IRIS: Intent Resolution via Inference-time Saccades for Open-Ended VQA in Large Vision-Language Models

Anonymous ACL submission

Abstract

Vision-Language Models (VLMs) have demonstrated remarkable capabilities in Visual Question Answering (VQA), yet they often struggle with referential ambiguity when multiple objects in an image could satisfy a given query. We introduce IRIS (Intent Resolution via Inference-time Saccades), a novel training-free approach that uses eye-tracking data in real-time to resolve ambiguity in open-ended VQA. Through a comprehensive user study with 500 unique image-question pairs, we demonstrate that fixations closest to the time participants start verbally asking their questions are the most informative for disambiguation in Large VLMs, more than doubling the accuracy of responses on ambiguous questions (from 35.2% to 77.2%) while maintaining performance on unambiguous queries. We evaluate our approach across state-of-the-art VLMs, showing consistent improvements when gaze data is incorporated in ambiguous image-question pairs, regardless of architectural differences. We release a new benchmark dataset to use eye movement data for disambiguated VQA, a novel real-time interactive protocol, and an evaluation suite.

1 Introduction

Visual question answering (VQA) represents a fundamental challenge at the intersection of Computer Vision and Natural Language Processing (NLP), requiring systems to understand both visual content and linguistic queries to generate appropriate responses. While recent Vision-Language Models (VLMs) have achieved impressive performance on standard VQA benchmarks (Chen et al., 2023; Peng et al., 2023; Chen et al., 2024; Liu et al., 2024; Wang et al., 2024b; Qin et al., 2025; Yao et al., 2025), they continue to struggle with a pervasive real-world challenge: referential ambiguity. When multiple objects in an image could plausibly satisfy a query, such as asking "What is that?", current VLMs lack the contextual grounding to identify the intended object referred to in the ambiguous query, defined as the "referent".

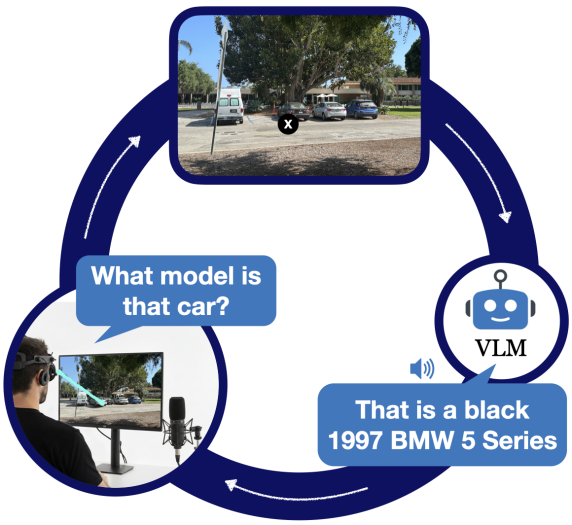

Figure 1: **IRIS overview**. Participant asks an ambiguous question about an image while their eyes are being tracked. The VLM uses the fixation data (marked as a white cross) to disambiguate the query and provide an accurate response in real-time.

Here, we present a solution that leverages eye movement fixations, a natural human behavior, to resolve referential ambiguity in open-ended VQA. Decades of cognitive science and psycholinguistics research show a tight coupling between eye movements, attention allocation, and linguistic planning (Just and Carpenter, 1976; Hayhoe and Ballard, 2005; Land, 2009; Yarbus, 1967). During natural viewing and questioning, fixations reliably precede verbal references by several hundred milliseconds, reflecting both planning and execution in speech production (Griffin and Bock, 2000). By aligning what is said with where (and when) people look, we obtain a time-locked, user-aligned signal that helps disambiguate referential intent in ambiguous VQA scenarios.

We present IRIS, a training-free approach that leverages human gaze data to enhance VLMs' abil-

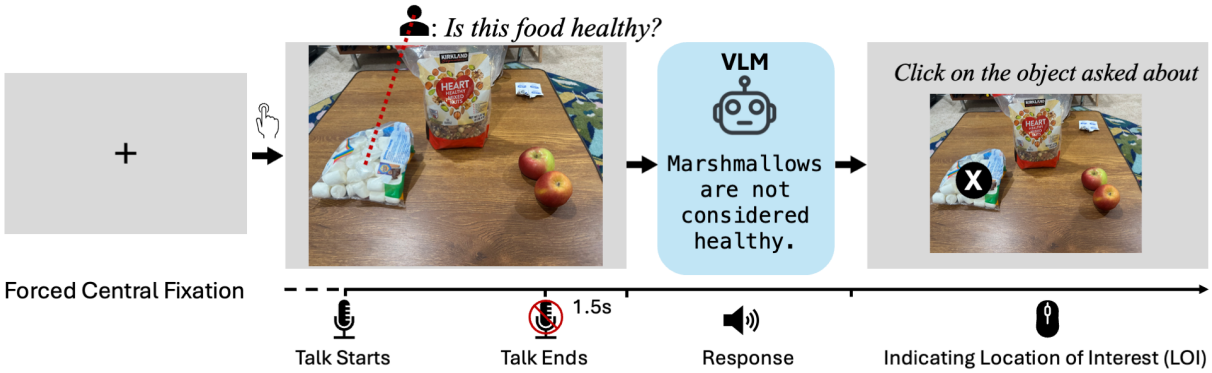

Figure 2: **Experimental procedure.** A central fixation check was enforced, after which participants freely viewed each image and asked any question aloud about it. Once 1.5s of silence elapsed following the question, the VLM was prompted with (i) the image, (ii) the transcribed question, and (iii) the same image with fixation data superimposed. Finally, participants reported the object they queried about (location of interest) by clicking the corresponding region of the image.

ity to resolve referential ambiguity in open-ended VQA. As illustrated in Figure 1, our system captures users' natural eye movements as they formulate questions about images, then provides these fixation patterns as additional context to guide VLMs toward the intended referent.

Prior work leverages human gaze across vision–language and recognition tasks. For image captioning, Sugano and Bulling (2016) integrated human fixations into an attention LSTM; for VQA, Inadumi et al. (2024) estimated within-image gaze to select regions of interest, and Sood et al. (2021) assessed human–model attention alignment without feeding gaze to the model. Beyond VQA, Karessli et al. (2017) encoded gaze as an auxiliary embedding to improve zero-shot image classification. In contrast, our approach operates at inference time without requiring model modification, making it immediately applicable to existing VLMs.

Through a comprehensive human study involving 500 unique image-question pairs, we collected a rich dataset of synchronized speech, gaze, and stimulus information. Participants formulated both ambiguous (for images containing multiple similar referents) and unambiguous questions (for images with a clear single referent), while we tracked their eye movements. Our experimental paradigm, detailed in Figure 2, captures the natural coupling between visual attention and linguistic formulation that occurs during human question-asking behavior.

Our temporal analysis reveals critical insights about the mechanism behind successful disambiguation using observers' gaze. We find that fixations occurring around the time of speech-onset provide the strongest disambiguation signals to VLMs. Moreover, we demonstrate that even a simple aggregation of all fixations during viewing significantly improves performance over image-only baselines ($p < .001$), suggesting that the concentration of gaze fixations itself carries disambiguating information. We evaluate our approach across 10 state-of-the-art (SOTA) VLMs, demonstrating consistent improvements when gaze data is incorporated. Our approach is not only grounded in literature linking eye movements and intent, but also practical and easily deployable. It finds immediate application in AR/VR systems that integrate eye tracking to deliver user-aligned ambiguity resolution in real-world interactive systems.

Our main contributions are:

- We introduce *IRIS*, a training-free approach that uses human eye data to steer VLM representations toward user-intended referents, resolving ambiguity without modifying model parameters or reliance on a specific architecture.
- Across 500 human-collected image–question pairs, conditioning on gaze around speech onset *more than doubles* accuracy for ambiguous questions (35.2% $\rightarrow$ 77.2%), while leaving unambiguous performance statistically unchanged.
- A controlled analysis identifies the most informative window around speech onset, evidence that gaze aligns with task-relevant features when the ambiguous question is formulated.
- We create a new real-time interactive experimental paradigm that synchronizes speech, gaze, and image information for VQA, as well as a corresponding offline evaluation suite and a new dataset on open-ended VQA.

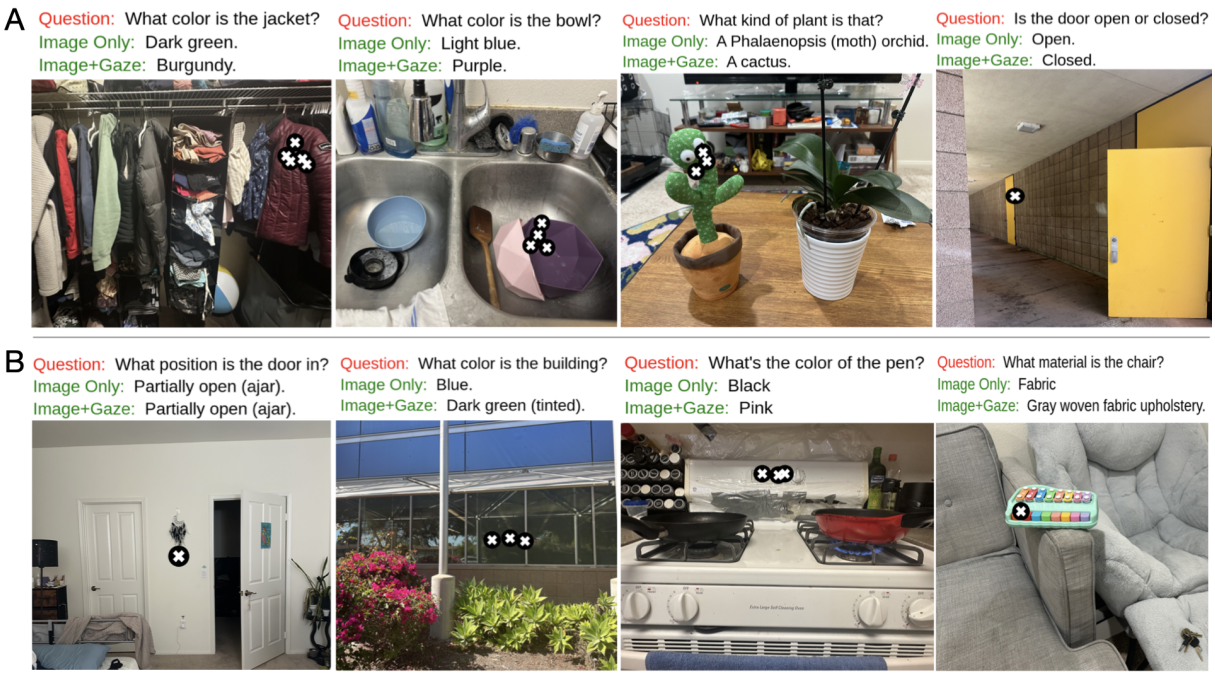

Figure 3: Qualitative results showing **successful (A) and failed (B) disambiguation using gaze data**. Black circles with white crosses mark temporally and spatially filtered fixation locations.

