VEGAS: Human-Aligned Video Caption Evaluation via Gaze

Anonymous authors
Paper under double-blind review

Abstract

Vision-language models excel at video captioning, yet typically generate descriptions that fail to capture individual viewers' attention. We propose VEGAS (**V**ideo caption **E**valuation via **GA**ze **S**core), a training-free metric that leverages test-time gaze to sample personalized, attention-aligned text. It is a cross-modal, information-theoretic metric that quantifies how well a candidate caption matches a viewer's focus. To evaluate VEGAS, we curate a dataset of egocentric activities and instructional slides paired with synchronized gaze and reference annotations. We then select captions based on VEGAS via rejection sampling without model retraining. Experiments show that VEGAS-selected captions align significantly better with human focus and improve downstream caption-to-video retrieval, demonstrating the practical utility of incorporating viewer attention during inference.

1 Introduction

In many applications, video captions serve not merely as descriptions but as interfaces to user intent. Vision-language models (VLMs) generate semantically accurate captions, yet they are typically trained on crowd-sourced annotations that aggregate across diverse human interpretations. As a result, they often describe what is broadly visible while ignoring viewer-specific attention and subjective perception. Consider an egocentric setting: two users observing the same kitchen scene may attend to different objects or actions, but a conventional VLM may produce a single globally correct caption that is pragmatically misaligned with what either user intends to retrieve, revisit, or act upon. In caption-indexed retrieval, users often search for videos by referring to attended objects or actions rather than the full scene. When an index caption emphasizes unattended background content, the relevant video can become harder to retrieve.

Gaze provides a measurable, though imperfect, proxy for visual attention Just & Carpenter (1976). We use gaze at test time to select among plausible captions, favoring descriptions that better reflect the viewer's attended referents without retraining or directly modifying the captioning model. Unlike prior work that uses human gaze only as a training-time supervision signal, we leverage gaze at test-time as a conditioning signal to generate personalized captions that reflect an individual's specific attention patterns. We introduce the **V**ideo caption **E**valuation via **GA**ze **S**core (VEGAS), a cross-modal, information-theoretic metric that quantifies how well a caption reflects a viewer's gaze patterns (Figure 1). We then use VEGAS to select gaze-aligned captions via rejection sampling, without requiring model retraining.

To enable controlled evaluation across diverse video stimuli, we curate a multimodal dataset that aligns human gaze signals with visual content and human-written captions. The dataset spans two complementary domains: egocentric videos from the Aria Everyday Activities dataset (AEA) Lv et al. (2024) and instructional slide decks from SlideVQA Tanaka et al. (2023). We choose these two domains because they represent two common gaze-driven scenarios: acting in the world and seeking information from dense visual media. While several prior datasets include video-gaze or video-language annotations, datasets that jointly provide synchronized video, gaze, and captions are scarce; existing examples such as VAS Yu et al. (2017) are relatively old and confined to movie clips. Our dataset supports analysis of attention-driven semantic variation, evaluation of gaze-conditioned captioning, and benchmarking of models that aim to generate personalized, attention-aware video descriptions.

Experiments on our curated dataset provide the strongest support for VEGAS on AEA, where gaze helps resolve concrete object ambiguities. Gaze-conditioned captioning yields a statistically significant increase

