric termed *error consistency* (analogous to Cohen's kappa κ), proposed by Geirhos et al. (2020) to evaluate the trial-by-trial agreement of error patterns between two decision-makers (i.e., how often they get the same thing right or get the same thing wrong). This metric adjusts for the potentially different accuracies of the two decision-makers, such that a value of 0 indicates the expected agreement solely by chance, and 1 indicates the maximal agreement.

We first established two baselines: To evaluate intra-human error consistency, for a stimulus list to which n participants were assigned (n ranging from 2 to 5), we enumerated all $\binom{n}{2}$ pairs of participants, calculated the error consistency for each pair, then average across first the pairs and then the 20 stimulus lists. Intra-VLM consistency is the average of consistency across all pairs of 111 VLMs, a simple calculation given that every VLM sees each stimulus exactly once.

To evaluate how well a VLM's error patterns align with all human participants, we compared the VLM with all participants one by one and calculated the metric on the 45 test stimuli that they both saw. We then investigate how such one-VLM-against-humans error consistencies linearly relate to the VLM's accuracy (shown in Fig. 5b; To remove outliers, 10 VLMs that are far from the mean accuracy by 1.96 standard deviations are not used for the linear regression) and size. The consistency does not become higher as VLMs get bigger ($R^2 < 0.1$) but does as VLMs achieve higher accuracy ($R^2 = 0.23, p < 10^{-6}$). Overall, there is a clear disalignment between VLMs and humans in terms of when they make errors, indicating that their computations likely approach the problem in qualitatively different ways.

This result contrasts with a previous finding by (Han et al., 2021) on a comparison between humans and convolutional networks specialized for gaze inference, where high alignment was found (not exactly error consistency though). Two explanations are available. One is that gaze-specialized methods align with human gaze inference better than VLMs do. Another explanation, which we preferred, is that the two reports measure different things. The previous reported alignment on their "in-the-wild" stimulus set requires very little inference of eye direction, but mainly head direction. The current stimulus set, on the other hand, requires eye direction inference (evident given the presence of *View effect* found in human responses in Analysis B). Indeed, as revealed in Analysis B, VLMs likely rely mainly on head direction but not eye direction, which can also explain this difference.



(a) The higher the accuracy, the better. Accuracy does not improve as the number of parameters increases.

(b) The higher error consistency, the more agreement of trial-by-trial error patterns between two decision-makers.

Figure 5: The 95% confidence intervals for linear regression are drawn as shaded areas. Standard deviations are reported for variables drawn as horizontal lines.

5 Analysis B: Behavioral Patterns in Five Top-tier VLMs and Humans

In this section, we aim to uncover the *kind* of gaze referent understanding that top-tier VLMs have by testing the following four hypotheses in VLMs and humans:

- *View effect*: The accuracy is closer to the chance level in images taken from a side view of the gazer than from a front view.
- *Proximity effect*: As referent candidates get closer, the accuracy becomes closer to the chance level.
- *Choice effect*: As `#Objects` increases (from 2 to 4, which means chance level drops from 50% to 25%), the accuracy becomes closer to the chance level.
- *Sensitivity*: VLMs are not robust in the sense that we will observe performance further from the chance level for a particular viewing angle (left vs. right), some specific `Object` combinations over others, or some specific prompt templates.

The measure of how close the accuracy is to the chance level is the ratio between their odds (i.e., the additive difference between their log-odds). This definition naturally fits in the use of logistic regression. As these are testable hypotheses, we pre-registered our study. [2]

Difference from Analysis A Analysis B features a focused group of top-tier VLMs, a larger sample size, slightly adjusted prompt templates, and mixed-effects modeling. Given that most VLMs perform close to chance, it is only meaningful to analyze the most performant ones closely. The five selected VLMs are: 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), InternLM-XComposer2-vl-7b Dong et al. (2024), and GLM-4V-9B (GLM et al., 2024). They are five of the top seven most-performant VLMs in Analysis A, while InternLM-XComposer2d5-7b and OpenAI-o1 were not selected for redundancy and cost considerations, respectively. For a comparison of prompt templates used for Analysis A and B, see Appendix A.2. Instead of repeating each stimulus once for each VLM as in Analysis A, we repeat each ten times. This leads to 900 $\times$ 10 trials per VLM from 5 selected VLMs (Analysis A VLM responses were excluded from Analysis B) and 45 trials per person from 65 participants. As sanity checks, we visualized all VLM and human data points we get in Appendix A.8 and Appendix A.9, respectively, as well as the human performance by condition for every participant in Appendix A.10. These initial explorations are consistent with our hypotheses.

Additional Variables In addition to the stimulus variables in Subsection 3.2, there are also:

- `Accuracy` (Binary): The outcome variable. Either a correct or incorrect inference of the gaze referent.
- `StimulusID` (Categorical): The index of the test stimulus. 900 unique numbers in total.
- `PromptID` (Categorical; VLM-only): The identifier of the prompt template, from 0 to 11. See Appendix A.2 for a list of templates.
- `ParticipantID` (Categorical; Human-only): The identifier of the human participant.

Mixed-Effects Modeling As Robitzsch (2020) suggested, we treated `Proximity` and `#Objects` as continuous variables due to the underlying scale they represent and mean-centered them before fitting the models. The following model skeleton is used:

$$\begin{aligned}\text{logit}([\text{Accuracy}]) \sim \text{logit}(&\frac{1}{\#\text{Objects}}) \\ &+\text{View} + \text{Proximity}_{\text{centered}} + \#\text{Objects}_{\text{centered}} \\ &+\text{Actor} + (1|\text{Random effects})\end{aligned}$$

[2] The experiment design, sample size, and analysis plan were pre-registered at Open Science Framework: https://osf.io/xdmr9 after the collection of VLM responses in Analysis A but before the collection of VLM responses in Analysis B and all human responses. Therefore, VLM responses collected for Analysis A are not used in this section for Analysis B. The number of human participants is 65 rather than the 80 indicated in the pre-registration, because large effect sizes were observed already. Although we pre-registered that four VLMs would be used, any four of the five VLMs will produce the same results. Due to these small deviations, we conducted post-hoc power tests and discussed the implications.

where $[\cdot]$ is the Iverson bracket and the log-odds function $\text{logit} := p \mapsto \ln(p/1-p)$ is the inverse of the logistic function $\sigma(x) = 1/\left[1+\exp(-x)\right]$. The second logit term effectively offsets the first term such that the fixed and random effects need to account for the variance of the log-odds of the ratio between the actual accuracy and the baseline expected accuracy. This skeleton is used because more complex models cannot converge. See Appendix A.12 for details of modeling.

Model Selection Possible random effects are: `Gazer:Objects` (nested under the gazer because a combination of objects corresponds to one of the gazer), `StimulusID`, `PromptID`, and `ParticipantID`. They are potential random effects because they are meant to be samples from a larger population (e.g., the 9 distinct objects are just samples of all possible objects). We explored the random effect structures individually for each VLM and all humans. For Gemini and Qwen, the eventual random effects are `StimulusID` and `PromptID`. For GLM, GPT, and InternLM, the random effect is `StimulusID`. For humans, the random effects are `StimulusID` and `ParticipantID`. Statistical models with other random effect structures either cannot converge, or do not yield a significantly better fit to the data under the parsimony principle according to ANOVA tests ($\alpha = .05$). A reproducible model selection procedure can be found at https://github.com/grow-ai-like-a-child/referential-gaze.

Main Effects We found strong *Proximity effect* and *Choice effect* in both VLMs and humans. Surprisingly, no *View effect* is observed in VLMs while the effect is strong in humans. Table 1 and Fig. 6 depict an intuitive summary of effects found, while the full statistical model output and trends in the log-odds space (where effects are detected) are in Appendix A.12. Had we not included `StimulusID` as a random effect to account for correlations between trials using the same stimulus and simply run a logistic regression, we would have found all effects significant. This highlights the importance of using mixed-effects modeling in repeated-measure experiments.

Table 1: Significance codes: "***" means $p < 0.001$, "**" means $p < 0.01$, and "*" means $p < 0.05$. "F", "L", and "R" short for `View=front, left`, and `right`, respectively. For insignificant effects, "true null" indicates that the effect is likely not true, while the "failed to find" ones are inconclusive due to insufficient sample size. Still, all the significant effects cannot be explained by a random-guessing account of VLMs' gaze referent inference.

Effects	**GPT**	**Gemini**	**GLM**	**InternLM**	**Qwen**	**Humans**
Proximity effect	***	***	**	*	True null	***
Choice effect	***	***	***	***	***	***
View effect (L. vs. F.)	Failed to find					***
View effect (R. vs. F.)	True null	Failed to find				***
Sensitivity to `Objects`	None					
Sensitivity to `Prompt`	None	Slight	None	None	Slight	N/A

Post-Hoc Power Test of the Sample We conducted Two One-Sided T-tests (TOST) to distinguish insignificant effects that are truly null effects and those that we failed to find due to the insufficient sample size. This is driven by some subtle deviations from the pre-registered plan. We used the eventual model of each group to calculate the Intraclass Correlation Coefficients (ICC) of `StimulusID` (the major source of random variance), which allowed us to approximate the effective sample size and Smallest Effect Size of Interest (SESOI). The results are reflected in Table 1, indicating that, except for the *View effect*, we have enough statistical power to conclude for other effects. Future works that investigate VLMs' use of eye direction cues, as hinted by our explanation of the lack of *View effect* in the discussion section, should use a larger sample size. The reproducible procedure is available along with the model selection code.

Sensitivity Analysis Including random effects allows us to capture the variability across random effects and reveals the *sensitivity* of VLMs' performance to random effects. The lack of the need to account for random variability across `PromptID` for GPT, for example, indicates GPT's performance is not sensitive to the prompt. For Gemini and Qwen, although taking the variability across `PromptID` into account leads to a better fit to the data, the variance of `PromptID` is less than 0.1 and therefore, they are practically not sensitive to the prompt as well. Similarly, they are not sensitive to the scene `Objects`, and the performance difference between `View=left` and `right` is not significant.