2 IRIS

2.1 System Architecture

Our IRIS system integrates three key components: (1) real-time eye-tracking to capture overt visual attention patterns (fixation locations), (2) speech recognition to identify question timing and content, and (3) MLLMs to generate responses. The system operates in a naturalistic setting where users freely explore images and formulate questions while their eye movements are recorded. The core insight driving our approach is that human gaze patterns during question formulation carry rich information about the intended referent of the question. When a person asks "What color is that?" while looking at a specific object among many, their fixations cluster around that object, providing a natural disambiguation signal. We capture these fixations and present them to the VLM as additional visual context, enabling the model to identify the referent and respond accurately.

2.2 Gaze Data Collection Protocol

We recruited 10 participants (5 female; age: 19-26 years, M=22±0.89 years) with normal or corrected-to-normal vision for a within-subjects study approved by the Institutional Review Board. All participants provided informed consent and received course credit for participation. Gaze was recorded monocularly from the left eye using EyeLink1000 Tower Mount (SR Research Ltd., 2005), at 1000 Hz (mean calibration error <1 dva; max <1.5 dva). Stimuli consisted of 50 photographs captured in everyday environments, with 40 scenes for ambiguous questions (containing multiple potential referents) and 10 for unambiguous ones (see Section A.1, Figure 8 for examples). The images were blocked, and ambiguous and unambiguous blocks were counterbalanced across participants. We also included 7 practice images from the MS COCO test set (Lin et al., 2014). Images were displayed on an Acer VG272X monitor (1920 × 1080 @ 60 Hz; 60 cm width) at 62 cm viewing distance (0.029 dva/pixel). Each session began with a 9-point calibration procedure. Participants completed a practice block with feedback on valid questions. Trials were gated by a central forced-fixation check (Figure 2): participants pressed space while fixating on a cross; progression required fixating within a 1.45 dva radius, with recalibration after three failures. Once the fixation check was successful, participants were shown the trial image with no time limit while they formulated an unambiguous or ambiguous question (based on the block assigned). Participants then asked their question out loud while continuing to view the image. OpenAI's speech-to-text and text-to-speech APIs were used to convert participants' voice to text questions and VLM's

text responses to voice output, respectively. Next, participants listened to the VLM's response. Finally, participants registered the Location of Interest (LOI), which was the location of the queried object using a mouse click. Sessions lasted 45 minutes per participant on average.

2.3 Gaze Data Processing

Gaze data typically consists of periods of relative stability, or fixations, which alternate with ballistic eye movements, also known as saccades. We identified fixations and saccades using EyeLink's online velocity and acceleration-based algorithm (SR Research Ltd., 2005) to parse saccades when eye velocity and acceleration exceed a threshold of $30°/s$ and $8000°/s^2$, respectively. The period between two saccades was classified as a fixation. The duration, start, and end locations of each fixation and saccade were recorded during the trial (see Section A.3 and Figure 12 for descriptive statistics of recorded fixations).

For each trial, we used a spatiotemporal filtering approach to denoise fixation signals. First, we identified a specific time window around speech onset, which was determined using voice activity detection on the audio stream (see Section A.3 and Figure 13 for details on speech onset latency and duration). Next, a spatial filter using the coordinate-wise median location of all temporally filtered fixations was applied. Any fixations that remained within a 2 dva radius of this median were kept, and the rest were discarded. If this step resulted in zero remaining fixations (10.8% of all 500 trials), we kept all the fixations that passed the temporal filtering stage. Finally, we overlaid the filtered fixations on the original image as white crosses on black circles (Figure 4). For more details on the efficiency of our filtering approach, see Section A.2 (Figures 9, 10, and 11).

2.4 VLM Integration

We employed a system prompt that instructed the VLM to use eye movement data for disambiguation. Importantly, we instructed the model not to mention the eye movement data in its response, maintaining a natural interaction flow. The complete set of system prompts for all our tasks, including accuracy evaluation and baseline response generation (as described in Section 2.7), is provided in A.5.

Figure 4: **Temporal-spatial filtering of eye gaze data.** Fixations colored by time. Black lines on the color bar mark speech onset and end. Diamonds represent fixations within ±1s of speech onset; all others are circles. Any diamond within 2 dva of the diamonds' median (red +) is spatially filtered and rendered as white crosses on black circles and passed onto IRIS.

2.5 Ground Truth Generation

Establishing ground truth for open-ended VQA presents unique challenges, as responses can vary in phrasing while conveying the same information. We developed a multi-stage process for ground truth generation. We first generated responses from three SOTA VLMs, including GPT-5 (OpenAI, 2025), Gemini 2.5 Pro (Gemini Team, Google DeepMind, 2025), and Claude Opus 4.1 (Anthropic, 2025), using the participants' LOI to identify the true object that was queried about. We then recruited five independent evaluators to review each image-question pair along with the model responses and the participants' LOI. They selected the most accurate response or provided a custom answer if none were satisfactory. This step minimized bias from any single model and incorporated human judgment for ambiguous cases. Because our semantic similarity metric (Section 2.6) is sensitive to textual length and detail, we standardized the final ground truth by selecting the shortest accurate response among those chosen by evaluators. This approach ensured consistent comparisons across trials and conditions, while avoiding deflated similarity scores due to verbose phrasing. Finally, all ground truth responses were manually verified for accuracy by the first two authors.

2.6 Evaluation Metrics

We employed two complementary metrics to assess system performance:

Accuracy: We evaluated response correctness using binary classification (correct vs. incorrect), assessed both manually and by an automated evalu-

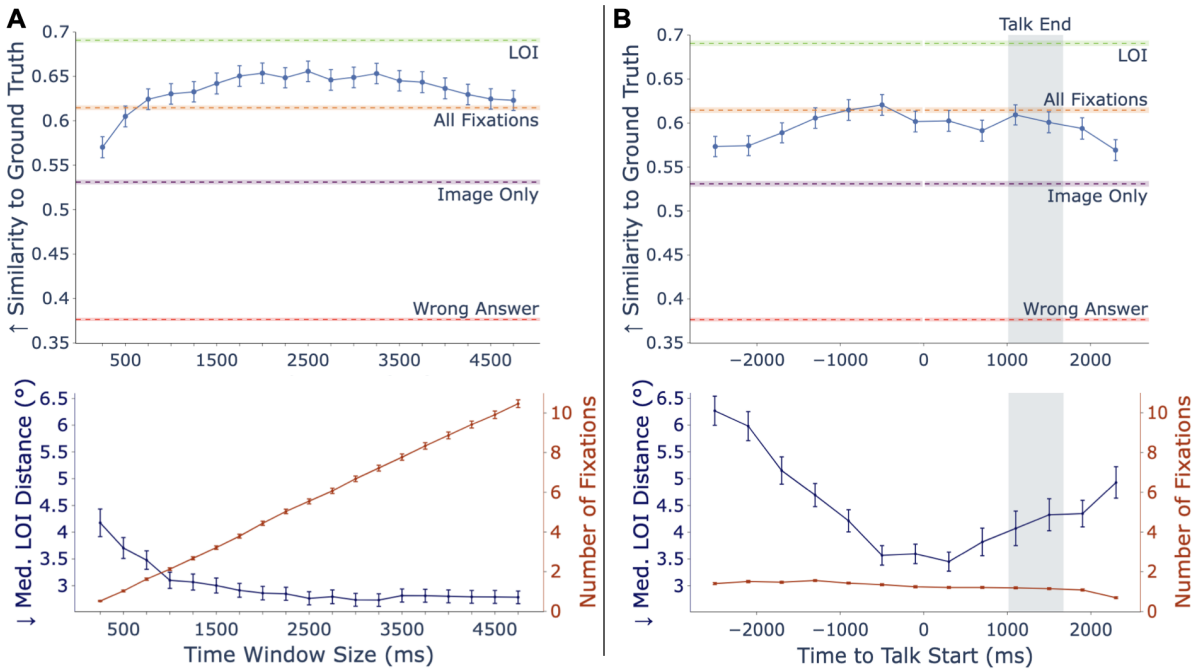

Figure 5: **Temporal dynamics of gaze informativeness** for ambiguous questions. **(A) Model performance increases as the temporal window expands around speech onset time**, converging to the "all-fixations" baseline similarity at ±4500ms around speech onset. The bottom panel shows decreasing distance between the fixation median and the Location of Interest (LOI) in dark blue and increasing fixation count with larger windows in red. **(B) Peak performance a few milliseconds before speech onset** is revealed in sliding window analysis (600ms window-width, 400ms sliding step-size), aligning with minimum fixation-to-LOI distance. Gray shaded region indicates the interquartile range of speech end times. Error bars represent SEM. (A) and (B) top: green - "LOI (perfect gaze)" upper bound; orange - "all-fixations" baseline; purple - "image-only" baseline; pink - "wrong answer" lower bound (see Section 2.7 for details).

ator VLM, Gemini-2.5-Flash (Gemini Team, 2025). The evaluator was provided with an accuracy evaluation prompt (see A.5), alongside the question–response pair and the corresponding image annotated with the Location of Interest (LOI). Based on this input, it classified each response as correct or incorrect. To assess the reliability of this automated evaluation, we manually compared responses generated by GPT-5-Mini (results in Figure 6) across all 500 image-question pairs 2and found an 88% agreement in accuracy ratings between human judgment and the VLM evaluator. Unless noted, all p-values reported are from two-tailed paired-samples t tests.