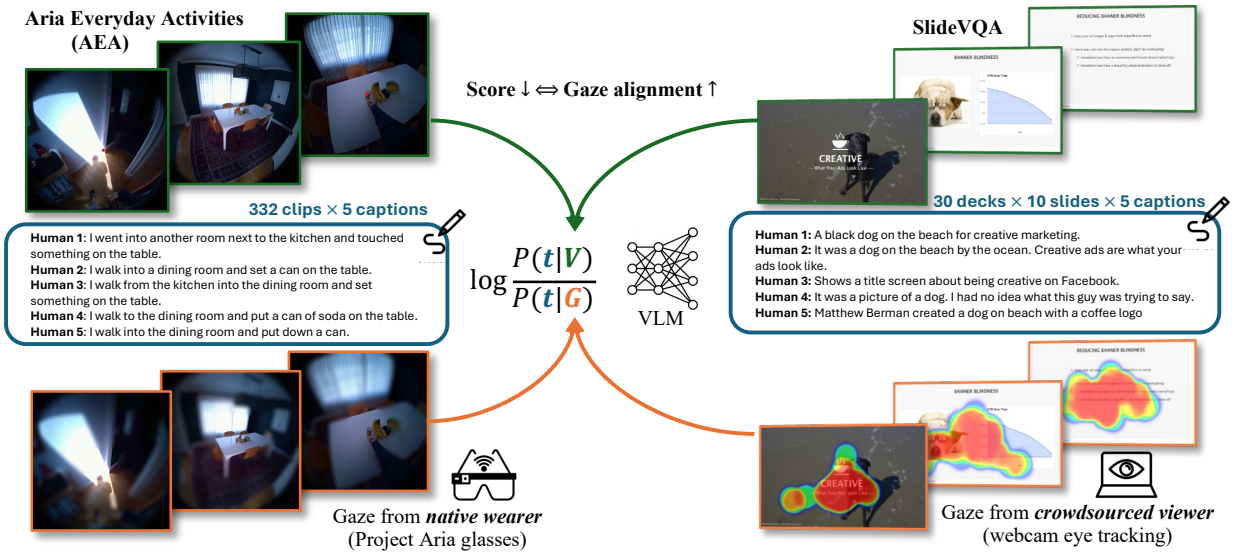


Figure 1: **Overview of our framework.** We curate a multimodal dataset pairing synchronized human gaze and annotated captions across egocentric video clips and instructional slides. VEGAS then leverages VLM's conditional probabilities to evaluate the alignment between a caption and a viewer's gaze.

in mean SBERT similarity (+0.0856), which translates to retrieval gains over conventional VLM-generated descriptions, improving mAP by +1.14%, +2.48%, and +2.46% at ranks 1, 5, and 10, respectively. The SBERT improvement is smaller and non-significant on SlideVQA (+0.0256). This domain variance indicates that VEGAS is most effective when visual attention can be used to disambiguate concrete referents, but less diagnostic when the stimulus supports many valid conceptual summaries from the same visual.

2 Related Works

Captioning with Gaze. Video captioning has traditionally relied on CNN-RNN architectures, such as LRCN Donahue et al. (2015), to model visual and temporal dynamics Aafaq et al. (2019). To enhance localization and relevance, human gaze has been integrated as a supervisory signal in both image Sugano & Bulling (2016) and video captioning Yu et al. (2017). Gaze-based supervision also increasingly supports object detection, action recognition, and skill assessment in egocentric and task-oriented contexts Lall & Liu (2025); Li et al. (2021); Wu et al. (2025), with recent efforts focusing on supervised fine-tuning with gaze annotations Deng et al. (2024). Beyond training-time labels, recent studies also employ human gaze during inference for egocentric video understanding Peng et al. (2025), intent resolution Madinei et al. (2026), and difficulty-controlled text generation Säuberli et al. (2026). While parallel research utilizes saliency prediction as a proxy for empirical gaze or evaluates the alignment between machine and human attention, our work treats gaze as a test-time conditioning signal to select captions aligned with individual attention patterns.

Video-Gaze-Language Datasets. The intersection of visual grounding and human attention has been facilitated by several key datasets that provide synchronized video, gaze, and annotated captions. Early efforts, such as VAS Yu et al. (2017) and Film Gaze Breeden & Hanrahan (2017), integrated movie sequences with human-provided descriptions and gaze traces. Action recognition datasets, *e.g.*, Actions in the Eye Mathe & Sminchisescu (2014) explores gaze patterns across task-specific viewing groups, and GTEA Gaze+ Li et al. (2018) provides fine-grained action annotations for egocentric video. More recently, EgoExo4D Grauman et al. (2024) offers a large-scale multimodal resource featuring egocentric perspectives paired with expert commentary and atomic action descriptions. In contrast, our dataset provides synchronized video, human gaze, and free-form captions across more diverse domains, including egocentric and presentation slides, enabling a broader study of gaze-conditioned language understanding.

Evaluating Captions. Traditional metrics rely on gold-standard references, which are costly to obtain and not always reliable. Metrics, such as BLEU Papineni et al. (2002), METEOR Banerjee & Lavie (2005), and CIDEr Vedantam et al. (2015), quantify the lexical similarity between candidate and reference captions. In contrast, metrics such as SPICE Anderson et al. (2016) and S-BERT Reimers & Gurevych (2019) assess deeper semantic similarities and structural alignment between captions. Recently, cross-modal evaluation approaches propose information-theoretic evaluation by measuring the mutual information between video frames and captions Chen et al. (2026); han Li et al. (2026). Inspired by these approaches, we propose an evaluation methodology for measuring the mutual information between captions and human gaze signals to assess the cognitive plausibility of texts.

3 Problem Formulation

We study the problem of aligning video caption generation with human visual attention during video viewing. Our goal is to control caption generation such that the produced descriptions reflect what a human observer attends to while watching a video. We use gaze signals as a proxy for human cognitive attention, since direct access to humans' internal perceptual or cognitive reasoning is not available. Let V denote a video and G denote the corresponding gaze signal over spatial–temporal regions of V.

We formulate gaze-conditioned caption selection as an information-theoretic objective between gaze signals and generated captions. Specifically, we want to find a set of captions (a set of sequences of tokens $\mathbf{t}$) that maximizes the mutual information, denoted by $\mathbf{I}(\cdot)$, between the gaze signal G.

$$\max_{\mathbf{t}} \;\; \mathbf{I}(\mathbf{t}; G) \;=\; \mathbb{E}_{\mathbf{t},G}\left[\log \frac{P(\mathbf{t}, G)}{P(\mathbf{t})P(G)}\right], \tag{1}$$

where the objective favors captions whose semantics exhibit strong statistical dependence with the viewer's gaze patterns. Here, $P(\mathbf{t}, G)$ denotes the joint distribution over captions and gaze signals, while $P(G)$ and $P(\mathbf{t})$ denote their corresponding marginal distributions under the natural data-generating process. The expectation $\mathbb{E}_{\mathbf{t},G}[\cdot]$ averages over caption–gaze pairs drawn from the joint distribution $P(\mathbf{t}, G)$, i.e., over the co-occurrences of gaze observations and captions.