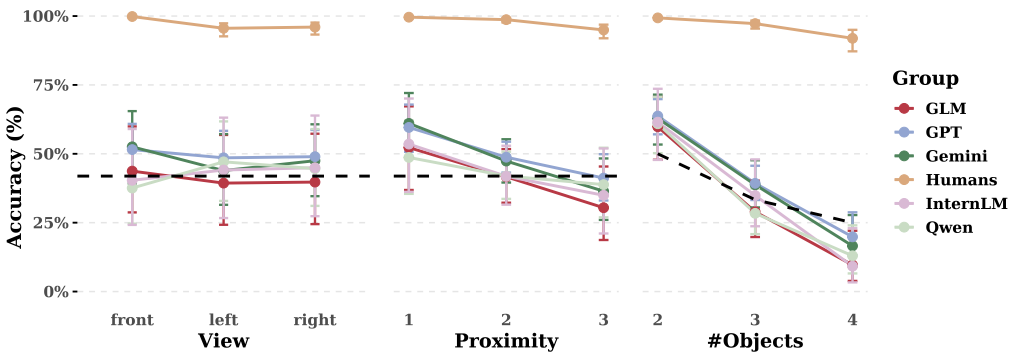


Figure 6: The estimated marginal means. The random-guessing baseline is indicated by dashed lines. Since different models fitted for different groups vary in their random effect structures, each model contributes to curves of the same color, and all 6 fitted models are used.

6 Discussion

We found that 94 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. In addition to the overall accuracy being close to chance level, they responded with every possible answer almost evenly frequently. Without controlled variables, we would have concluded that they are random guessers, not capable of gaze referent inference. Instead, we focused our analysis on five VLMs that far outperformed chance level and found interesting patterns: Contrary to expectations of brittle behavior (e.g., the last page in Gupta et al. 2024), top-tier VLMs' performance varied only slightly across irrelevant elements like the specific scene objects and prompts. This suggests that their gaze referent inference is shifting toward one that is robust to perceptual variations, such as object visual features. More interestingly, as task difficulty increases, their accuracy moves closer to the random-guessing level, which means their computation is *about* the image and the task. These findings cannot be explained by reducing them to random-guessers (with performance as well as who respond without looking at the image), or approximate retrievers (Kambhampati, 2024) (with performance depend on similarity to training data), both of which will predict almost constant performance as task difficulty changes, particularly given our stimuli that have controlled for superficial similarity to training data, in the sense that they only meaningfully differ if the observer is capable of gaze referential inference, and one that is incapable will perceive almost uniform similarity.

What will be the best explanation of these behavioral patterns? A good explanation needs to explain both (1) the positive finding of *Proximity effect* and *Choice effect*, and (2) the lack of *View effect* in VLMs that is strong for humans. The most probable one to us is that they use head direction without using eye direction jointly. This heuristic works better when the difficulty is low, but severely breaks down as the number of objects increases or when objects become closer to each other (but not much when the view is from the side rather than the front), leading to a performance that is closer to the chance level. Since humans use eye direction for inference, as the view changes from the front to the side, the angle becomes harder to approximate, making it more challenging to track the line of sight, and thus, the performance is impacted. VLMs do not use this cue in the first place and therefore, do not show *View effect*. This also explains why previous research found strong performance correlation between humans and early deep learning methods (Han et al., 2021) while we do not (between humans and VLMs). Our stimuli require heavy use of eye details for inferring gaze direction, while head directions are sufficient to solve most problems in many naturalistic evaluations of gaze inference. Still, these speculations are only our best explanation so far, calling for further investigations.

In summary, this work uses gaze as a proxy to demonstrate how controlled studies can help us investigate behavioral patterns behind benchmark scores (that would have marked them as random guessers) and reductionism ("approximate retrievers"), constrain hypotheses (they are unlikely to be just guessing), and provide new hypothesis (that they might primarily use head directions but not eye directions) that is subject to further mechanistic investigations.

Limitations This study, like most controlled studies, suffers from the limited quality, amount, and diversity of the evaluation cases. The photos are of size 448 by 448, a resolution that is noticeably lower than what the human vision system provides. This collection of photos does not and is not meant to comprehensively resemble the rich visual experience that people will perceive in their daily life: All photos are taken in the same office space; only two actors serve as the gazer; only 9 objects in total (See Appendix A.1). Still, the stimulus pool includes a range of gaze-irrelevant elements that increase the realisticness, including backgrounds with irrelevant objects, referent candidate objects with different visual salience, and facial decorations (for example, Actor X wore head decorations, and Actor Y wore false eyelashes and colored contact lenses).

Future Works This work raises more questions than it answers. If top-tier VLMs approximate a gaze referent inference heuristic (e.g., one that uses head direction but not gaze direction, given the absence of *View effect*), what is the mechanism of the approximation, and how does it emerge from the data? Given that gaze understanding likely bootstraps human cognitive development and language acquisition, in order to have AI that learns like humans and acquires Theory of Mind from naturalistic data, will a curriculum that encourages early development of gaze-referential understanding be helpful?

7 Acknowledgment

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.

For stimuli collection, our human actors granted us permissive, royalty-free rights to employ their visual materials. For human response collection, all methods were approved by the Johns Hopkins Institutional Review Board and were carried out per relevant guidelines and regulations. Both actors and participants received financial compensation. We thank them for their contribution.

The authors would also like to thank all anonymous reviewers for their valuable feedback.

References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48. (Back to section A.12)

Brooks, R. and Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science*, 8(6):535–543. (Back to sections: 1, 2)

Caron, A. J., Butler, S., and Brooks, R. (2002). Gaze following at 12 and 14 months: Do the eyes matter? *British Journal of Developmental Psychology*, 20(2):225–239. (Back to section 2)

Csibra, G. and Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, 13(4):148–153. (Back to section 1)

de Leeuw, J. R., Gilbert, R. A., and Luchterhandt, B. (2023). jspsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, 8(85):5351. (Back to section 3.4)

Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Wei, X., Zhang, S., Duan, H., Cao, M., Zhang, W., Li, Y., Yan, H., Gao, Y., Zhang, X., Li, W., Li, J., Chen, K., He, C., Zhang, X., Qiao, Y., Lin, D., and Wang, J. (2024). Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. (Back to section 5)

Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., Dong, X., Zang, Y., Zhang, P., Wang, J., et al. (2024). Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201. (Back to section 3.3)

Farroni, T., Massaccesi, S., Pividori, D., and Johnson, M. H. (2004). Gaze following in newborns. *Infancy*, 5(1):39–60. (Back to section 2)

Geirhos, R., Meding, K., and Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc. (Back to section 4)

Gemini, T. et al. (2023). Gemini: A family of highly capable multimodal models. (Back to section 1)

Gemini, T. et al. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. (Back to section 5)

GLM, T. et al. (2024). Chatglm: A family of large language models from glm-130b to glm-4 all tools. (Back to section 5)

Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., and Gordon, L. (1987). The eyes have it: lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*, 14(1):23–45. (Back to section 1)

Grossmann, T., Johnson, M. H., Lloyd-Fox, S., Blasi, A., Deligianni, F., Elwell, C., and Csibra, G. (2008). Early cortical specialization for face-to-face communication in human infants. *Proceedings of the Royal Society B: Biological Sciences*, 275(1653):2803–2811. (Back to section 2)

Gupta, A., Vuillecard, P., Farkhondeh, A., and 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*, pages 615–624. (Back to sections: 2, 6)

Han, N. X., Wang, W. Y., and Eckstein, M. P. (2021). Gaze perception in humans and cnn-based model. (Back to sections: 2, 4, and 6)

Kambhampati, S. (2024). Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 1534(1):15–18. (Back to sections: 1, 6)

Kendrick, K. H., Holler, J., and Levinson, S. C. (2023). Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378(1875):20210473. (Back to section 1)

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121. (Back to section 1)

Land, M. F. (2006). Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research*, 25(3):296–324. (Back to section 1)

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning. (Back to section 1)

Ma, Z., Sansom, J., Peng, R., and Chai, J. (2023). Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031. (Back to section 1)

Mikulincer, M., Shaver, P. R., Brooks, R., and Meltzoff, A. N. (2014). *Gaze following: A mechanism for building social connections between infants and adults.* American Psychological Association,, Washington, DC, first edition. edition. (Back to sections: 1, 2)

Muradoglu, M., Cimpian, J. R., and and, A. C. (2023). Mixed-effects models for cognitive development researchers. *Journal of Cognition and Development*, 24(3):307–340. (Back to section 4)

OpenAI (2024). Gpt-4o system card. (Back to section 5)

Pi, Z., Vadaparty, A., Bergen, B. K., and Jones, C. R. (2025). Dissecting the ullman variations with a scalpel: Why do llms fail at trivial alterations to the false belief task? (Back to section 1)

Posner, M. I. and Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, 13(Volume 13, 1990):25–42. (Back to section 1)

Prasov, Z. and Chai, J. Y. (2008). What's in a gaze? the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 20–29. (Back to sections: 1, 2)

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526. (Back to section 1)

Qian, K., Zhang, Z., Song, W., and Liao, J. (2023). Gvgnet: Gaze-directed visual grounding for learning under-specified object referring intention. *IEEE Robotics and Automation Letters*, 8(9):5990–5997. (Back to sections: 1, 2)

Qwen, T. et al. (2025). Qwen2.5-vl. (Back to section 5)

Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education*, Volume 5 - 2020. (Back to section 5)

Sander, J., Çetinçelik, M., Zhang, Y., Rowland, C. F., and Harmon, Z. (2024). Why does joint attention predict vocabulary acquisition? the answer depends on what coding scheme you use. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0). (Back to section 1)

Sap, M., Le Bras, R., Fried, D., and Choi, Y. (2022). Neural theory-of-mind? on the limits of social intelligence in large LMs. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. (Back to section 1)

Senju, A. and Csibra, G. (2008). Gaze following in human infants depends on communicative signals. *Current Biology*, 18(9):668–671. (Back to section 2)

A Appendix / supplemental material

A.1 Stimulus Cleansing Procedure, Examples, and Distribution

Fig. 7 shows more examples of actor Y. Fig. 8 illustrates all 9 objects in the stimulus pool. See Fig. 9 for a list of the 18 combinations of these objects and the distribution of stimuli. The stimulus distribution is not perfectly uniform due to the stimulus cleansing for quality control. Note that within every variable assignment, the number of stimuli with each ground truth is also roughly the same (i.e., correct answer counterbalanced).