Embedding-Based Similarity Metric: Let g_i denote the ground truth response for trial i and $r_i^{(m)}$ the response of model m on the same trial. We use a fixed, frozen sentence encoder $E(\cdot)$ to embed both texts (OpenAI, 2024). Our primary score is cosine similarity outlined in 1. For each model m, we report the mean and standard deviation of $s(\cdot,\cdot)$ across trials.

$$s\left(r_i^{(m)}, g_i\right) = \frac{\langle E(r_i^{(m)}), E(g_i)\rangle}{\|E(r_i^{(m)})\|_2 \, \|E(g_i)\|_2}. \quad (1)$$

2.7 Semantic Similarity Baselines, Upper and Lower Bounds

Semantic similarity values range between 0 (dissimilar) and 1 (identical) (Equation 1). To interpret the effects of gaze on ambiguous and unambiguous VQA, we compared the model's response to several baselines and upper and lower bounds. We generated baselines by sending specific information to the model (detailed here), and computing the similarity of the model response thus generated to the ground truth. Upper bounds: (i) mouse-click LOI as "perfect gaze" with image and question, and (ii) inter-rater agreement between the five human-generated ground truth statements (mean pairwise similarity). We also included a baseline that sends all fixations (no temporal or spatial filtering) to the model. Lower bounds: (i) only image and question only (no gaze) baseline and (ii) a "wrong-answer"

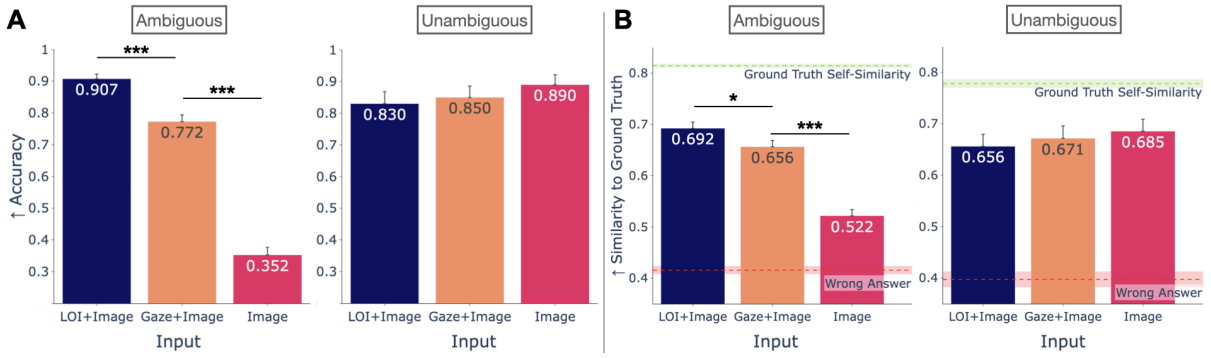

Figure 6: **Effect of gaze input on performance** compared to Image-Only and Image with Location of Interest (LOI). (A) By adding gaze to the input, **accuracy increases from 35.2% to 77.2%,** $p < .001$) for ambiguous questions, while remaining non-significant for unambiguous questions ($p = .52$). (B) Semantic similarity shows similar significant improvements only in ambiguous conditions. Error bars represent SEM shown on the upper side.

baseline, generated by prompting the model to respond incorrectly given the image, LOI, and question (see Section A.5), to control for object-level differences within each image.

3 Results

3.1 Qualitative Analysis

Figure 3 qualitatively showcases when gaze helps and when it does not (see also Section A.2, Figures 9 and 10). In Successful Cases (Figure 3A), fixations narrow the model to the intended referent: e.g., gazing at the burgundy jacket corrects "Dark green," fixating the purple bowl overrides a default to the salient blue one, fixations on the toy cactus disambiguate "plant," and fixations on the closed door spatially resolve a two-door scene. In Failed Cases (Figure 3B), gaze is uninformative or misleading: fixations fall between two doors, land on tinted glass yielding a wrong "Dark green" building color, reflect a voice-to-text error ("pen" vs "pan") with vague gaze, or miss the relevant chair entirely. Overall, success requires fixations that are (1) concentrated on the target around speech onset, (2) show clear separation from distractors, and (3) aligned to question intent. Otherwise, gaze can have no effect, or even harm model performance. For a detailed taxonomy and quantification of errors, see Section A.4 (Figure 14 shows examples of error types and Figure 15 quantifies that across all trials and participants).

3.2 Temporal Analysis of Gaze Information

To quantify features of gaze data that are: (i) crucial for successful disambiguation, and (ii) correlated with question intent, we systematically varied a temporal window centered on speech onset, to identify the optimal time window that contains the most informative gaze disambiguation signal (Figure 5). In Figure 5A, we observe that similarity to ground truth increases monotonically as we expand the temporal window from ±250ms to ±3000ms around speech onset. Performance starts at a similarity of 0.57 for the smallest window and reaches a peak of 0.65 at ±2500ms, after which it converges with the "all-fixations" baseline (0.615) at ±4500ms. This convergence suggests that fixations beyond ±4500ms around speech onset contribute minimal additional disambiguation value. The bottom panel reveals the mechanism underlying this performance curve: median fixation distance to LOI decreases from 4.0° to 2.7° as the window expands, while fixation count increases linearly. The inverse relationship between fixation-to-LOI distance and performance (correlation $r = -0.89, p < .001$) confirms that fixations closer to the intended object provide stronger disambiguation signals. Notably, even the all-fixations baseline similarity is significantly greater than that of image-only (0.615 vs 0.531, $p < .001$), indicating that the mere concentration of fixations during viewing helps resolve ambiguity compared to the absence of any gaze data. The LOI upper bound similarity (0.688) represents "perfect gaze" information, while the "wrong answer" baseline similarity (0.463) shows the lower bound by deliberately prompting the model to generate an incorrect response given image, LOI, and question (see Section 2.7).

Figure 5B employs a sliding window approach (600ms width) to pinpoint when gaze is most informative in relation to speech onset. Performance peaks at a similarity of 0.62 for windows centered near speech onset (-200ms to +400ms to speech on-

set), precisely when the median fixation-to-LOI distance is minimal (3.5°). This temporal alignment between optimal performance and closest fixations to the intended object validates that speakers naturally look at what they are asking about. The performance decline and median fixation-to-LOI distance increase after speech onset suggest that later fixations introduce noise rather than a disambiguation signal. Manual accuracy analysis further confirmed that our temporal filtering approach yields a 17.7% accuracy gain over the "all-fixations" baseline.

3.3 Impact of Gaze on Ambiguous vs Unambiguous Questions

We use the optimal time window identified in our analysis in Figure 5 to demonstrate the differential impact of gaze information on ambiguous and unambiguous questions (Figure 6). Gaze data within a 2-second window centered on speech onset provide substantial benefits when referential ambiguity exists, but offer minimal value when questions have clear, unique referents. For ambiguous questions (Figure 6A, left), accuracy increases dramatically from 35.2% (image-only) to 77.2% (gaze+image), representing a 115% improvement ($p < .001$). The large gap between image-only and gaze-augmented performance confirms that VLMs struggle with referential ambiguity but can effectively utilize gaze disambiguation signals when provided. In contrast, unambiguous questions show no significant improvement with gaze data (83.0% to 86.0%, $p = .52$). The minimal difference between conditions (LOI: 89.0%, Gaze: 86.0%, Image: 83.0%) suggests a ceiling effect where the questions are sufficiently clear that additional context provides marginal benefit, as expected for unambiguous scenarios.

The semantic similarity analysis (Figure 6B) corroborates these findings. Ambiguous questions show significant improvement from 0.531 (image) to 0.650 (gaze+image, $p < .001$), approaching the LOI performance of 0.691 ($p = .01$). The ground truth self-similarity of 0.82 represents the upper bound of achievable similarity given variation in human-generated ground truth responses. For unambiguous questions, similarity scores remain statistically unchanged across conditions (image: 0.656, gaze+image: 0.671, $p = .685$), reinforcing that gaze primarily benefits ambiguous scenarios. Importantly, the similarity-to-ground-truth measures for all conditions were greater than the "wrong answer" baseline (see Section 2.7), indicating that deliberately guiding the model to answer incorrectly given the LOI leads to worse model performance than the image-only condition.

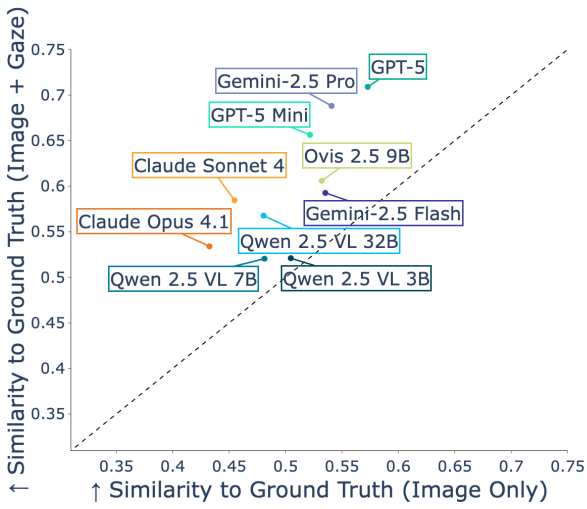

Figure 7: **Semantic similarity to ground truth across VLMs**. Models above the diagonal line show improvement with the inclusion of eye gaze data in ambiguous trials.

3.4 Performance Across VLM Architectures

Table 1 shows consistent accuracy gains on incorporating gaze data across diverse VLMs (OpenAI, 2025; Gemini Team, Google DeepMind, 2025; Anthropic, 2025; Lu et al., 2025; Bai et al., 2025a).

Across diverse architectures and model sizes, *Image+Gaze* uniformly outperforms *Image-only*, with improvements that are largest for frontier models and remain substantial for compact open-source ones. The effect correlates with instruction-following rather than parameter count alone, suggesting that gaze acts as a robust, architecture-agnostic disambiguation prior. Figure 7 shows the results of semantic similarity evaluation on the same models. Each point represents a model's performance with image-only similarity (x-axis) versus image+gaze similarity (y-axis) on ambiguous image-question pairs. Remarkably, all models show improvement with gaze data, falling above the diagonal identity line. GPT-5 achieves the highest absolute performance, reaching 0.71 similarity when augmented with gaze. These consistent across-family gains without fine-tuning or architectural changes suggest our method taps general visual-processing abilities and makes our approach immediately applicable to future VLMs as they continue to improve.

Model	Image Only (%)	Image+Gaze (%)	Δ (pp)	p-value
GPT-5 Mini	49.7	**83.0**	+33.2	< 0.001
Gemini 2.5 Flash	54.2	**83.3**	+29.0	< 0.001
Gemini 2.5 Pro	59.3	**82.3**	+23.0	< 0.001
GPT-5	53.7	**76.5**	+22.8	< 0.001
Claude Sonnet 4	54.2	**74.0**	+19.8	< 0.001
Ovis 2.5 9B	52.5	**69.5**	+17.0	< 0.001
Claude 4.1 Opus	55.2	**72.0**	+16.7	< 0.001
Qwen 2.5 VL 32B	56.2	**73.0**	+16.7	< 0.001
Qwen 2.5 VL 7B	51.7	**63.7**	+12.0	< 0.001
Qwen 2.5 VL 3B	52.0	**54.5**	+2.5	0.48

Table 1: **VLM Accuracies on ambiguous trials.** "Image+Gaze" uses fixations from ±1s around speech onset; Δ is the difference in percentage points. p-values test *Image+Gaze* vs. *Image-Only*.

Method	Similarity
Crosses (Ours)	$\mathbf{0.830 \pm 0.020}$
Heatmap	0.820 ± 0.019
Bounding Box	0.780 ± 0.021
Coordinates (Text)	0.748 ± 0.022
Cropped	0.703 ± 0.023

Table 2: **Ablation study on gaze representation methods.** Semantic similarity to ground truth using GPT-5-Mini on ambiguous trials. Our cross-marker approach outperforms alternative representations.