Intractability. The optimization in (1) is intractable because it requires access to the true marginal distributions $P(G)$ and $P(\mathbf{t})$, which correspond to the natural data-generating processes of human gaze and language. These distributions lack closed-form expressions and cannot be directly evaluated, as they implicitly encode unknown human perceptual and linguistic processes that would require learning full distributions over gaze behavior and caption space, which is essentially the same goal as VLMs and saliency models.

In practice, when we only obtain samples from these distributions, machine learning models approximate the joint behavior through surrogate objectives defined on conditional likelihoods. For example, a pretrained VLM already optimizes a negative log-likelihood objective of the form $-\log P_\theta(\mathbf{t} \mid V)$, which replaces direct modeling of $P(\mathbf{t})$ by learning the conditional distribution of captions given visual input. Similarly, alternative approximations for missing marginals often rely on contrastive or variational objectives, such as InfoNCE-style Oord et al. (2018) losses in representation learning or KL-regularized objectives in variational inference, which replace intractable mutual information terms with tractable upper or lower bounds.

In the next section, we explain how VEGAS approximates the inaccessible Equation (1) by leveraging pretrained VLM likelihoods together with gaze-conditioned visual decomposition, thereby avoiding explicit estimation of the intractable marginal distributions.

4 Video caption Evaluation via GAze Score (VEGAS)

Our framework operates in two stages. First, we introduce the *Video caption Evaluation via GAze Score (VEGAS)*, a cross-modal, information-theoretic metric that quantifies how well a caption reflects a viewer's visual attention as indicated by their gaze (Section 4.1). Second, we use this score to perform gaze-conditioned rejection sampling over candidate captions, selecting those most aligned with the viewer's attentional focus without any model retraining (Section 4.2).

4.1 VEGAS: A Gaze-Aware Caption Evaluation Metric

Setup and Notation. Let V denote the video frames, and let $G \subseteq V$ denote the *gaze-attended* spatial–temporal regions. Operationally, we construct G by preserving regions near the viewer's fixation in each frame and suppressing regions away from fixation. The complement $\bar{G} = V \setminus G$ contains the *non-attended* regions. We assume access to a frozen pretrained VLM that outputs the conditional probability $P_\theta(\mathbf{t} \mid V)$ of generating a token sequence given a video input. Given a candidate caption $c = \mathbf{t} = (t_1, \ldots, t_L)$ of length L, we evaluate its token likelihoods under both the original video V and the gaze-attended video G. VEGAS requires no additional training or fine-tuning.

Intuition. Intuitively, we approximate the intractable distributions $P(\mathbf{t}, G)$, $P(\mathbf{t})$, and $P(G)$ using the VLM next-token distribution $P_\theta(\mathbf{t} \mid V)$, parameterized by the VLM parameters θ. A gaze-aligned caption should remain highly predictable from the attended regions G alone, while the non-attended regions $\bar{G}$ should contribute minimal additional information. VEGAS captures this principle by measuring the contribution of non-attended regions to predicting the caption conditioned on the attended regions.

Also, since we cannot compute the expectation in (1) directly, we instead use the pointwise mutual information $\mathcal{I}(X, Y)$, defined over per data point as the log-ratio $\log \frac{P(X,Y)}{P(X)P(Y)}$. Unlike mutual information, which averages the pointwise mutual information and is non-negative, pointwise mutual information is unbounded and can take both positive and negative values, reflecting local agreement or disagreement between gaze and caption.

Definition. We define *VEGAS* of caption c with respect to gaze G as the pointwise conditional mutual information $\mathcal{I}$ between the caption tokens $\mathbf{t}$ and the non-attended visual regions $\bar{G}$, conditioned on the attended regions G:

$$\mathrm{VEGAS}(c, G) \triangleq \mathcal{I}\big(\mathbf{t}; \bar{G} \mid G\big) = \log \frac{P_\theta(\mathbf{t} \mid G, \bar{G})}{P_\theta(\mathbf{t} \mid G)} = \log \frac{P_\theta(\mathbf{t} \mid V)}{P_\theta(\mathbf{t} \mid G)} \quad \text{(Lower is better)}, \tag{2}$$

where the last equality follows from $G \cup \bar{G} = V$. Intuitively, a *low* VEGAS score indicates that the caption is well-explained by the gazed regions alone (*i.e.*, it is gaze-aligned), as the token sequence $\mathbf{t}$ is well-predicted by the conditioned gaze G relatively to conditioning on the raw video V, and vice versa. In contrast to Equation (1), where higher values indicate stronger gaze alignment, VEGAS captures the information gain in caption prediction when incorporating non-attended regions ($\bar{G} = V \setminus G$), thereby quantifying the extent to which a caption depends on visual content outside the viewer's focus. A lower VEGAS score is therefore preferred, as it indicates that non-attended regions contribute little additional predictive information beyond the attended gaze regions, meaning the caption is primarily driven by what the viewer actually looks at.