Proximity rubrics For `Proximity`, a value of one corresponds to putting objects farthest away between each other on the table, while a value of three means they are placed the closest possible, but not touching. It represents the relative sparseness objects can be, rather than absolute distance.

Stimulus Cleansing Procedure We manually examined every photo we took and dropped all photos taken in the middle of a blink. All cases with occluded eyes can be approached by considering the head position, and there is no significant occlusion of parts of the objects. Black padding is added when resizing photos to 448 by 448 pixels.

Generalizability Note that sometimes the background is messy: this increases the credibility of our evaluation. Messy backgrounds mainly appear when `View=right`, yet we do not observe significantly worse performance in the `View=right` condition, so such a background introduces minimal confounding effect while adds realistic noise to the stimuli and increases diversity. We thus expect our results to have reasonable generalizability.

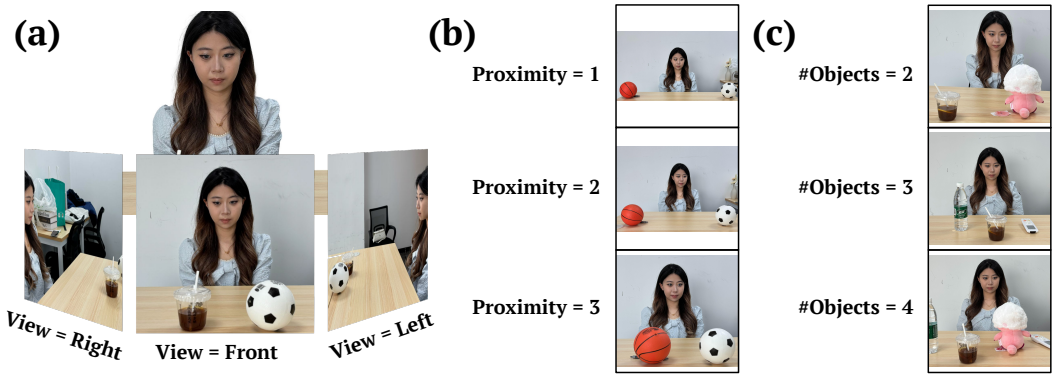


Figure 7: Examples of stimuli for Actor Y with different `View`, `Proximity` and `#Objects`.

Figure 8: All 9 objects in the stimulus pool with different sizes and visual salience.

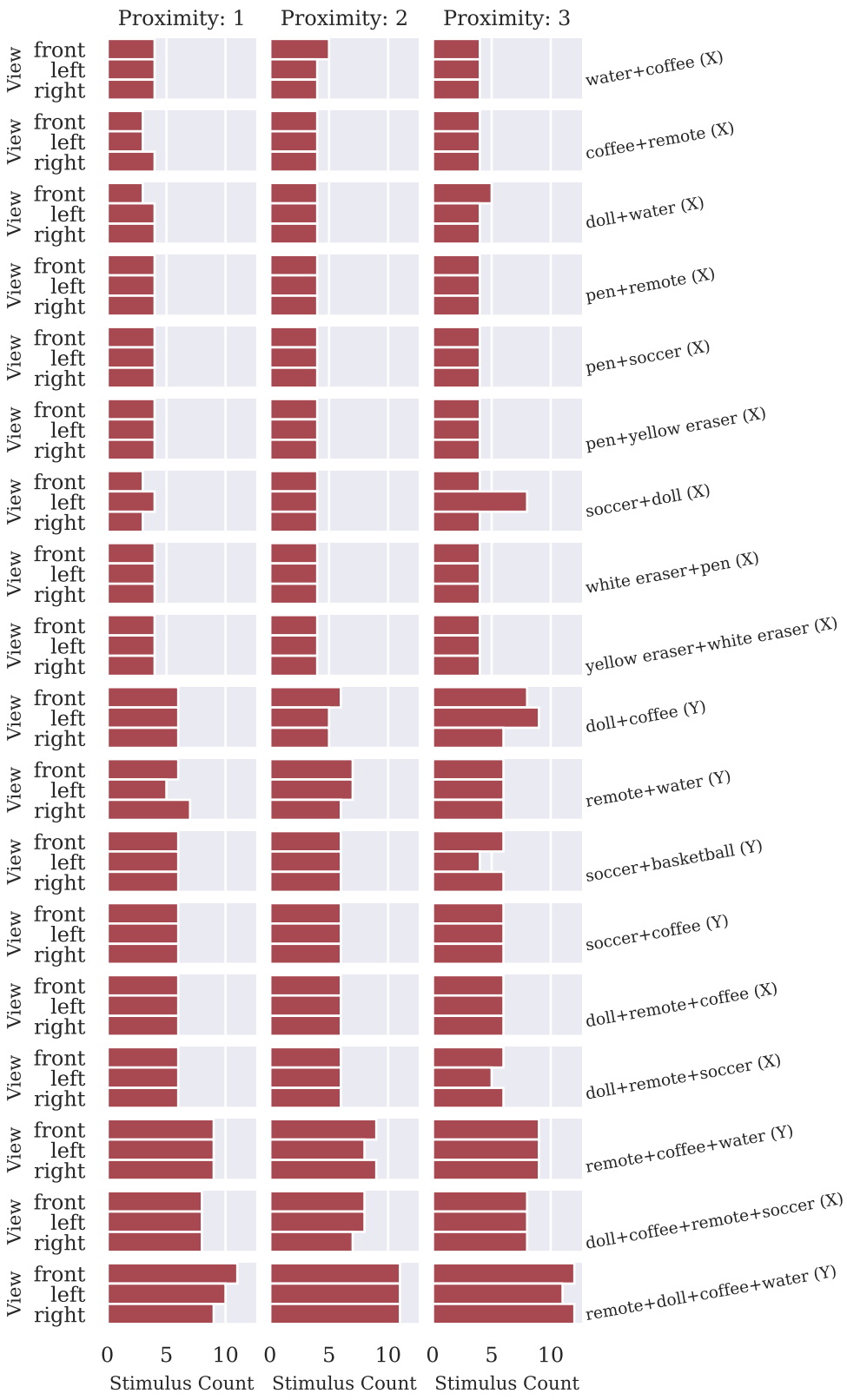


Figure 9: Number of stimuli nested within `View`, `Proximity`, and `Objects`. The brackets in row names denote whether the gazer is Actor X or Y for the combination of `Objects`.

A.2 VLM Prompt Variants

See Table. 2 and 3 for the list of prompt templates we used for Analysis A and B, respectively. The image tokens were inserted into the place marked by the placeholder in the template. `\n` indicates a change of line. The determined attitude and the guessing proposal in these templates are necessary to lower the refusal rate. Due to the unexpected close-to-chance performance we observed in Analysis A, we removed the wording of guessing in some prompts for Analysis B.

Table 2: The full list of 12 prompt templates in Analysis A, assuming that `#Objects=3`.

ID	Prompt Template
1	<image-placeholder> What is this person looking at?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.
2	<image-placeholder> Where is this person looking?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.
3	<image-placeholder> Which object is this person looking at?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.
4	<image-placeholder> What is she looking at?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.
5	<image-placeholder> Where is she looking?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.
6	<image-placeholder> Which object is she looking at?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.
7	<image-placeholder> What is this person looking at in the image?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.
8	<image-placeholder> Where is this person looking in the image?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.
9	<image-placeholder> Which object is this person looking at in the image?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.
10	<image-placeholder> What is she looking at in the image?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.
11	<image-placeholder> Where is she looking in the image?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.
12	<image-placeholder> Which object is she looking at in the image?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.

Table 3: The full list of 12 prompt templates in Analysis B, assuming that `#Objects=3`.

ID	Prompt Template
1	<image-placeholder> What is this person looking at?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so make your best guess.
2	<image-placeholder> Where is this person looking?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.
3	<image-placeholder> Which object is this person looking at?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. You cannot refuse to choose.
4	<image-placeholder> What is she looking at?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. There is no need to reason. If you don't know, you still must choose one, so make your best guess.
5	<image-placeholder> Where is she looking?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. There is no need to reason. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.
6	<image-placeholder> Which object is she looking at?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. There is no need to reason. You cannot refuse to choose.
7	<image-placeholder> What is this person looking at in the image?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so make your best guess.
8	<image-placeholder> Where is this person looking in the image?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.
9	<image-placeholder> Which object is this person looking at in the image?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. You cannot refuse to choose.
10	<image-placeholder> What is she looking at in the image?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. There is no need to reason. If you don't know, you still must choose one, so make your best guess.
11	<image-placeholder> Where is she looking in the image?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. There is no need to reason. If you don't know, you still must choose one, so you might select randomly. You cannot refuse to choose.
12	<image-placeholder> Which object is she looking at in the image?\n A. xxx\n B. xxx\n C. xxx\n Please answer with the option's letter A, B, C directly. There is no need to reason. You cannot refuse to choose.

A.3 Human Response Collection Details

Attention Checks The attention checks are the same across participants and are meant to be the simplest cases that a focused person should not fail. Indeed, within participants who are correct on all attention checks, the non-attention-check questions with `#Objects=2` and `Proximity=1` (meaning that they are as difficult as the attention checks) have a mean accuracy of around 99.30%.

Human Participant Demographics See Fig. 10 for the demographics of the 65 valid participants (those who passed all attention checks). Demographic data of some participants is missing and hence not plotted. All participants are paid an hourly rate of 12 USD regardless of their residence.