3.5 Ablation Studies

We conducted ablation studies to evaluate alternative methods for representing eye movement data to the VLM. Using GPT-5-Mini, we compared our baseline approach (white crosses overlaid on the image) against four alternative gaze representation methods (Table 2). **Heatmaps:** We generated Gaussian heatmaps from the eye movement data within the optimal time window and overlaid them on the original image with transparency. While heatmaps provide a continuous density representation of gaze attention, the smoothing inherent in heatmap generation may dilute the precise spatial information that discrete fixation markers preserve. **Bounding Boxes:** We clustered the filtered fixations using DBSCAN and drew bounding boxes around the resulting clusters on the image. **Raw Coordinates as Text:** Instead of visual overlays, we provided the raw (x, y) pixel coordinates of filtered fixations as text in the prompt alongside the unmodified image. **Cropped Images:** We cropped the image to a region centered on the median fixation location, discarding peripheral content. This approach performed the worst among all methods, as cropping removes potentially important contextual information and may eliminate parts of the queried object itself when fixations are offset from the object center.

4 Conclusion

We present IRIS, a training-free approach that leverages human eye-tracking data to resolve referential ambiguity in vision-language models. Across 10 participants and 50 images, fixation data from a critical temporal window around speech onset more than doubled accuracy on ambiguous questions while maintaining performance on unambiguous ones. The consistent effectiveness across 10 diverse large VLMs demonstrates the generality and immediate applicability of our method. As eye-tracking technology becomes ubiquitous through AR/VR devices and accessibility tools, gaze-augmented VQA represents a promising paradigm for intuitive human-AI communication. By releasing our benchmark dataset, real-time experiment code, and evaluation suite, we provide the research community with tools to further explore gaze-informed VQA. Future work should explore extending this paradigm to other modalities and tasks, potentially revolutionizing how we interact with intelligent systems.

Limitations

Our work represents an initial proof-of-concept for gaze-augmented VQA, and we intentionally scoped our study to controlled conditions to rigorously establish the core finding that natural gaze carries disambiguating information for VLMs. While we used research-grade eye-tracking equipment, the rapid advancement of consumer eye-tracking in devices suggests that the hardware requirements for IRIS-style systems are becoming increasingly accessible. Our participant sample, though sufficient to demonstrate statistically significant effects, reflects a university population; future work could expand demographic coverage to further validate generalizability.

References

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

Anthropic. 2025. Claude opus 4.1. https://www.anthropic.com/news/claude-opus-4-1. Accessed 2025-09-22.

S. Bai and 1 others. 2025a. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*. Single reference for Qwen2.5-VL 3B/7B/32B variants.

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

Ali Borji and Laurent Itti. 2013. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207.

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

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, and 1 others. 2023. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Moreno I Coco and Frank Keller. 2012. Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive science*, 36:1204–1223.

Heiner Deubel and Werner X. Schneider. 1996. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, 36(12):1827–1837.

Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google DeepMind. Gemini 2.5 Flash model.

Gemini Team, Google DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google DeepMind. Covers Gemini 2.5 Pro and 2.5 Flash; Accessed 2025-09-22.

Zenzi M Griffin and Kathryn Bock. 2000. What the eyes say about speaking. *Psychological science*, 11(4):274–279.

Mary Hayhoe and Dana Ballard. 2005. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194.

Mary M Hayhoe, Travis McKinney, Kelly Chajka, and Jeff B Pelz. 2012. Predictive eye movements in natural vision. *Experimental brain research*, 217(1):125–136.

James E. Hoffman and Babu Subramaniam. 1995. The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, 57(6):787–795.

Muhammet Furkan Ilaslan, Chenan Song, Joya Chen, Difei Gao, Weixian Lei, Qianli Xu, Joo Hwee Lim, and Mike Zheng Shou. 2023. GazeVQA: A video question answering dataset for multiview eye-gaze task-oriented collaborations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Shun Inadumi, Seiya Kawano, Akishige Yuguchi, Yasutomo Kawanishi, and Koichiro Yoshino. 2024. A gaze-grounded visual question answering dataset for clarifying ambiguous Japanese questions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 558–571.

Laurent Itti and Christof Koch. 2001. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203.

Robert J. K. Jacob. 1991. The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Trans. Inf. Syst.*, 9(2):152–169.

Marcel Adam Just and Patricia A Carpenter. 1976. Eye fixations and cognitive processes. *Cognitive psychology*, 8(4):441–480.

Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. 2017. Gaze embeddings for zero-shot image classification. *Preprint*, arXiv:1611.09309.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Michael F Land. 2009. Vision, eye movements, and natural behavior. *Visual neuroscience*, 26(1):51–62.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Hsin-Hung Li, Nina M. Hanning, and Marisa Carrasco. 2021. To look or not to look: dissociating presaccadic and covert spatial attention. *Trends in Neurosciences*, 44(8):669–686.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755, Cham. 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. *arXiv preprint arXiv:2304.08485*.

S. Lu and 1 others. 2025. Ovis2.5 technical report. *arXiv preprint arXiv:2508.11737*.

Päivi Majaranta and Andreas Bulling. 2014. Eye tracking and eye-based human–computer interaction. In Stephen H. Fairclough and Kiel Gilleade, editors, *Advances in Physiological Computing*, Human–Computer Interaction Series. Springer, London.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. *Preprint*, arXiv:1511.02283.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, and 13 others. 2024. MM1: Methods, analysis & insights from multimodal llm pre-training. *Preprint*, arXiv:2403.09611.

OpenAI. 2024. Openai text embeddings. `https://platform.openai.com/docs/guides/embeddings`. Accessed: September 2025.

OpenAI. 2025. Introducing gpt-5. `https://openai.com/index/introducing-gpt-5/`. Accessed 2025-09-22.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Michael I. Posner. 1980. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32(1):3–25.

Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2024. Rephrase, augment, reason: Visual grounding of questions for vision-language models. *Preprint*, arXiv:2310.05861.

Xiaofei Qin, Ruiqi Pei, Changxiang He, Fan Li, and Xuedian Zhang. 2025. Efficient knowledge distillation and alignment for improved kb-vqa. *Scientific Reports*, 15(1):20682.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Daniel Reich and Tanja Schultz. 2024. On the role of visual grounding in VQA. *arXiv preprint arXiv:2406.18253*.

Karthikeyan Shanmuga Vadivel, Thuyen Ngo, Miguel Eckstein, and BS Manjunath. 2015. Eye tracking assisted extraction of attentionally important objects from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3241–3250.

Ekta Sood, Fabian Kögel, Florian Strohm, Prajit Dhar, and Andreas Bulling. 2021. VQA-MHUG: A gaze dataset to study multimodal neural attention in visual question answering. *CoRR*, abs/2109.13116.

SR Research Ltd. 2005. *EyeLink 1000 [Apparatus and software]*. SR Research Ltd.

Yusuke Sugano and Andreas Bulling. 2016. Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203*.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2018. Anaphora and coreference resolution: A review. *Preprint*, arXiv:1805.11824.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Alberto Testoni, Barbara Plank, and Raquel Fernández. 2024. Racquet: Unveiling the dangers of overlooked referential ambiguity in visual llms. *Preprint*, arXiv:2412.13835.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *CVPR*, pages 9568–9578.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, and 1 others. 2024b. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Chi Chen, Haoyu Li, Weilin Zhao, and 1 others. 2025. Efficient gpt-4v level multimodal large language model for deployment on edge devices. *Nature Communications*, 16(1):5509.

Alfred L. Yarbus. 1967. *Eye Movements and Vision*, 1 edition. Springer, New York, NY. SpringerLink eBook edition published 2013.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. *Preprint*, arXiv:1511.05099.

A Appendix

A.1 Examples of Ambiguous and Unambiguous Questions

Each of our 10 participants formulated their own questions for the same 50 images (10 for unambiguous questions and 40 for ambiguous ones), thus resulting in 500 unique image-question pairs. Figure 8 shows sample questions on a few example images.

The top panel shows two examples from the Unambiguous condition. For instance, in Unambiguous Example 1, participant AR asks about "the purple object," and there is only one such object present in the image. In Unambiguous Example 2, participant KM asks about the "cactus," again unambiguously referring to the only cactus in the image.

In contrast, the bottom panel shows examples of ambiguous questions. For instance, in Ambiguous Example 1, participants AR and KM both ask about a mug, but it is unclear from their questions which one of the two mugs present in the image is being referred to, making their questions ambiguous given the image information. Similarly, in Ambiguous Example 2, participants KM and VR inquire whether the "food" is healthy, but there are several food items present, of which some are healthy (apples) but others are not (marshmallows), thus making their question ambiguous about the intended referent. KV's question on this image is not ambiguous and, thus, inappropriate for our study (see Section A.4)

A.2 Temporo-Spatial Visualization of Raw Fixation Data

We plot the raw fixation start locations on trial images to visualize the temporal and spatial features of our eye movement data with respect to task-related events such as speech onset time and mouse click location. Figure 9 shows example successful trials, where the eye gaze data helped the VLM disambiguate the object being asked about in ambiguous scenarios. Figure 10 shows example unsuccessful trials, where the inclusion of eye gaze data did not disambiguate the object in question for our VLM and algorithm. Finally, Figure 11 describes examples of a few rare cases (1% of all trials) where our spatial filtering approach resulted in discarding all fixations and lead to an incorrect VLM response, but keeping all fixations that occurred within ±1s of speech onset rescued the trial. Importantly, the data shown are from different participants, emphasizing the generality of our approach to eye gaze data and questions collected across subjects.

In the figure above, we show all fixations made by a participant during the trial as circle and diamond markers, colored according to the temporal order of their occurrence. Specifically, bluer markers occurred earlier in the trial, whereas yellower markers occurred later. Diamond markers show all fixations that occurred within ±1s around speech onset (temporal filter). The magenta circle indicates the mouse-click location registered by the participant at the end of the trial. The red cross shows the coordinate-wise median location of all diamonds. The red dashed circle marks a region of 2 dva radius centered on the median. This is our spatial filter. Any diamonds that were within this red dashed circle were sent to the VLM, and the rest were discarded. Except for those trials where the spatial filtering process resulted in zero remaining fixations, in which case all temporally filtered fixations, or diamonds, were sent to the model.

This visualization enables us to easily grasp how eye gaze evolves through time as a function of task instructions. Several features of the gaze data are immediately apparent. First, subjects take different amounts of time for different trials; however, the duration of speech remains more or less the same across trials. Second, the longer the time taken in a trial, the more fixations are made. Third, although fixations closer in time usually cluster together, there are also fixations made later on the same objects visited previously. Finally, and most importantly, fixations made around the time of speech (or question asked) generally indicate the location of the object eventually reported by the subject with a mouse click. In other words, people look at the object they are asking about around the time they utter the question.