Token-level Decomposition. Because the VLM factorizes the caption likelihood autoregressively, VEGAS decomposes over tokens:

$$\mathrm{VEGAS}(c, G) = \sum_{\ell=1}^{L} \log \frac{P_\theta(t_\ell \mid t_{<\ell}, V)}{P_\theta(t_\ell \mid t_{<\ell}, G)}, \tag{3}$$

which allows us to identify which tokens (and, by extension, which semantic concepts) are and are not supported by the viewer's gaze. Namely, it captures the lift in the VLM's per-token confidence when inputting on the full video V versus inputting on the video with the non-attended regions $\bar{G}$ removed.

Properties. VEGAS is (i) *cross-modal*, requiring no human-written reference caption as gold standard and jointly grounding language and vision through the VLM's internal distribution; and (ii) *personalized*, as it evaluates a caption with respect to a specific individual's gaze pattern.

4.2 Gaze-Conditioned Rejection Sampling

Given a pretrained VLM and a viewer's gaze signal G, our goal is to produce a caption that is gaze-aligned without retraining the model. We achieve this via rejection sampling over the VLM's own caption distribution.

We first draw N candidate captions $\{c_1, \ldots, c_N\}$ by sampling from one or multiple VLMs with various temperature and reasoning settings conditioned on the full video V. For each candidate c_i, we compute $\text{VEGAS}(c_i, G)$ using Equation (3). We then select the caption with the *lowest* VEGAS score:

$$c^* = \arg\min_{c_1,\ldots,c_N} \text{VEGAS}(c_i, G),\ \forall i \in [1, N] \quad \text{(VEGAS-based Rejection Sampling)}. \tag{4}$$

The problem in (4) selects the caption whose content is best explained by the viewer's attended regions, effectively filtering out captions that describe unattended content.

Practical Advantages. Selecting captions with VEGAS has several practical advantages. First, it requires *no retraining or fine-tuning* of the VLM; gaze is used at inference time. Second, because the base distribution is the VLM's own caption prior, selected captions inherit the fluency and coherence of the pretrained model. Third, the framework is agnostic to the specific VLM architecture, making it applicable to any model that provides token-level likelihoods conditioned on visual input, even online APIs like GPT and Gemini.

5 Dataset Curation

We curate a multimodal dataset for gaze-conditioned caption selection and evaluation, pairing visual content, synchronized human gaze, and free-form captions to capture distinctive attention-language alignments. Our curation spans two distinct domains: (1) egocentric daily activities using the Aria Everyday Activities (AEA) dataset, and (2) instructional slide presentations using SlideVQA. Participant instructions and user study interfaces are detailed in Appendix A and B.

We recruited human annotators via Prolific (Prolific, 2026), restricting eligibility to individuals aged 18 years or older who reside in the United States and are fluent in English. To ensure accurate data collection, participation was restricted to desktop computer users. For the SlideVQA user study, participants were additionally required to use a webcam to enable remote eye tracking.

5.1 Egocentric Video Captioning Dataset (Aria Everyday Activities)

The AEA subset captures gaze-conditioned language in dynamic, real-world environments. Project Aria smart glasses provide the first-person video streams alongside hardware-synchronized human gaze measurements natively recorded from the wearer.

Video Segmentation. We first segment raw egocentric recordings into shorter, action-centric clips to isolate discrete tasks. We leverage `Gemini-3.1-Pro-Preview` to identify boundaries, prompting the model to output precise start and end timestamps along with a short, first-person summary beginning with "I..." in the simple present tense. This minimizes semantic ambiguity and establishes a tighter temporal mapping between visual attention and language. The Gemini-generated segmentation summaries are used only for boundary proposals. They are not shown to human annotators and has no effect on human caption quality.

Video Clip Selection. Our goal is to evaluate on gaze-sensitive clips where personalization is expected to matter, namely segments where access to gaze substantially changes the generated caption. To identify such segments, we run captioning inference with `Gemini-3.1-Pro-Preview` under two different settings: (1) using gaze-conditioned inputs, where we mask frame regions centered at fixation points, and (2) using the raw video without gaze information. We measure the semantic divergence between these pairs using SBERT similarity and evaluate them via the VEGAS metric using `Gemma-4-31B-IT`. We retain the top 25% of segments exhibiting the lowest SBERT similarity and highest VEGAS scores, yielding 332 highly gaze-dependent video clips for human annotation.

Caption Annotation. Using the Qualtrics XM Platform (Qualtrics, 2024), recruited annotators viewed each segmented clip overlaid with a synchronized gaze visualization (indicated by a red dot). We instructed participants to write a single-sentence caption exceeding 30 characters that explicitly describes the action emphasized by the wearer's visual attention. This protocol enforces target-driven descriptions rather than a passive, exhaustive narration of the entire background scene.