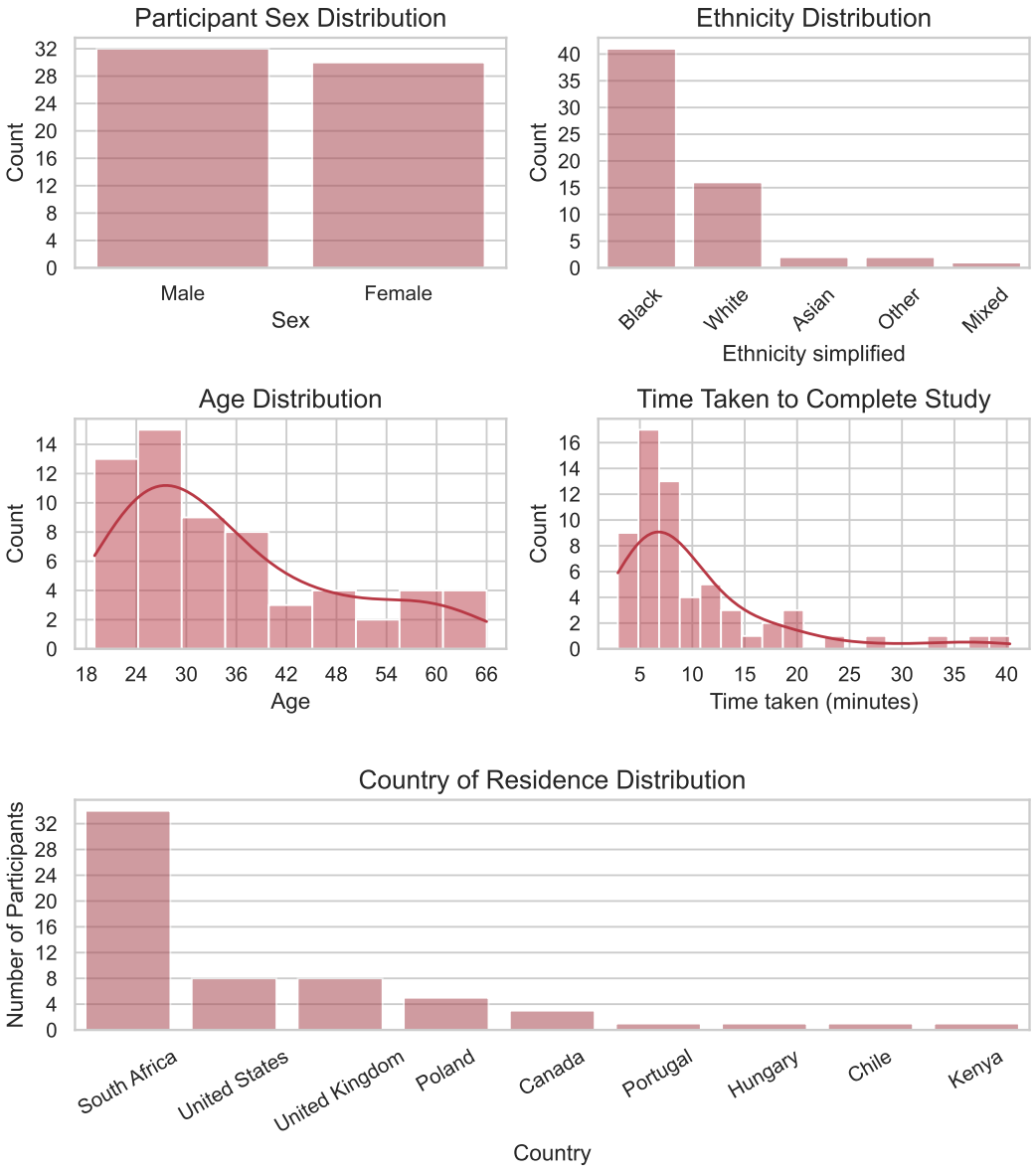


Figure 10: The demographics.

Human Accuracy vs. Response Time See Fig. 11. We submitted the response time details for every trials of valid participants (who pass all attention checks). The higher the response time of the trial is, the lower the likelihood of getting the correct response ($p < .001$). This indicates that people tend to spend more time on harder cases and still tend to fail on them.

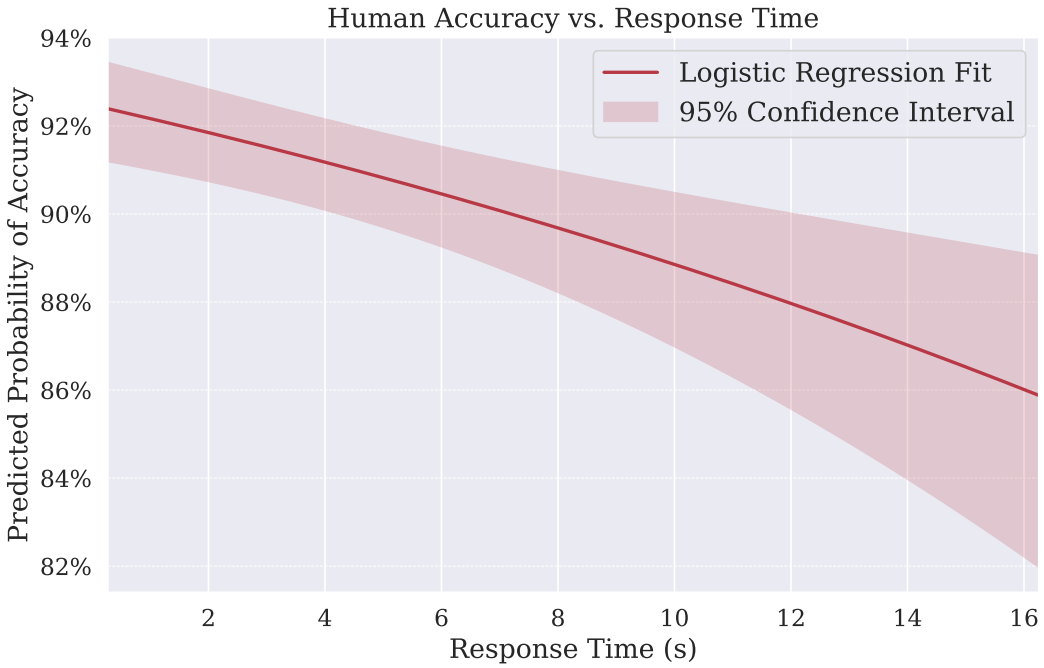
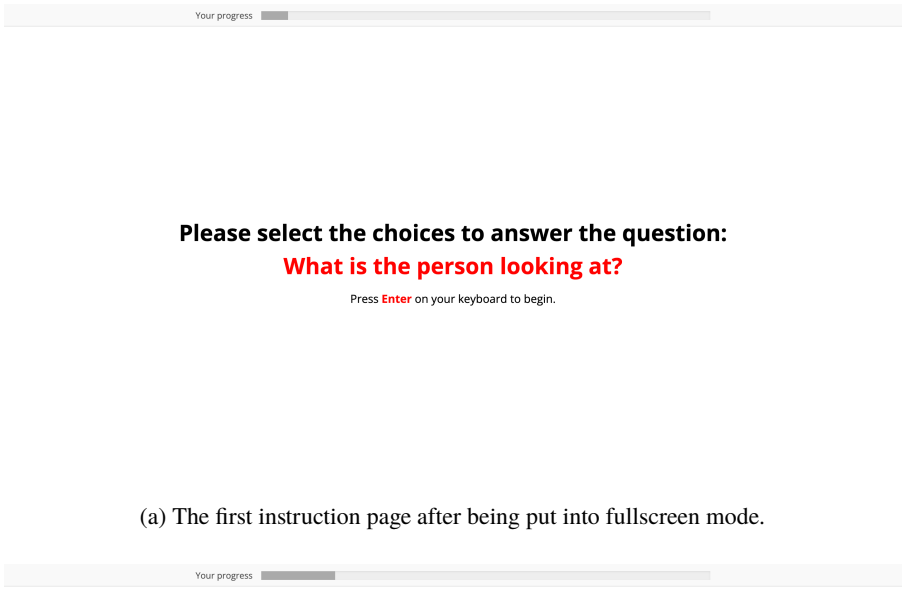
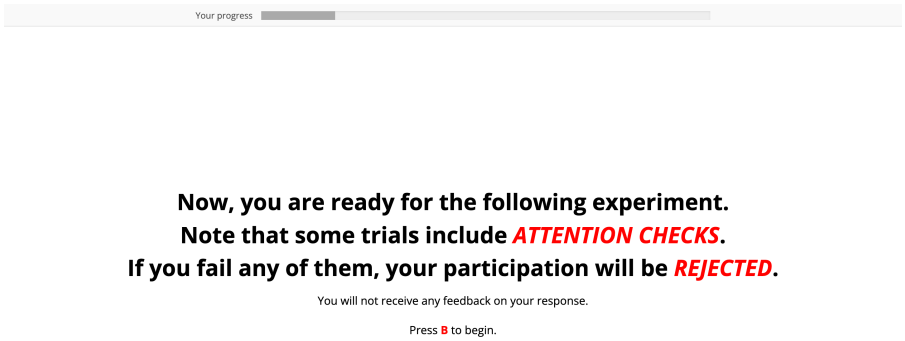


Figure 11: The prediction made by a logistic regression model (response time shown here as the 99th percentile range).

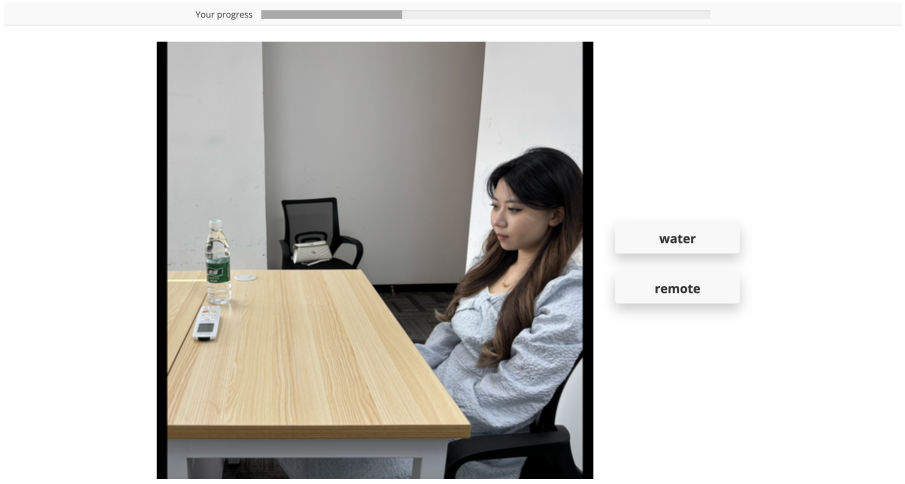
Details of the Survey See Fig. 12 for screenshots of the survey. The participants click a button to enter full-screen mode and start reading instructions (they have to press different keys on the keyboard to ensure they read the instructions). Then, they go through three practice trials based on 3 of the seven attention checks with correctness feedback (in the form of a check mark or a cross mark with no sound). After that, they complete the 45 (test stimuli) + 7 (attention check) questions with a progress bar to motivate them, as the questions are relatively easy, but no correctness feedback is available. The stimuli in the three practice trials will appear again as attentional checks that look the same as the other trials. Four other pre-determined attention checks will also appear. Participants do not know which are attention checks, and they are warned to try their best for all questions. Even if they failed attention checks, they were still paid, regardless of the warning. Participants cannot go back and forth to make changes, but they can take as long as they want to answer the questions.