Although our temporal and spatial filtering algorithm did well for the most part, achieving successful disambiguation in 88% of images overlaid with eye gaze data, there were a few cases where it failed (see Section A.4). As described in the main text, some of the original failures were caused by our spatial filtering process, resulting in zero remaining fixations. For these cases, all fixations that remained after temporal filtering were sent to the VLM. This led to a 1.5% improvement in model performance, measured as the accuracy of model response.

Next, we show some cases of failures in Fig-

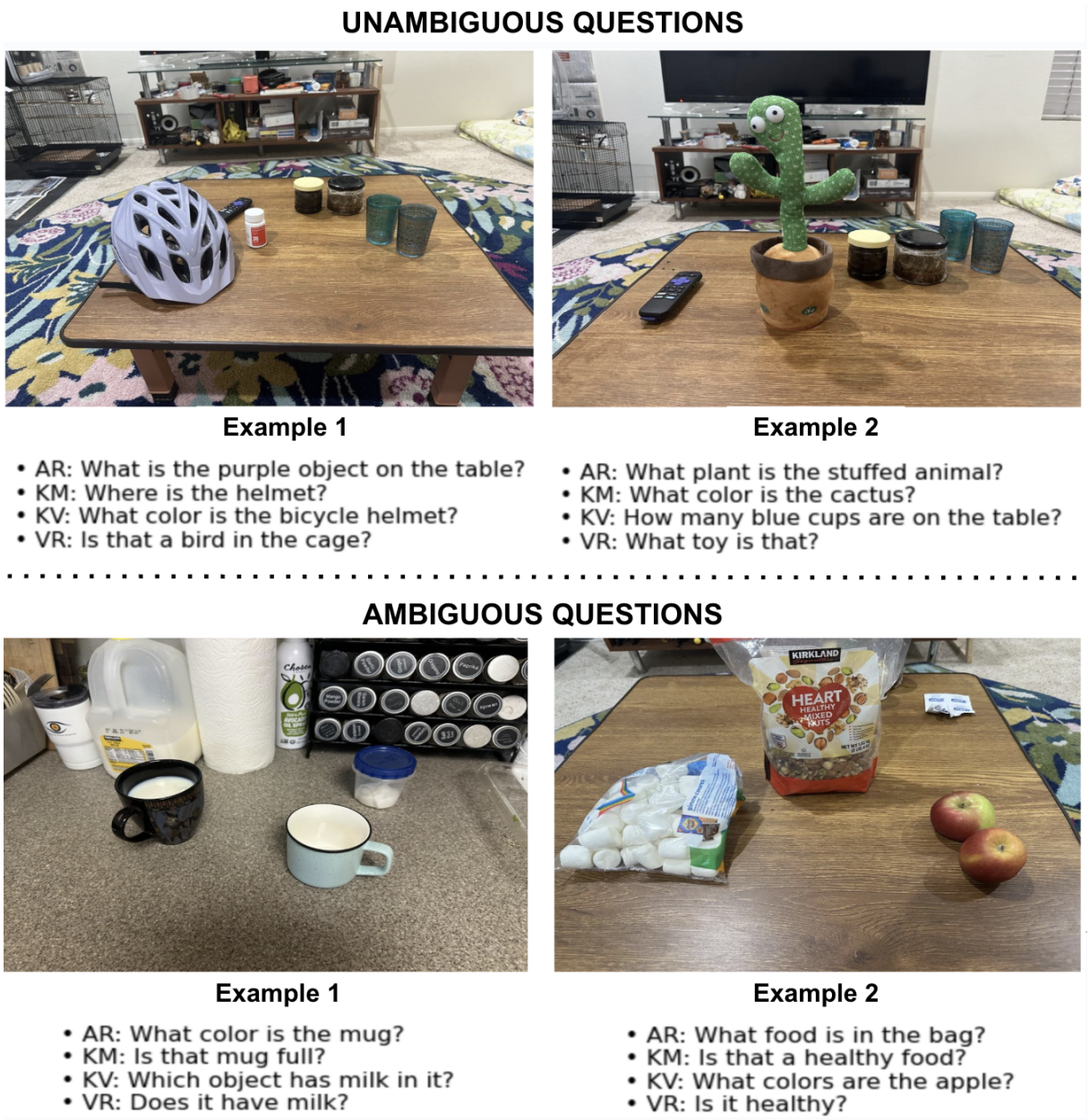

Figure 8: **Sample unambiguous and ambiguous questions** asked by four participants on four example images.

ure 10, including cases where there were zero remaining fixations after employing a median-based spatial filter (right top and right bottom panels). There are several causes of these failures: first, the object in question may have been fixated too few times leading to a shift in the median towards the wrong referent (top left panel in Figure 10), or there might have been too few fixations close to speech onset on the correct object (bottom left), or too many fixations overall leading to an uninformative median (top and bottom right), or the model might have been incorrect even with the right eye gaze data (detection error, as described in Section A.4).

Finally, we present examples of the 2% cases where median filtering resulted in zero fixations, and retaining all fixations that occurred within ±1s of speech onset (essentially removing the spatial filter) yielded a more accurate VLM response (Figure 11).

A.3 Speech Behavior and Fixation Descriptive Statistics

We present more insights into the speech (questioning) behavior of the subjects in the study, and also discuss the descriptive statistics of the fixations collected in our data.

First, Figure 12 shows the average time taken by participants to ask a question once the trial

starts (left) and the average time between speech onset and end (in other words, the mean duration of speech during asking the question) (right). Although participants took longer to ask a question in the Ambiguous condition (M=11.51s, SE=2.45s) than in the Unambiguous condition (M=7.25s, SE=1.02s) on average, this difference was not significant ($t(9) = -1.87, p = 0.09$). The mean duration of questions asked was also not significantly different between Ambiguous (M=1.47 s, SE=0.11 s) and Unambiguous (M=1.56 s, SE=0.22 s) conditions ($t(9) = .46, p = 0.66$).

Next, Figure 13 shows some descriptive statistics for the fixations observed in our study. The leftmost panel shows the average fixation number for our two conditions. This aligns with the longer time between the stimulus onset and the start of questioning observed for ambiguous trials: a longer time allows for more fixations. Although the Ambiguous condition has a higher mean number of fixations (M=39.71, SE=6.95) than the Unambiguous condition (M=28.91, SE=3.79), this difference is not significant ($t(9) = -1.71, p = 0.121$). The middle panel shows the mean duration of a given fixation on unambiguous and ambiguous trials. On average, fixation durations were highly similar between Ambiguous (M=268.31ms, SE=12.74ms) and Unambiguous conditions (M=265.06ms, SE=17.30ms) ($t(9) = -0.43, p = 0.675$). Finally, the rightmost panel shows the mean latency of the first fixation after stimulus onset. As observed for fixation duration, the average first fixation latency was very similar between Ambiguous (M=388.62ms, SE=14.47ms) and Unambiguous conditions (M=381.32ms, SE=15.96ms) ($t(9) = -0.51, p = 0.621$).

A.4 Taxonomy of VLM Response Errors: Unsuccessful Trials

We analyze the "unsuccessful" trials to quantify why some trials fail in terms of the various types of errors observed in the different stages of the question-answering process. To do this, we manually categorized each of the 1500 VLM responses (50 stimulus images X 10 subjects X 3 categories - Image+LOI, Image+Gaze, Image Only) into four types of errors. The types of errors, listed roughly in the temporal order of their potential occurrence, are elaborated below along with some examples (see Figure 14):

- Error Type 1. Inappropriate Question. In this case, the failure happened because the participant asked a question that was not suitable for the purposes of testing whether eye gaze data can help disambiguate visual information in the presence of ambiguity. For example, asking about the color of a bottle when there is only one bottle in the scene for a trial in the Ambiguous condition. Around 5.2% of the trials had this kind of error.

- Error Type 2. Detection Error. In this case, an appropriate question was asked by the participant, yet the model failed to answer the question accurately when eye gaze data or mouse click locations for ambiguous trials were sent to the VLM, or for unambiguous trials with or without eye gaze data and mouse click locations. Around 5.1% of the trials had this kind of error.

- Error Type 3. At this stage, the participant had asked an appropriate question, but one of the following things happened, depending on whether eye gaze data or mouse-click location was sent to the model along with the image:

  - Hallucination/Detail Error (Image+LOI or Image+Gaze). In this case, the VLM correctly detected the object category being asked about, yet responded with an incorrect detail and/or hallucinated. Approximately 2% of all 'Image+LOI' and 'Image+Gaze' trials had this type of error.
  - Referent Bias (ambiguous Image Only trials). In this case, the VLM assumes one of the multiple referents was being asked about and answers with the details of that object. This reflects the model's tendency to select a particular instance as the intended referent when multiple same-category instances are present. To be clear, when no eye gaze data is sent to the VLM, an accurate response of the VLM could either be a follow-up clarification question or a listing of the details of all the instances of a category (see bottom right "No Error" example in Figure 14). Around 57% of all 'Image Only' trials had this kind of bias.

- Error Type 4. Eye Data Error. The final form of failure is observed when the VLM response

is accurate for the Image+LOI subtype but inaccurate for the Image+Gaze subtype for the same trial (image-question pair). This implies that the portion of the gaze data sent to the VLM for that trial might not have accurately captured the object being asked about. 11.6% of all 'Image+Gaze' trials had this kind of error. Importantly, only ambiguous trials could have this kind of error because eye gaze data would be uninformative for unambiguous trials.

Figure 15 shows the proportion of trials within a specific condition (Ambiguous or Unambiguous) and a specific VLM input type (Image+LOI, Image+Gaze, and Image Only) with the different error types as discussed above, both as subject-wise and across subjects measures. It is worthwhile to note that most of the errors of type 1 (inappropriate questions) were contributed by only a few of the subjects, such as IC and KZ. Secondly, for the Unambiguous condition, Detection Error forms the major contributor to unsuccessful trials, when mouse click location or eye gaze data is sent to the VLM. For the Ambiguous condition, a major portion of the unsuccessful trials is caused by error type 4 (Eye Data Error), which is when the VLM responds accurately when prompted with Image+LoI but not with Image+Gaze. Finally, for ambiguous trials, when only the image without any mouse clicks or eye gaze data is sent to the model, a major portion (70%) of the responses contain type 3B errors (Assumption Error). This is in line with the findings of Testoni et al. (2024) who show that LLMs have a bias in responding with stereotypical, and sometimes risky, assumptions when prompted with inherently ambiguous questions.