5.2 Slide Deck Captioning Dataset (SlideVQA)

The SlideVQA subset contains instructional and educational slide decks spanning diverse topics and visual layouts. Since SlideVQA does not provide gaze annotations, we collect viewer attention data through a controlled viewing study. We select 30 slide decks from the dataset and present the first 10 slides of each deck to participants for captioning and gaze recording. We recruit 5 participants per slide deck. Each participant views a slide sequence and provides a caption per slide while we record their gaze behavior.

Data Collection. We recorded participant gaze trajectories using the webcam-based eye-tracking model on the RealEye platform (RealEye, 2025). The platform natively processes visual streams locally on the user's device, capturing and storing only anonymized, timestamped coordinate vectors without saving any raw camera video. Participants were instructed to view each slide naturally for a minimum of 20 seconds before proceeding to the annotation phase. Directly after viewing a slide, participants compose a free-form caption between 30 and 200 characters detailing the prominent text blocks, visual elements, or core semantic concepts that anchored their attention.

Data Processing. Gaze coordinates were normalized relative to the native slide resolutions and aggregated into continuous spatial attention maps. RealEye provides operational quality scores ranging from 1 (*Very Low*) to 6 (*Perfect*)[1]. For our experiments, we enforce two quality filters: (1) we exclude participants with a RealEye quality grade below 3, and (2) we purge participants who submitted duplicate text responses across different slides within a single deck. Ultimately, 142 out of 150 total participants were retained, yielding an average calibration score of 4.88.

5.3 Dataset Statistics

Our final curated dataset comprises 632 total visual samples paired with 2,981 high-quality, human-written captions (see Table 1). The egocentric AEA subset (332 action clips derived from 91 unique parent videos, totaling 46.78 minutes) captures everyday tasks like cooking, computing, and object manipulation. The presentation-based SlideVQA subset (300 individual slides across 30 decks) spans dense academic, business, and scientific layouts. Captions average 14.8 ± 5.6 words for AEA and 16.0 ± 7.1 words for SlideVQA. The notable variation in reading behaviors and linguistic styles between the two subsets underscores the value of cross-domain gaze-aware evaluation.

Table 1: Statistics for the Aria Everyday Activities (AEA) and SlideVQA sub-datasets. Caption statistics are reported for human-written annotations (after quality filtering for SlideVQA).

Metric	**AEA (Egocentric)**	**SlideVQA (Presentation)**
Number of Parent Source Units (Videos / Decks)	91	30
Number of Annotated Samples (Clips / Slides)	332	300
Total Retained Human Captions	1,660	1,321
Mean Caption Length (Words)	14.8 ± 5.6	16.0 ± 7.1

5.4 Ethical Considerations

We recognize that gaze trajectories can reveal sensitive information regarding individual cognitive patterns and implicit user intent. For SlideVQA, participants were informed during consent that gaze would be collected and reused for research. We will publish both raw gaze traces and processed attention heatmaps to enable reproducibility. To mitigate privacy risks, all personally identifiable information (PII) in metadata and annotations was stripped during post-processing. For AEA, the gaze data are used directly from the source dataset, whose videos were already processed with EgoBlur (Raina et al., 2023) to automatically detect and

[1]RealEye assigns a quality grade from 1 (*Very Low*) to 6 (*Perfect*) based on eye-tracking sampling rate, fixation computability, percentage of items with gaze data, total gaze data length, and gaze-on-screen coverage. Grades ≥ 3 (*Average* and above) achieve a sampling rate of at least 10 Hz and $\geq 30\%$ gaze-on-screen coverage.

obscure PII such as faces and license plates. All data collection and annotation procedures are approved by the Institutional Review Board (IRB) and follow institutional ethical guidelines.

6 Experiments

Candidate Caption Generation. We generate candidate video captions using multimodal large language models spanning closed-source models (`Gemini-3.1-Pro-Preview` and `GPT-5.5`) and an open-source model (`Qwen3.6-35B-A3B`). All models use the same candidate-captioning prompt, provided in Appendix D. Candidate generation uses original visual inputs: for AEA, we sample frames from the original extracted clip videos at 1 fps; for SlideVQA, we provide the slide images directly. For each sample, we generate one caption from each model at decoding temperatures in $\{0.0, 1.0\}$, yielding up to $N = 6$ candidates per sample before post-processing (see Appendix E). We choose these models because they are among the strongest performers on multimodal understanding tasks and provide a diverse pool of caption candidates.

VEGAS Scoring. For VEGAS evaluation, we deploy `Gemma-4-31B` locally to score captions in a controlled setting, avoiding API-side model updates and serving variability. We set the model temperature to $T = 1$ to score captions under the model's native, untempered next-token distribution. The random seed only affects token sampling, not the underlying next-token probabilities, and therefore does not alter the model's intrinsic likelihood estimates. Per Equation (3), throughout the experiments, we retain only the top-20 token probabilities at each decoding step and use a fallback log-probability of -30 for tokens outside this set. Appendix H ablates these choices and an alternative Qwen2.5-VL-32B-Instruct scorer, showing that downstream SBERT and retrieval performance remain stable across the implementation grid. Additional details on compute resources and caption post-processing are provided in Appendix F and E, respectively.