(a) The first instruction page after being put into fullscreen mode.



(b) The second instruction page explains the existence of attention checks.



(c) An example of the question pages where participants click one of the buttons to make their choice and proceed to the next question.

Figure 12: Screenshots from the human survey.

A.4 Full Comparison of the Overall Accuracy

See Fig. 13. Note that responses collected in Analysis B (in Section 5) for the 5 selected VLMs are excluded to preserve equal sample sizes among VLMs for fair comparison.

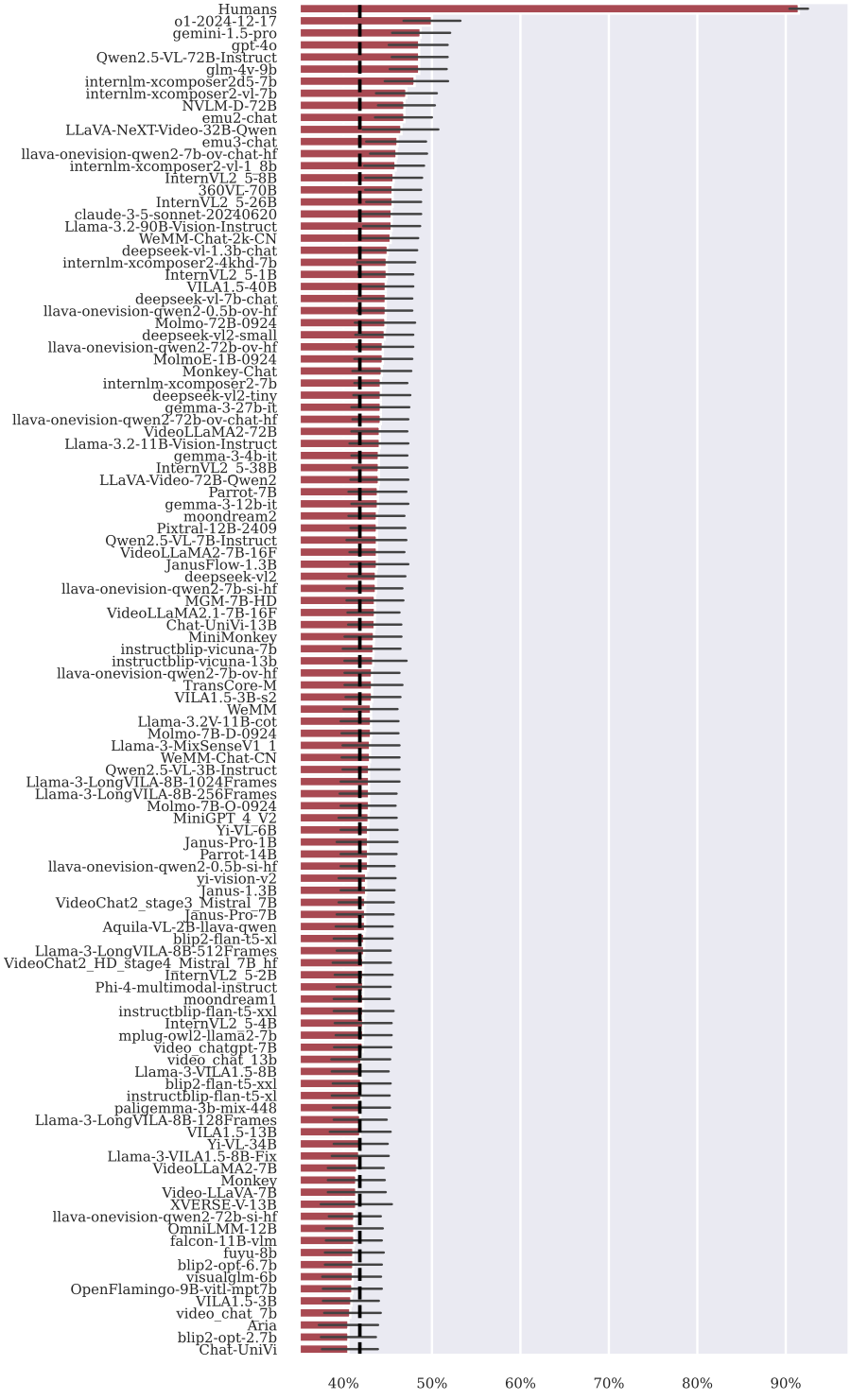
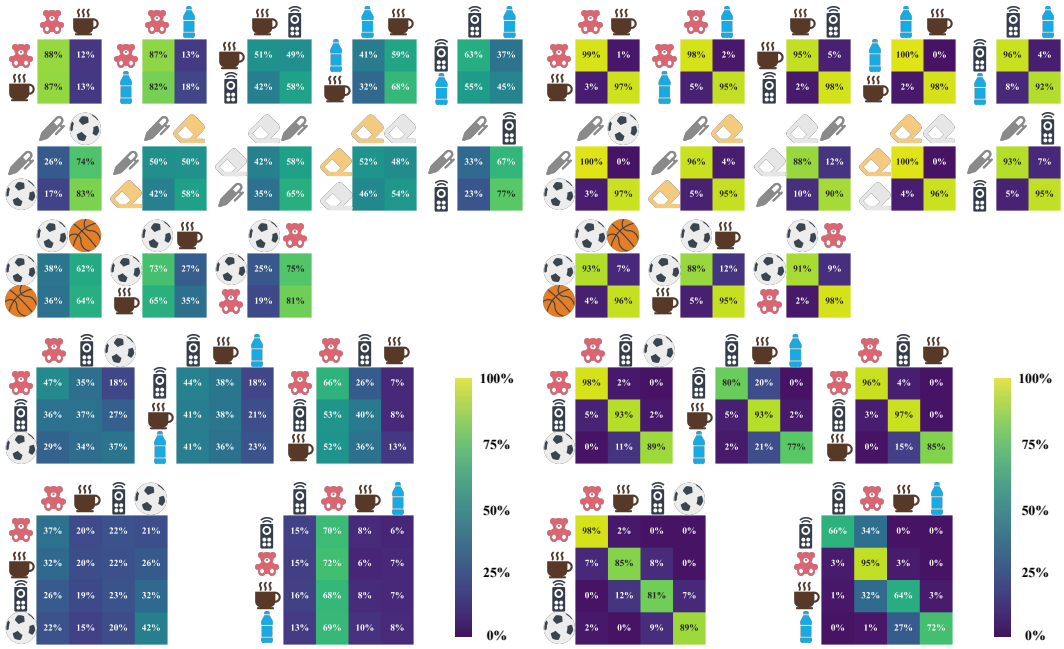


Figure 13: Full comparison of the overall accuracy of Humans and VLMs. 95% CIs are drawn horizontally, while the random guessing baseline of 42% is drawn vertically.

A.5 Confusion Matrices of VLMs and Humans

See Fig. 14 for the confusion matrices for each of the 18 combinations of `Objects`, across all 111 VLMs and 65 valid human participants. For each combination, the order of objects matches the order they appear on the table, such that nearby objects in the combination visualization are nearby in the stimuli (we did not counterbalance the order when constructing the scenes). Indeed, human results show that cells near the diagonal have a higher frequency than cells that are further away. For example, the remote and the doll are nearby in stimuli with a remote, a doll, a coffee, and a bottle of water, likely causing the second cell in the first row to achieve a high frequency in the bottom-right confusion matrix for humans. VLMs also show similar effects: as the correct answer moves from the first *row object* (e.g., the doll for the bottom-left matrix) to the last *row object* (e.g., the soccer), the proportion of trials that were responded with the first *row object* decreases, and the proportion of trials that were responded with the last *row object* increases.

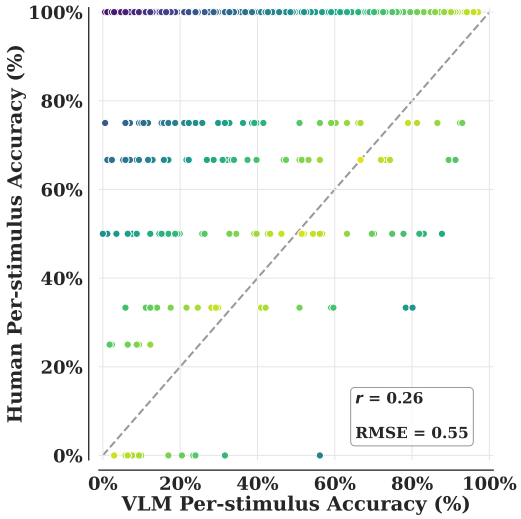


(a) Confusion matrices for VLMs. (b) Confusion matrices for humans.

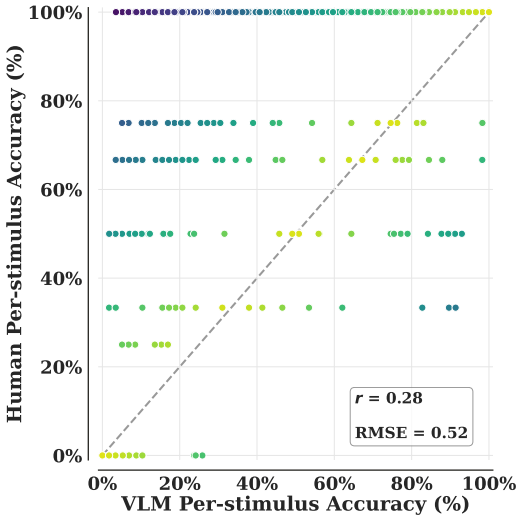
Figure 14: Confusion matrices. Each row corresponds to the ground truth object, and each column corresponds to the selection made by the VLM or human participant.