A.5 List of System Prompts

VQA prompt for Image+Gaze/LOI input: *"You are an expert visual interpreter specialized in identifying and describing features of specific objects in images by considering the eye movement data. The user will provide an original image alongside the same image with eye movement data, where fixation points indicated by white X signs. Alongside these two images, the user will ask a specific question regarding an object's features, such as its color, size, shape, or spatial location relative to other objects present in the image. Your task is to carefully analyze the image, identify the specific object in question considering the eye fixation points, and answer the user's question precisely and concisely. Your response should only describe the referred object closest to the fixation points indicated by white X signs. You should provide clear, factual answer without any extra output. Do not speculate or include details unrelated to the indicated object or question. Do not mention the eye movement data in your response. Your response should be very short, concise, but accurate."*

VQA prompt for Image-Only input: *"You are an expert visual interpreter specialized in identifying and describing features of specific objects in images. The user will provide an image, and alongside that image, the user will ask a specific question regarding an object's features, such as its color, size, shape, or spatial location relative to other objects present in the image. Your task is to carefully analyze the image, identify the specific object in question, and answer the user's question precisely and concisely. You should provide clear, factual answer without any extra output. Do not speculate or include details unrelated to the indicated object or question. Your response should be very short, concise, but accurate."*

VQA prompt for generating "wrong answer": *"You are an expert visual interpreter specialized in identifying features of specific objects in images by considering the eye movement data. The user will provide an original image alongside the same image with eye movement data, where fixation points indicated by white X signs. Alongside these two images, the user will ask a specific question regarding an object's features, such as its color, size, shape, or spatial location relative to other objects present in the image. Your task is to carefully analyze the image, identify the specific object in question considering the eye fixation points, and answer the user's question concisely but wrongly. Your response should only be the wrong answer about the referred object closest to the fixation points indicated by white X signs. You should provide clear, wrong answer without any extra output. Do not speculate or include details unrelated to the indicated object or question. Do not mention the eye movement data in your response. Your response should be very short, concise, but wrong."*

VLM accuracy evaluation prompt: *"You are an expert visual interpreter specialized in identifying the correctness of an answer to a question about an object in the image. The user will provide an image alongside the same image with the referent object in question indicated by a white X sign.*

Alongside these two images, the user will provide the specific question and answer. Your task is to analyze the images and decide whether the answer about the referent object is correct or not. Your response should only include one word: "correct" or "incorrect". Do not include any other words or details in your response."

A.6 Participant Task Instructions

Participants received the following verbal instructions at the start of the experiment: *"In this study, you will view a series of images on the screen. For each image, your task is to formulate a question about a specific object in the scene and ask it out loud. In some blocks, you will be asked to formulate questions that could refer to multiple objects (ambiguous), while in other blocks, your questions should clearly refer to only one object (unambiguous). After you ask your question, you will hear a response from an AI assistant. Finally, you will click on the object you were asking about. Please keep your head still and positioned on the chinrest throughout the experiment. You may take breaks between blocks if needed."* Participants then completed 7 practice trials with feedback before beginning the main experiment.

A.7 Related Work

Vision-Language Models and Visual Question Answering. VLMs have progressed from early two-stream transformers (VisualBERT (Li et al., 2019), LXMERT (Tan and Bansal, 2019)) and contrastive pretraining (CLIP; (Radford et al., 2021)) to instruction-tuned systems (Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023), LLaVA (Liu et al., 2023)) and recent open models that markedly advance VQA and multimodal reasoning: LLaVA-OneVision (Li et al., 2024), Qwen2-VL and Qwen2.5-VL (Wang et al., 2024a; Bai et al., 2025b), InternVL 2.5 (Chen et al., 2025), IDEFICS2 (Laurençon et al., 2024), and large-scale pretraining analyses such as MM1 (McKinzie et al., 2024). Yet grounding issues persist, including failures owing to visual shortcomings in MLLMs (Tong et al., 2024) or bypassing visual grounding in shortcut learning (Reich and Schultz, 2024). When questions are underspecified, VQA performance may be helped by prompt-level fixes (e.g., visually grounded rephrasing) (Prasad et al., 2024). Complementing these trends, emerging work integrate human gaze cues within images to clarify referents by introducing new model architectures (Inadumi et al., 2024). In contrast, our approach leverages the viewer's natural gaze at inference, without architectural changes or retraining, to guide the model toward the intended referent and reduce ambiguity in open-ended VQA.

Eye Movements and Attention. Foundational research links eye movements to covert attention (Posner, 1980; Hoffman and Subramaniam, 1995; Deubel and Schneider, 1996; Li et al., 2021). In natural tasks, gaze anticipates actions and speech, revealing time-resolved planning during sentence production (Coco and Keller, 2012), and scene and event description (Hayhoe et al., 2012; Griffin and Bock, 2000). Classic studies have demonstrated that what observers look at depends strongly on visual and linguistic context and goals (Yarbus, 1967; Tanenhaus et al., 1995). While this temporal coupling is well established, it is rarely *operationalized* to select referents at the precise moment of question formulation in VQA. This work aligns a short window around speech onset with fixations to harvest the most informative disambiguation cue.

Eye-Tracking in Vision-Language and Computer Vision. Eye movements have a rich history of informing models of perception and learning. Yarbus (1967) pioneered the use of eye movements to understand visual perception, showing that fixation patterns vary with task demands. In the context of language processing, Tanenhaus et al. (1995) demonstrated that eye movements during spoken language comprehension reflect real-time semantic processing. Interactive systems have long explored gaze as an input modality: Jacob (1991) introduced early gaze-based interaction paradigms, and Majaranta and Bulling (2014) surveyed the evolution of eye-tracking interfaces. Building on these foundations, the computer vision community has increasingly recognized the value of gaze data: classical saliency models codify attention mechanisms inspired by human vision (Itti and Koch, 2001) (see Borji and Itti (2013) for a review), and gaze has been used to directly guide concrete vision tasks. For example, Shanmuga Vadivel et al. (2015) used multi-viewer eye-tracking to guide video object segmentation, showing that gaze can directly support object localization and extraction. More recently, gaze has been integrated as a supervisory signal in learning-based systems. For instance, Sugano and Bulling (2016) utilized gaze supervision to enhance image captioning models.

Within VQA, Sood et al. (2021) introduced VQA-MHUG, a gaze-annotated dataset designed to

study multimodal neural attention, and Ilaslan et al. (2023) presented GazeVQA, a video QA dataset capturing multiview gaze in task-oriented collaboration scenarios. Building on these directions, our work creates a real-time system that naturally integrates speech, gaze, and vision to resolve ambiguous questions in open-ended VQA. Inadumi et al. (2024) used gaze-target estimations to improve VQA performance, and Karessli et al. (2017) used human gaze embeddings as auxiliary information for zero-shot image classification. However, most prior work leverages gaze during *training* or as estimated gaze target annotations, and model-internal attention is not a reliable substitute for human referential intent. Our work creates a real-time system that naturally integrates human fixations, speech, and visual inputs directly *at inference* to bias VLM reasoning without retraining or architectural changes.

Multimodal Ambiguity Resolution. Ambiguity resolution has been studied across modalities. In NLP, Sukthanker et al. (2018) surveyed approaches to anaphora resolution (identifying what a referring expression refers to in context). In computer vision, Mao et al. (2016) addressed ambiguous referring expressions by generating unambiguous descriptions, and Kazemzadeh et al. (2014) created a game to collect referring expressions for objects in images. At the VQA intersection, Zhang et al. (2016) introduced a balanced dataset for binary ("yes/no") VQA on abstract scenes rather than ambiguous questions, while more recently Testoni et al. (2024) highlighted issues with overconfidence and bias in visual LLMs under ambiguity. Against this backdrop, our approach uniquely combines real-time gaze tracking with open-ended VQA to address ambiguity in a naturalistic setting.

A.8 Ethics statement

We conducted this research in accordance with the ACL Code of Ethics and applicable regulations. This study involving human participants was reviewed and approved by the human subject research protocols (to maintain anonymity, details will be provided upon acceptance). All participants provided informed consent prior to participation and could withdraw at any time without penalty. The consent form explained that: eye movements and voice recordings would be collected during the session, data would be pseudonymized and used for research, participation was voluntary and could be discontinued at any time without penalty, and anonymized data may be shared publicly for research reproducibility, excluding any personally identifiable information such as raw audio recordings.

Recruitment occurred via our university subject pool; participants received course credit. We collected monocular eye-tracking data at 1000 Hz, participants' voices, and mouse click locations for research purposes only. Personally identifying information was not collected; data were pseudonymized before sharing in compliance with institutional and legal requirements. The public release data will include only fixation coordinates and the transcribed text format of the audio to reduce re-identification risks; raw audio or any Personally Identifiable Information (PII) will not be released. Potential risks to participants were minimal and limited to transient discomfort/fatigue; these were mitigated by calibration breaks, screen distance guidelines, and data minimization. The participant sample (primarily students) may limit demographic diversity; we report results with this limitation in mind and caution against over-generalization.

A.9 Reproducibility Statement

To preserve anonymity, we currently release only the practice trials of our VQA benchmark (containing only MS COCO images), as well as semantic embeddings of model responses, and an example gaze-visualization code. The full stimulus image set, VLM responses, real-time interactive VQA demo, and the complete code of the project will be released upon acceptance. Our anonymized supplementary material can be accessed from: https://drive.google.com/file/d/1Ly5WGakEy_1MnjqW2WPjGQ-DVMAg05TS/view?usp=share_link

A.10 LLM Use Disclosure

We used LLMs to correct grammatical mistakes, polish writing, clean software code, and search related works.