Semantic Alignment Protocol. We evaluate whether lower-VEGAS captions are semantically closer to human annotations using SBERT similarity to each sample's human-caption centroid. First, we compare naive Gemini captions against VEGAS-selected captions and test paired similarity differences with a Wilcoxon signed-rank test. For AEA, the naive Gemini caption is the first-person clip summary produced during video segmentation; for SlideVQA, it is the Gemini caption generated directly from the slide image. We then control for candidate-pool quality by comparing selectors within the same caption pool: random selection, raw VEGAS, length-normalized VEGAS, and a best-in-pool oracle. The length-normalized score is defined as $\text{VEGAS}_{\text{norm}} = \text{VEGAS}/L$, where L is the number of scored tokens, and the best-in-pool oracle selects the caption with highest SBERT similarity to the human-caption centroid. Finally, we run two diagnostics: a gaze-corruption ablation to test whether VEGAS depends on the correct gaze signal, and a caption-length analysis to test whether improvements are explained by shorter captions.

Caption-to-Video Retrieval Protocol. To evaluate whether VEGAS turns wearer attention into a better semantic index, we perform caption-to-video retrieval on AEA. Each query is a human-written caption describing what the camera wearer attended to in a clip, and the retrieval index is constructed from captions automatically generated from the candidate videos. The candidate captions are pooled across three VLMs (`Gemini-3.1-Pro-Preview`, `GPT-5.5`, and `Qwen3.6-35B-A3B`) and two decoding temperatures ($T = 0$ and $T = 1$). Candidate captions are pooled across the three VLMs and two decoding temperatures. We compare random VLM caption indexing, VEGAS-selected VLM caption indexing, and a human pairwise reference in which one human caption queries another human caption for the same video. Retrieval metrics are averaged over human-query runs and reported as Precision@K, Recall@K, and mAP@K.

7 Results

How diverse are human captions for identical visual content? To quantify human annotation diversity, we embed the human captions for each AEA clip using `sentence-transformer/all-mpnet-base-v2` Song et al. (2020). For each clip, we fit a local 3-dimensional PCA to the human-caption embeddings, center the coordinates at the human-caption centroid, and pool the standardized coordinates across clips. We use the same encoder for all SBERT-based analyses in this section.

Across 332 AEA clips, the local PCA retains 89.29% of human-caption variance on average. In this PCA space, the mean pairwise human-human distance is 1.53, compared with a mean Gemini-human distance of 1.07 and a mean Gemini–centroid distance of 0.48. As shown in Figure 2, naive Gemini captions lie close to the human consensus centroid but cover only a small portion of the semantic variation among human annotations. This indicates that human captions for the same visual content are substantially more diverse than generic VLM captions, motivating personalized captioning methods that can reflect individual perceptual focus.

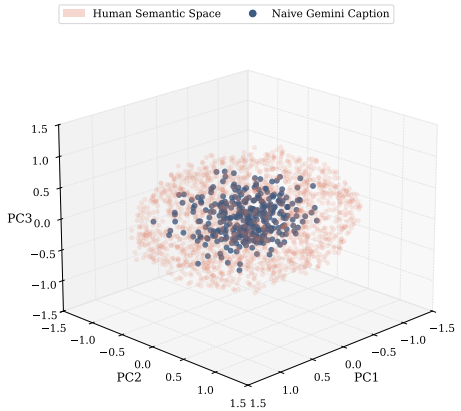


Figure 2: **PCA visualization.** Human captions are much more diverse than naive Gemini captions in semantic space.

Are individual differences in gaze patterns reflected in human captions? Given this high diversity, we evaluate whether individual spatial gaze patterns can act as a reasonable proxy to guide personalized captioning. In Figure 3 (center panel), the x-axis represents the average fixation spatial distribution distance between participant pairs, and the y-axis represents the average SBERT caption similarity. We observe a significant negative Pearson correlation ($r = -0.23, p < 0.001$), indicating that as visual gaze patterns diverge spatially, their written captions also diverge semantically. This directly demonstrates that individual visual attention patterns are significantly but partially associated with their semantic interpretation of visual content, validating gaze as a reasonable proxy for personalized human captions.