A.6 Per-stimulus Accuracy between Human and VLM

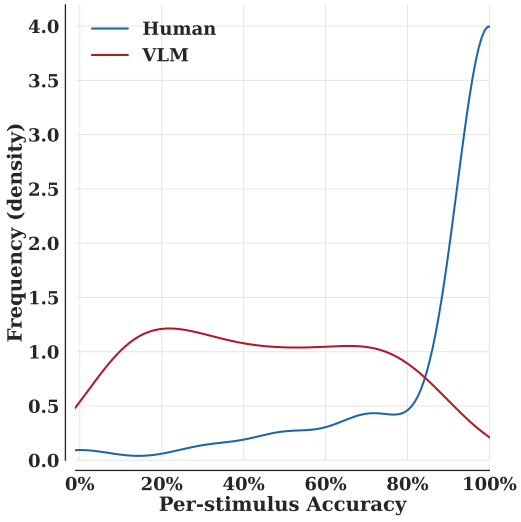
Fig. 15 shows per-stimulus accuracy (n = 900, each reflecting the mean accuracy of VLMs and humans on a particular stimulus) for VLMs versus human participants. From the plot, it is evident that a number of stimuli are trivial for human participants but remain challenging for VLMs. The Pearson correlation and RMSE between human and VLM accuracies indicate only modest correspondence but substantial overall differences in accuracy distributions between VLMs and humans. This result is consistent with the error consistency analysis in Analysis A.



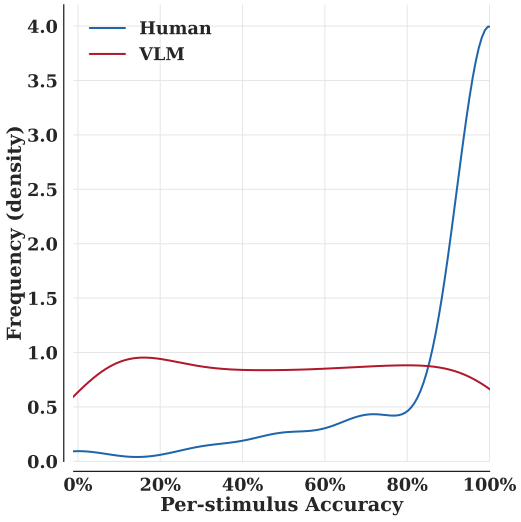
(a) Per-stimulus accuracy comparison between all VLMs and human participants. 851 datapoints lie above the diagonal, 48 fall below the diagonal, and 1 datapoint falls exactly on the diagonal.



(b) Per-stimulus accuracy comparison between 5 top-tier VLMs and human participants. 771 datapoints lie above the diagonal, 56 fall below the diagonal, and 73 datapoint falls exactly on the diagonal.



(c) Distribution of Per-stimulus accuracy comparison between all VLMs and human participants.



(d) Distribution of Per-stimulus accuracy comparison between 5 top-tier VLMs and human participants.

Figure 15: Per-stimulus accuracy and distribution comparison between VLMs and human participants. For subfigure (a) and (b), the color of each datapoint reflects the agreement between the two groups. Specifically, lighter colors indicate greater similarity in performance (and being closer to the diagonal), and vice versa. For subfigure (c) and (d), the distributions are visualized using kernel density estimation (KDE) curves.

A.7 VLM Response Distribution

Note that in the creation of prompts, we counterbalanced the correct answer by randomly shuffling the options in the multiple-choice questions. If VLM are randomly guessing, one way they can do it is by always outputting `A` (or any of the choices) blindly. The distribution of their response indicates that it is not the case, as shown in Fig. 16, although there is a slight tendency towards outputting `A` (likely due to the wording of guessing in Analysis A, as described in Appendix A.2).

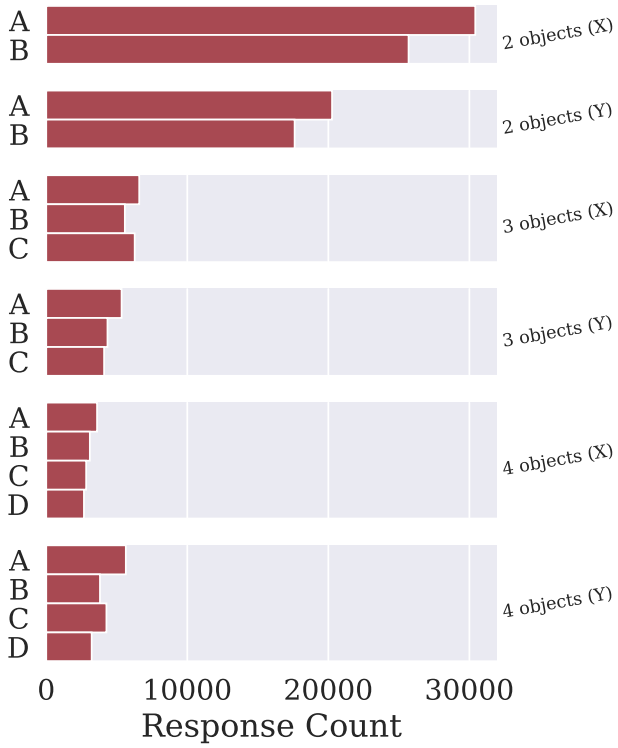


Figure 16: The distribution of VLM responses in terms of the options, combining trials in Analysis A and B. The brackets in row names denote whether the gazer is Actor X or Y for the combination of `Objects`.

A.8 VLM Data Point Overview

See Fig. 17. Each trial is represented as a colored cell. There are 900 columns, the same number as the number of tesst stimuli. Since VLMs involved only in Analysis A are presented with each stimulus once, and ones in Analysis B are presented with each stimulus ten more times, they occupy one and eleven rows respectively. Note that besides the five VLMs mentioned in the main text, Llama-3.2V-11B-cot was also evaluated using the Analysis B pipeline, such that each stimulus was presented 10 times, in addition to once in Analysis A. However, since its overall performance is close to the random-guessing baseline, we did not analyze its behavior.

Note that the main difference between the five top-tier VLMs in Analysis B and others is how they perform on the harder cases, especially when there are more `Objects`. Hints of *Proximity effect* can be found, while the finding of the null effect of VLMs size is also consistent with the visualization. Results of Analysis B VLMs also indicate that their performance is relatively robust: their eleven rows are very similar to each other (within each VLM).

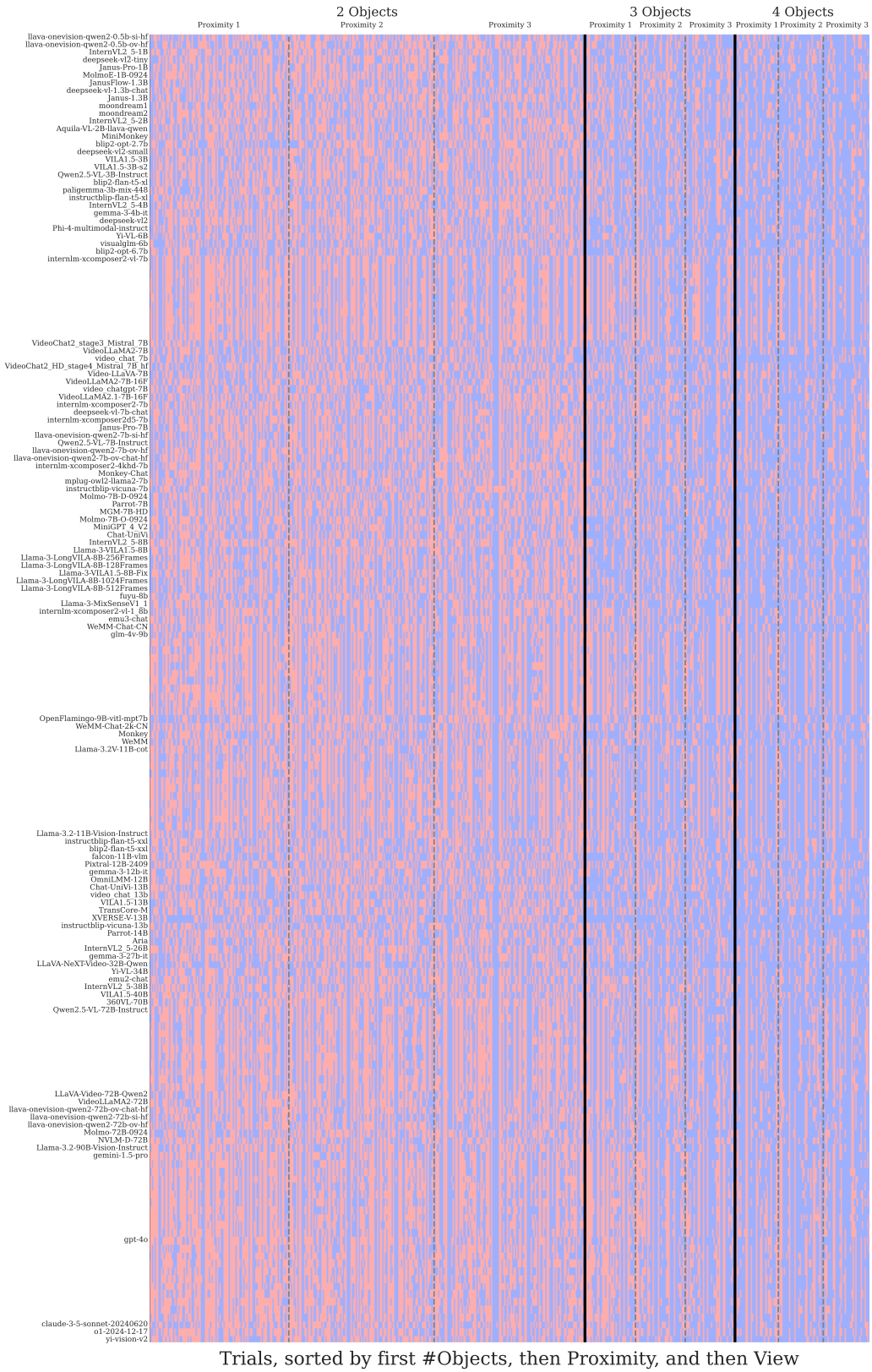


Figure 17: X axis represents stimuli. Blue cells are trials where the response is incorrect. Rows are sorted such that, roughly speaking, larger models are placed lower.