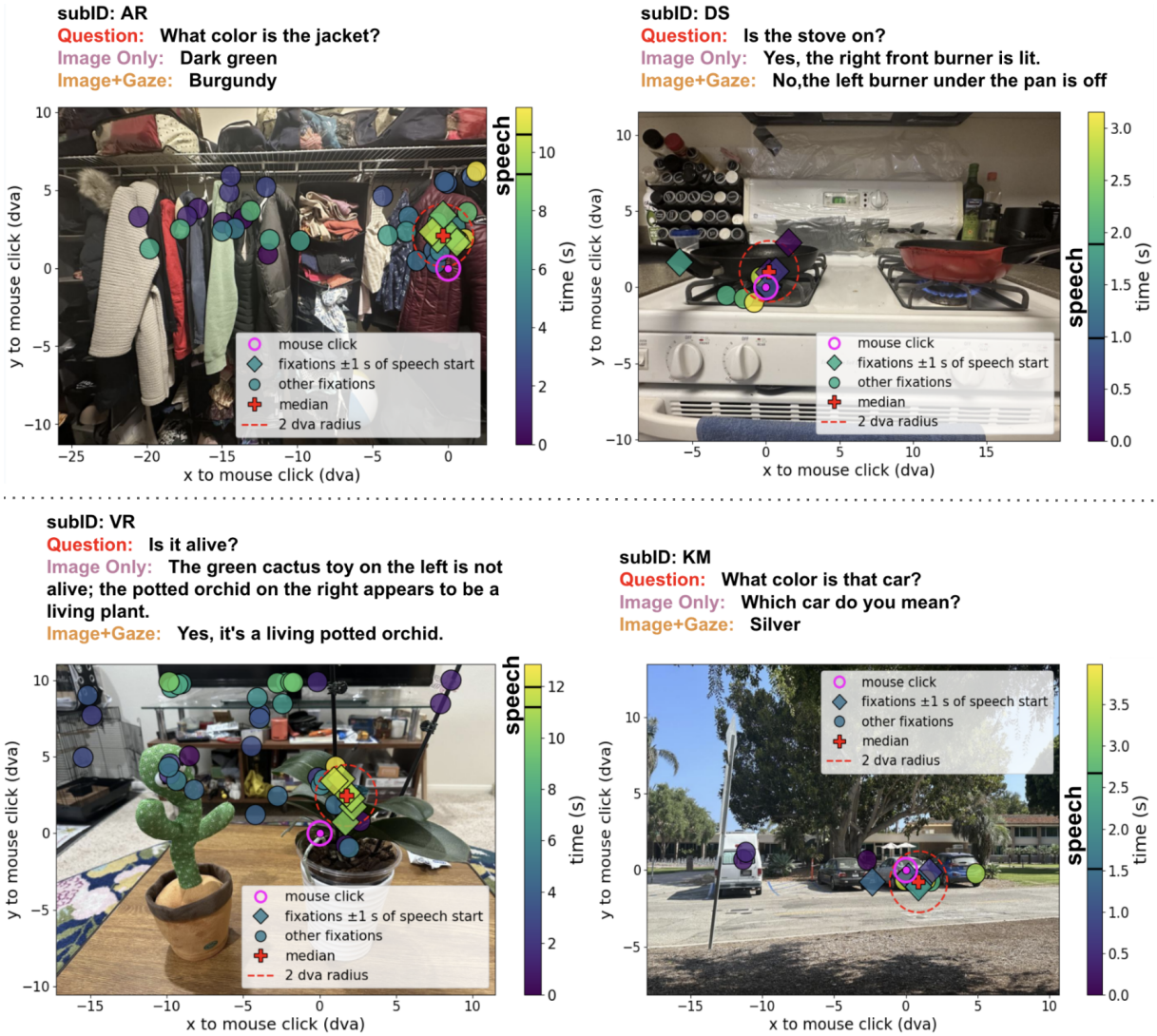

Figure 9: **Visualizing and filtering fixations: examples of "successful" ambiguous trials.** Inclusion of gaze data successfully disambiguated the object in question. The colored circles and diamonds are fixations shaded by their temporal order (yellow - more recent, blue - older). The magenta circle represents the mouse click location registered by the participant at the end of the trial. The red dashed circle indicates a region of 2 dva radius around the median of all fixations within ±1s around speech onset (shown as diamonds). Any diamond within the red dashed circle was sent to the VLM as eye gaze data. The x and y-axes indicate distance to the mouse click location. The times at which the participant started and ended their question are marked as black lines on the color bar. The fixation markers are enlarged for illustrative purposes.

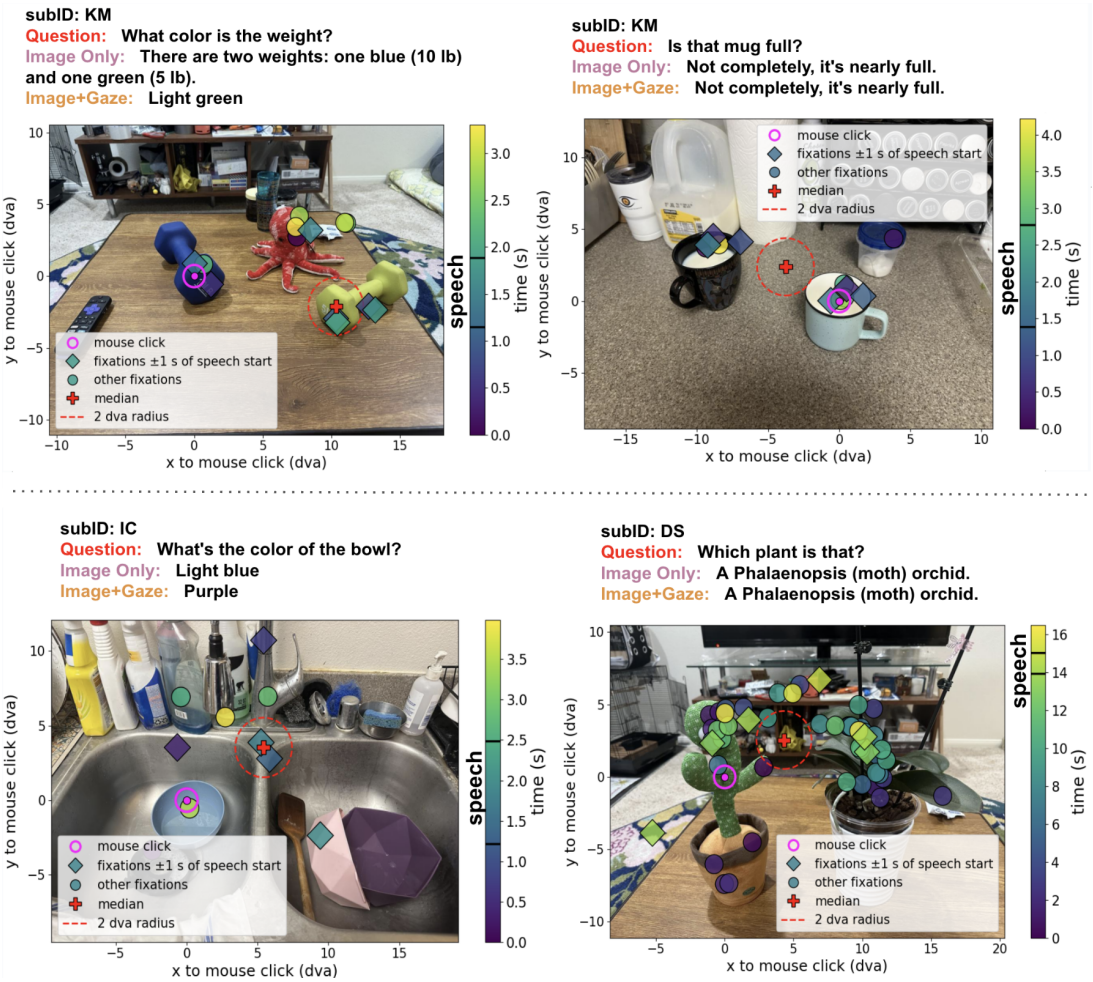

Figure 10: **Visualizing and filtering fixations: examples of "unsuccessful" ambiguous trials.** Inclusion of eye gaze data did not successfully disambiguate the object in these trials. Plotting conventions follow those described in Figure 9.

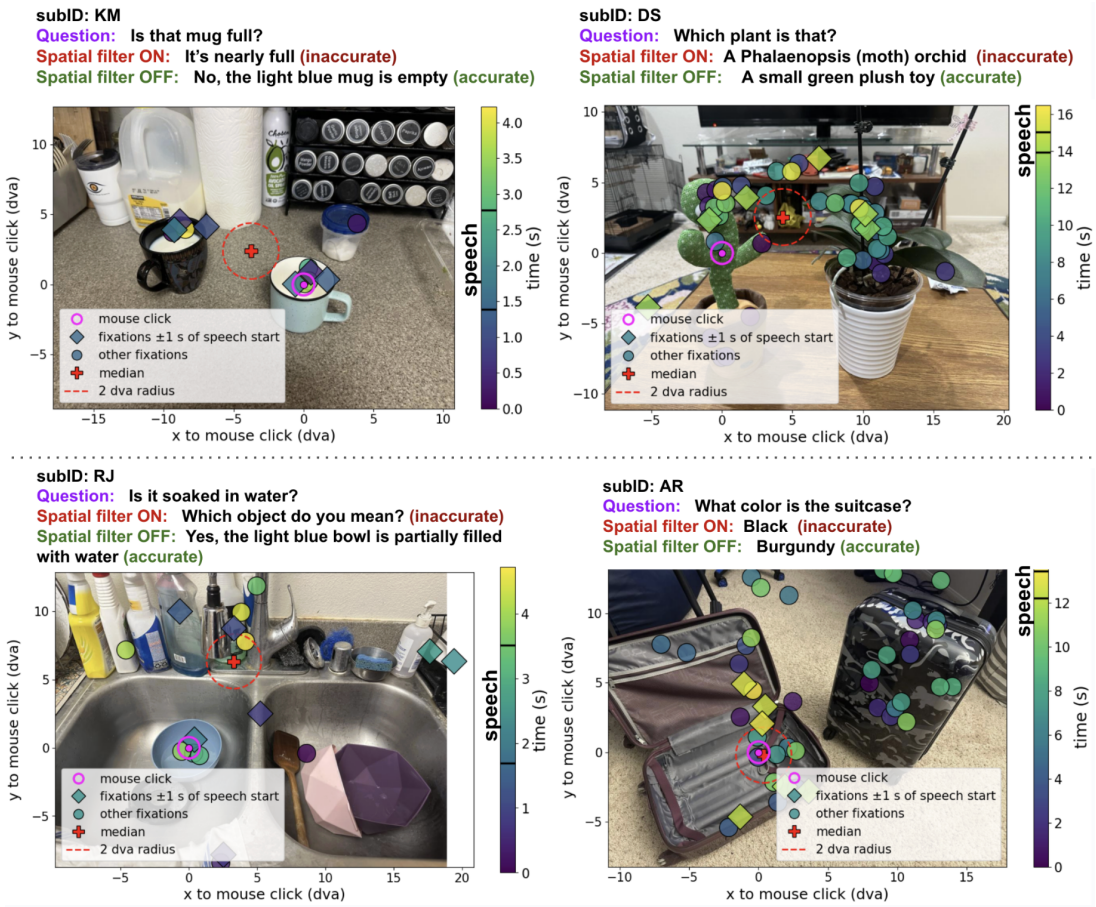

Figure 11: **Effect of spatial filtering.** Visualizing ambiguous Image+Gaze trials where spatial filtering resulted in zero retained fixations. In these examples, removing the spatial filter (but retaining the temporal filter) led to accurate disambiguation (2% trials). In other words, all fixations marked as diamonds were sent to the VLM during inference. Plotting conventions follow those described in Figure 9.