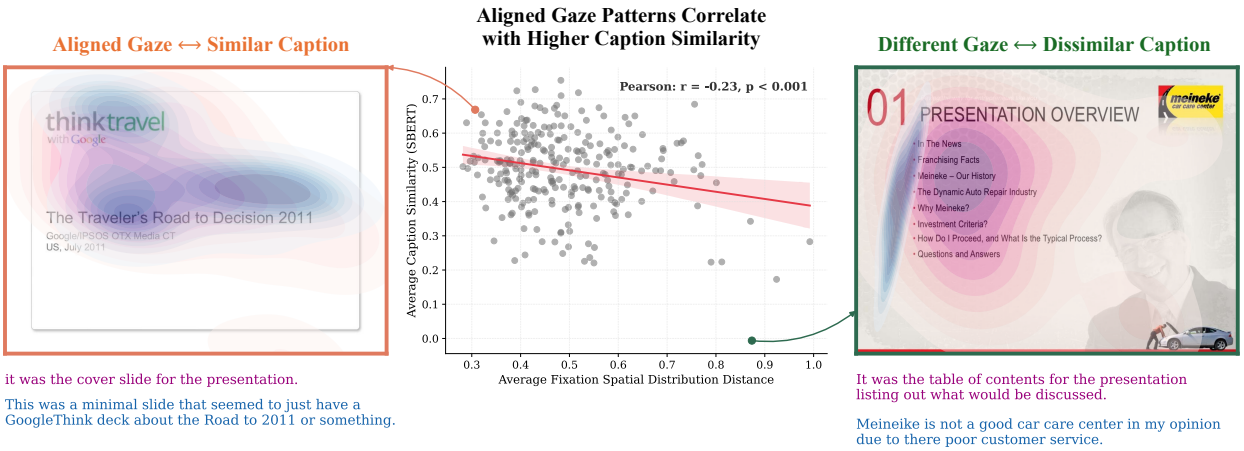


Figure 3: **Gaze-caption personalization in SlideVQA.** The center panel displays a scatter plot of average caption semantic similarity (SBERT) against average fixation spatial distribution distance between participant pairs. The left and right panels present qualitative slides where two participants (one in pink, the other in blue) exhibit distinct visual gaze distributions on identical slides.

Our method uses this gaze-language association to personalize caption selection. Figure 4 shows a qualitative example where the visual input is fixed, but the gaze pattern changes; VEGAS correspondingly assigns lower scores to different candidate captions whose tokens better match the regions attended by each viewer.

Are captions with lower VEGAS semantically closer to human captions? For AEA, the naive Gemini captions are the first-person clip summaries produced during the video-segmentation step in Section 5.1. As shown in Figure 5, captions selected by minimizing VEGAS (light blue) shift toward higher SBERT similarity relative to these naive Gemini captions (dark blue), increasing mean similarity by +0.0856 (+13.53%) on the $[-1, 1]$ SBERT scale. This rightward shift is statistically significant under a Wilcoxon signed-rank test ($Z = 10.04, p < 0.001$), indicating that VEGAS selection improves semantic alignment over generic segmentation summaries in egocentric activity videos, where gaze helps identify the actions and objects most

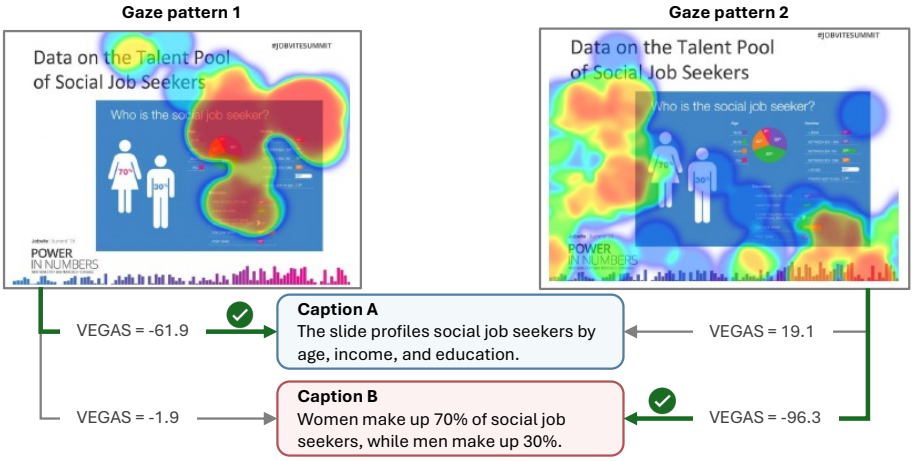


Figure 4: **VEGAS scoring in SlideVQA.** For the same slide and candidate captions, VEGAS assigns lower scores to captions whose tokens align with each viewer's attended regions.

relevant to human annotators. On SlideVQA, VEGAS selection produces a smaller positive mean shift of +0.0256 (+3.88%), but this difference is not statistically significant ($Z = 1.67, p = 0.0952$). We therefore interpret the SlideVQA result cautiously: because much of the slide text is directly visible, generic VLM captions can already match one plausible human description, while different annotators may still emphasize different valid aspects of the same slide. This makes SlideVQA a noisier target for single-reference semantic similarity and reinforces the value of human annotations that preserve diverse, attention-driven interpretations.