A.9 Human Data Point Overview

See Fig. 18. Again, participants were randomly assigned to one of the stimulus lists, each consisting of 45 test stimuli (shown here) and 7 predetermined attention-check stimuli (not shown here). Only participants who passed all attention checks are shown.

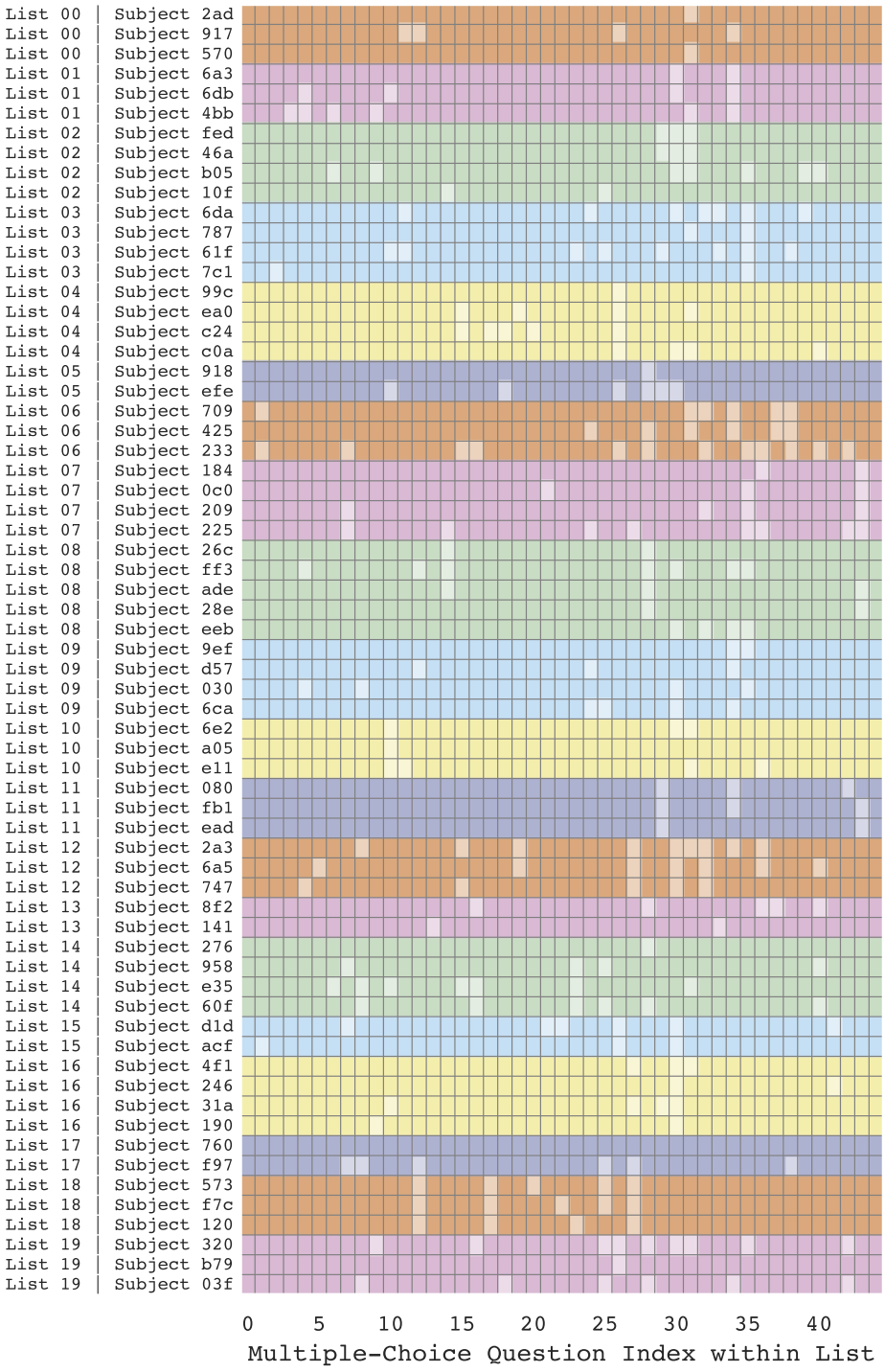


Figure 18: Each row depicts all the data points we collect from a human participant (who passed all attention checks), and each cell is a trial. Cells with a lighter color indicate wrong responses.

A.10 Human Performance by Condition

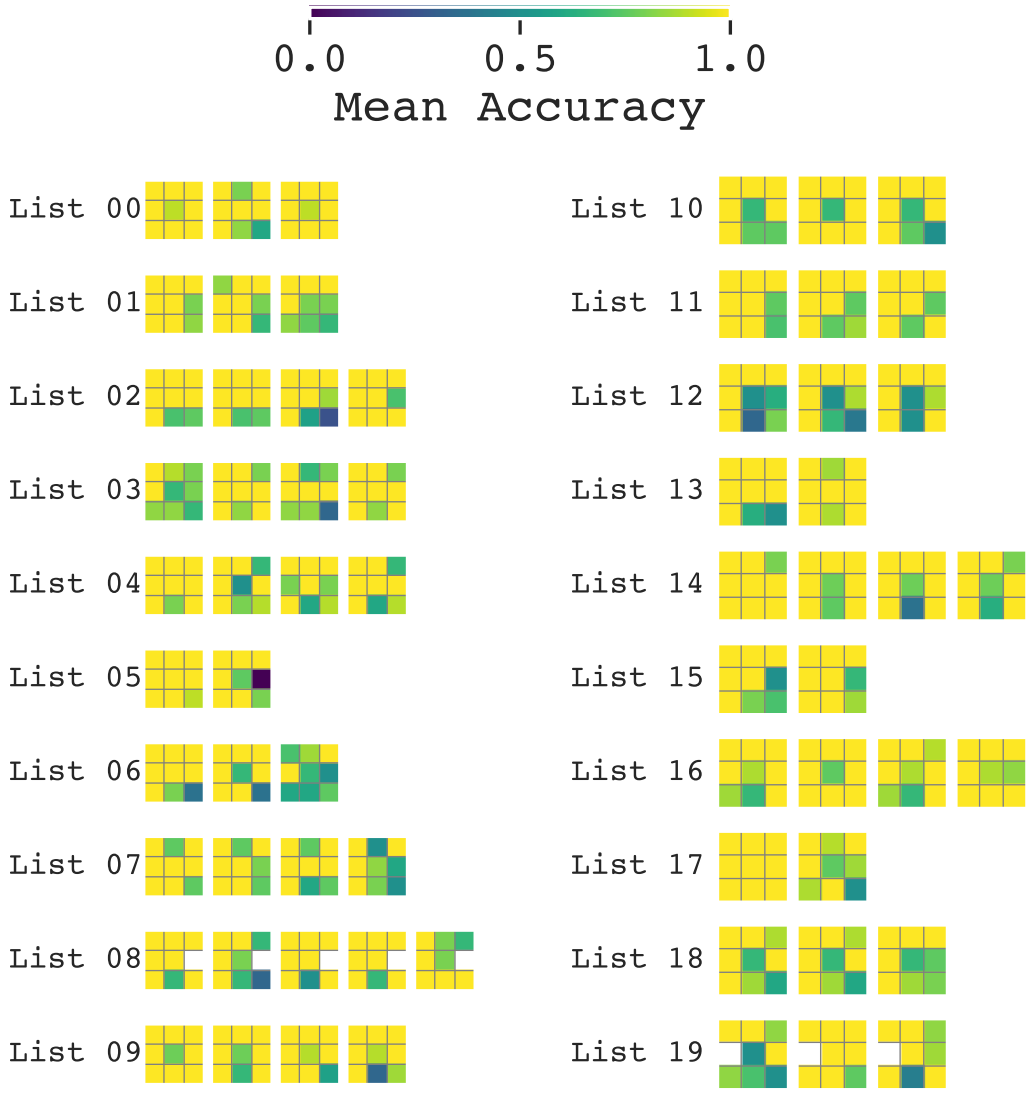


Figure 19: Each 3 by 3 matrix is a summary of a participant's performance. The rows from top to bottom represent `Proximity=1,2,3` respectively, and the columns from left to right represent `View=front, left, right` respectively. There is a general tendency of lower performance as `Proximity` gets higher (i.e., objects become closer to each other), especially when the `View`$\neq$`front`.

A.11 Accuracy Distribution of VLMs

Aggregated accuracy across the five VLMs against different controlled variables is shown in Fig. 20.