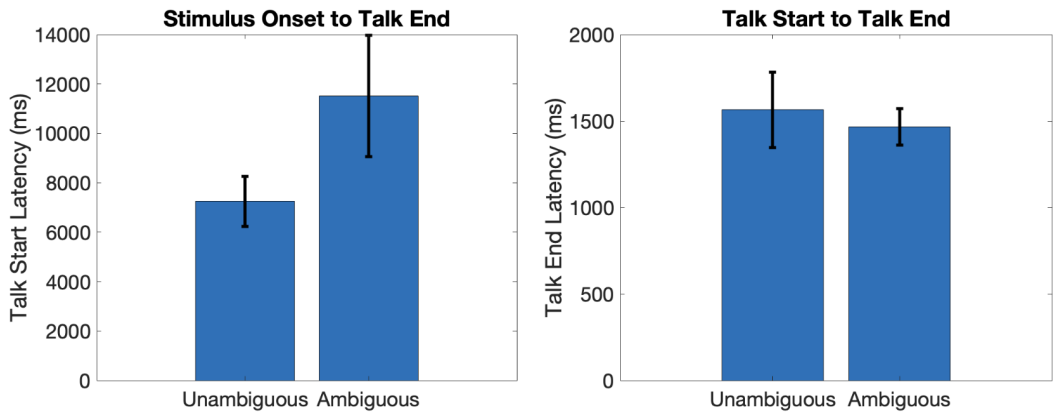

Figure 12: **Participants' speech (questioning) behavior.** The left panel shows the mean time taken by participants to formulate and ask a question from the time of stimulus onset. The right panel shows the mean duration between speech onset and end, as a function of experiment condition. The error bars indicate the standard error of the mean.

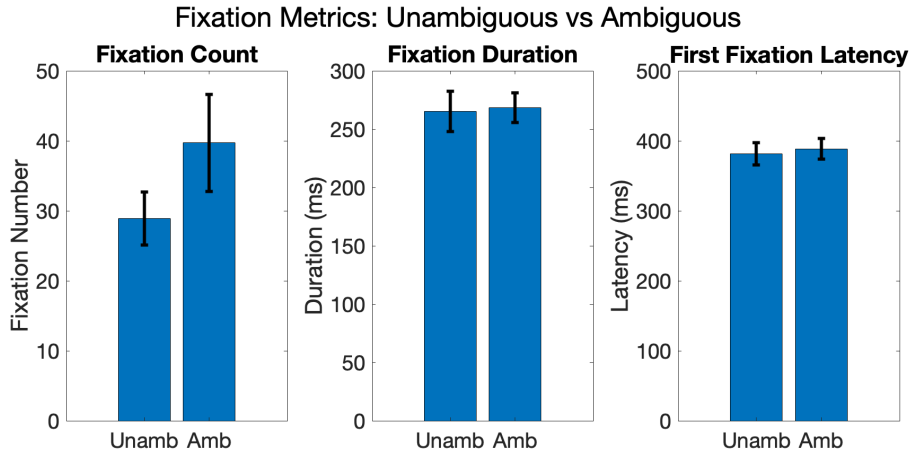

Figure 13: **Fixation metrics.** From left to right, the panels show average fixation count, fixation duration, and first fixation latency for Unambiguous and Ambiguous conditions. The error bars indicate the standard error of the mean.

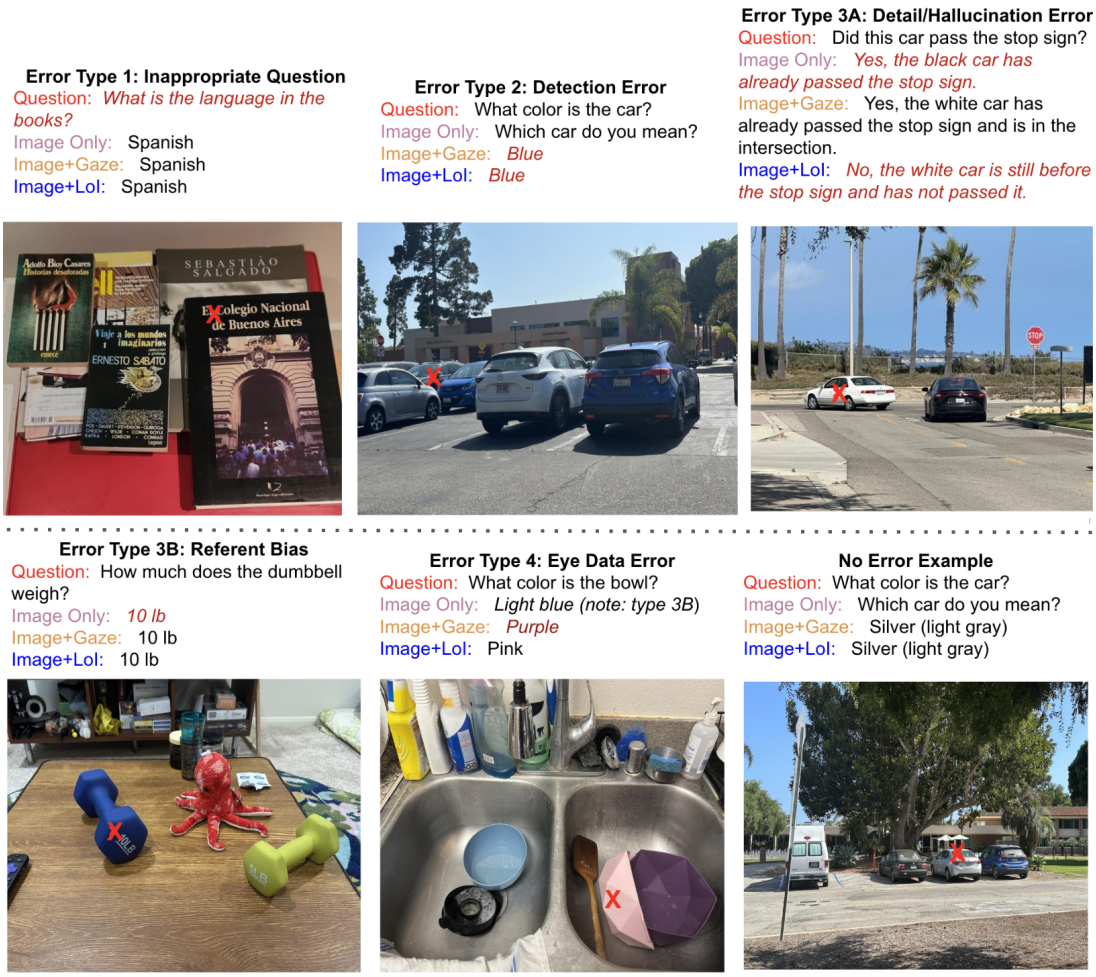

Figure 14: **Types of errors.** Example trials showing Inappropriate Question, Detection Error, Detail/Hallucination Error, Referent Bias, and Eye Data Error (discussed in Section A.4 in detail). The bottom right panel shows an example of a trial where none of the types of errors occurred. The red cross in each image indicates the mouse click location, depicting the object asked about in a given trial.

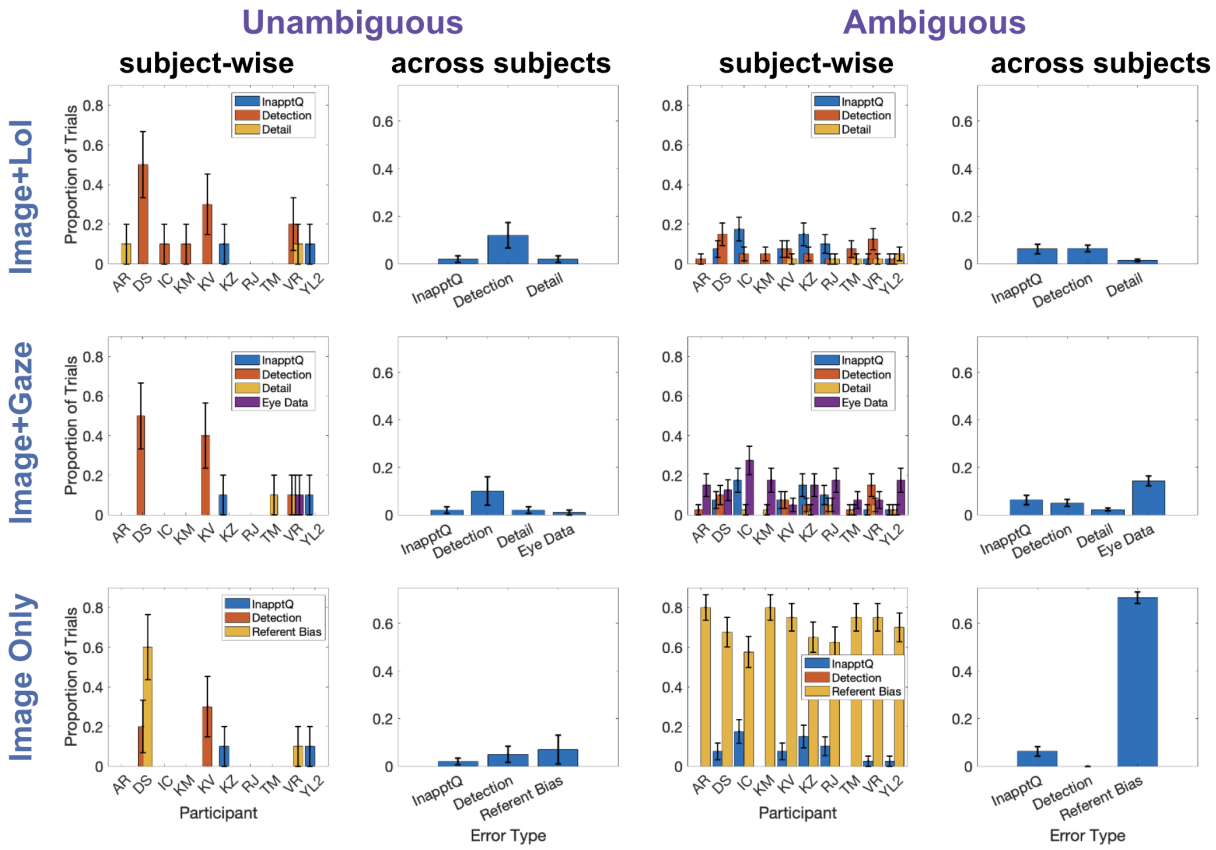

Figure 15: **Quantification of error types.** Proportion of trials that suffered from the error types discussed above and exemplified in Figure 14. The three rows show data from Image+LOI, Image+Gaze, and Image Only input categories from top to bottom, respectively. The four columns show data from Unambiguous and Ambiguous conditions, each separated into subject-wise and across-subject measures. InapptQ - Error Type 1 (Inappropriate Question), Detection - Error Type 2, Detail - Error Type 3A (Detail/Hallucination Error), Referent Bias - Error Type 3B, Eye Data - Error Type 4.