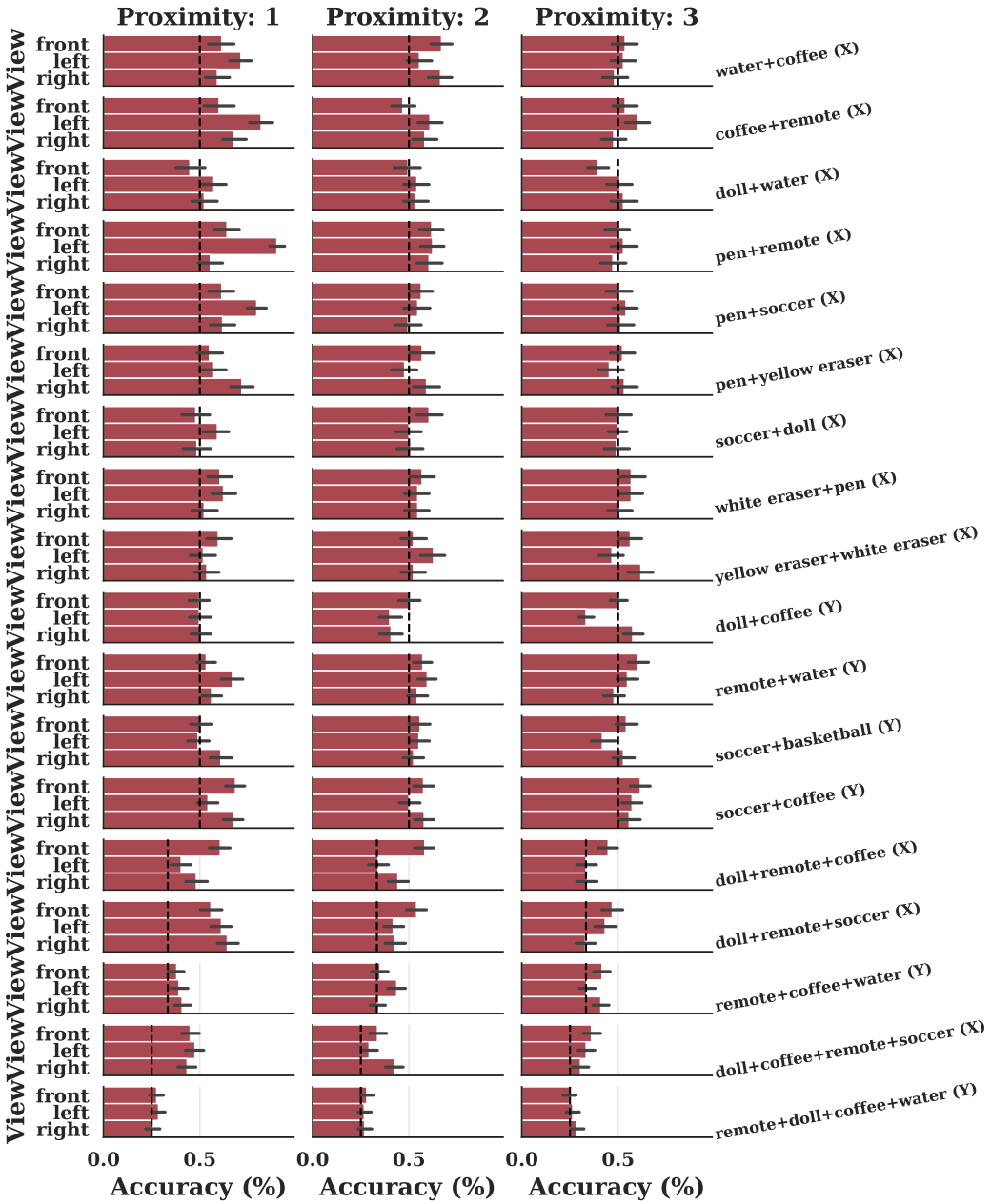
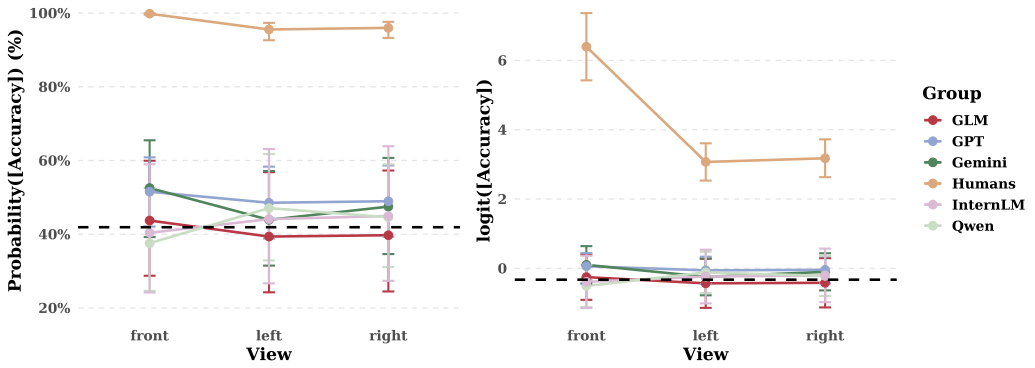


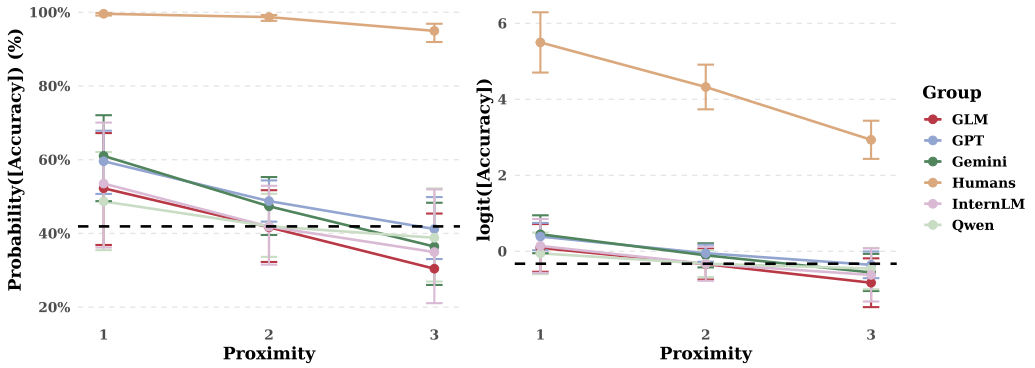
Figure 20: VLM Accuracy nested within `View`, `Proximity`, and `Objects`. Based on responses from the five top-tier VLMs. Dashed lines represent the random-guessing baseline. 95% CI drawn.

A.12 Mixed-Effects Modeling Details

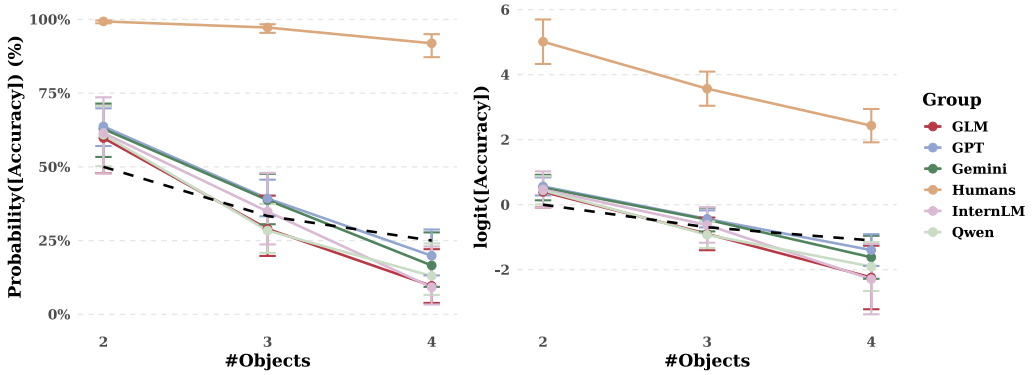
We fitted mixed-effects logistic models in R (Version 4.3.3) using the lme4 package (Version 1.1-35.5; Bates et al. 2015). Figure below shows the estimated marginal means obtained by fitting a mixed-effects model for each group, which then contributed to all curves of the same color in the figure. Dashed lines are the random-guessing baselines. The figures on the left depict variable relations with the accuracy in the probability space, while the figures on the right depict them in the logit space. Note that averaging is always performed in the logit space, as this is part of the reason why the link transformation is used in the first place. They also depict how probability space and logit space manifest the degradation differently: degradation is amplified when the accuracy is high (the case for humans) and reduced when accuracy is low (the case for VLMs) in logit space. No significant difference between marginal means for `View=left` and `right`.



(a) *View effect*



(b) *Proximity effect*



(c) *Choice effect*

Table 4: The estimated fixed and random effects statistics for each group. Significant effects are in black while others are in gray (using $\alpha = 0.05$). b = estimate; SE = standard error. The effect sizes are obtained by dividing the coefficient estimate by the effect size denominator (obtained by first adding $\pi^2/3$ to the sum of variances for all random effects, then taking a square root, following Muradoglu et al. 2023 All these significant effects cannot be explained by a random-guessing account of VLM's performance. Therefore, by manipulating these controlled variables, we can find behavioral features that constrain hypotheses of the inference underlying VLMs' emergent computation (i.e., excluding the random-guessing account).

Term	Statistic	Gemini	GPT	GLM	InternLM	Qwen	Humans
StimulusID	Variance	16.350	9.350	25.228	30.021	19.458	2.744
ParticipantID	Variance	NA	NA	NA	NA	NA	0.404
PromptID	Variance	0.025	NA	NA	NA	0.036	NA
Effect Size Denominator		4.434	3.555	5.340	5.772	4.773	2.537
Intercept	b	0.544	0.843	0.526	0.237	-0.059	7.414
	SE	0.307	0.218	0.375	0.432	0.347	0.527
View=left	b	-0.444	-0.170	-0.316	0.173	0.594	-3.236
	SE	0.381	0.280	0.488	0.548	0.429	0.422
	p	0.244	0.544	0.517	0.752	0.166	<0.001
	Cohen's d	-0.100	-0.048	-0.059	0.030	0.124	-1.275
View=Right	b	-0.252	-0.126	-0.293	0.245	0.443	-3.218
	SE	0.382	0.278	0.489	0.548	0.422	0.424
	p	0.509	0.650	0.550	0.655	0.293	<0.001
	Cohen's d	-0.057	-0.036	-0.055	0.042	0.093	-1.268
Proximity (centered)	b	-0.722	-0.464	-0.726	-0.657	-0.304	-1.305
	SE	0.193	0.140	0.249	0.278	0.211	0.168
	p	<0.001	0.001	0.004	0.018	0.150	<0.001
	Cohen's d	-0.163	-0.130	-0.136	-0.114	-0.064	-0.514
#Object (centered)	b	-0.906	-0.587	-1.459	-1.644	-1.145	-0.670
	SE	0.208	0.153	0.294	0.331	0.233	0.152
	p	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	Cohen's d	-0.204	-0.165	-0.273	-0.285	-0.240	-0.264
Gazer=Y	b	-0.166	-0.938	-1.348	-1.144	-0.744	-1.003
	SE	0.315	0.233	0.418	0.466	0.349	0.238
	p	0.598	<0.001	0.001	0.014	0.033	<0.001
	Cohen's d	-0.037	-0.264	-0.252	-0.198	-0.156	-0.395