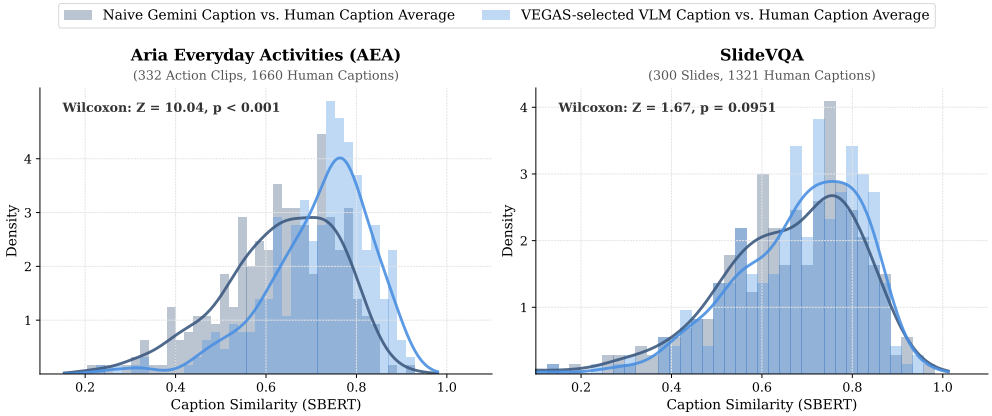


Figure 5: **Density distribution of SBERT similarity.** The x-axis denotes caption semantic similarity (SBERT) relative to the average human caption, and the y-axis shows empirical probability density. We compare naive Gemini captions (dark blue) against VEGAS-selected captions (light blue) from the pooled `Gemini-3.1-Pro-Preview`, `GPT-5.5`, and `Qwen3.6-35B-A3B` candidates. VEGAS selection shifts mean SBERT similarity by +0.0856 (+13.53%) on AEA and +0.0256 (+3.88%) on SlideVQA.

We next compare selectors within the same VLM candidate pool (Table 2). On AEA, raw VEGAS improves over random selection by +0.013 SBERT ($p < 0.05$), while length-normalized VEGAS shows a smaller positive trend. On SlideVQA, VEGAS is tied with random selection, consistent with slides admitting multiple valid semantic interpretations and providing a weaker gaze-conditioned signal. The best-in-pool oracle remains substantially higher in both domains, indicating room for stronger future selection objectives.

Table 2: **Selection from the same candidate pool.** Each selector chooses one caption per item from the same pool. SBERT is mean ± standard deviation to the human-caption centroid. Δ is relative to Random; significance uses a paired Wilcoxon test ($^*p < 0.05$, $^{**}p < 0.01$).

Domain	n	Selector	SBERT	Δ vs Random
AEA	332	Random (seed=42)	0.705 ± 0.124	–
		VEGAS	0.718 ± 0.115	$+0.013^*$
		VEGAS / token	0.714 ± 0.117	$+0.009$
		Best-in-pool	0.800 ± 0.075	$+0.095^{**}$
SlideVQA	300	Random (seed=42)	0.685 ± 0.139	–
		VEGAS	0.685 ± 0.138	-0.000
		VEGAS / token	0.685 ± 0.137	$+0.000$
		Best-in-pool	0.742 ± 0.114	$+0.057^{**}$

To illustrate VEGAS at the token level, Figure 6 shows an AEA example where the attended object disambiguates the best caption from generic and unrelated alternatives. An additional example is provided in Figure 12 of Appendix G.

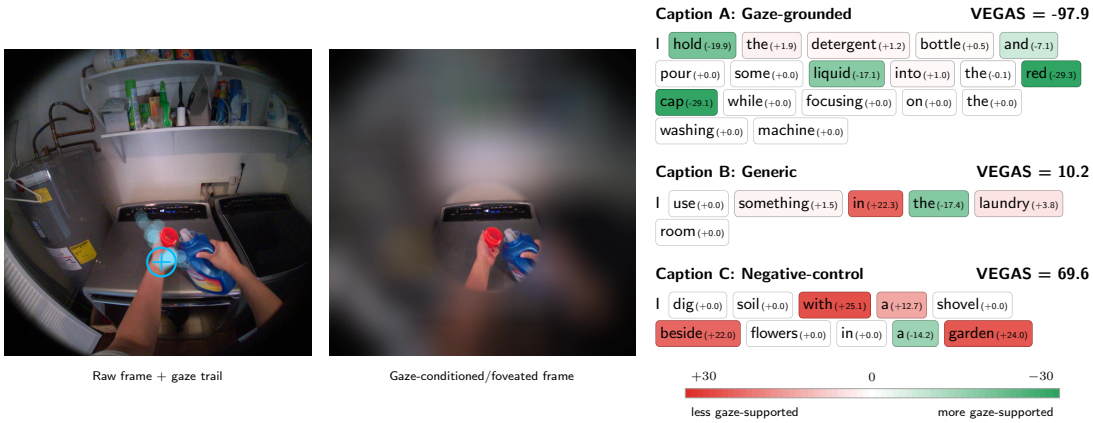


Figure 6: **AEA qualitative example.** VEGAS assigns the lowest score to the caption aligned with the attended object, compared with a generic caption and an unrelated negative control. Token scores show that terms tied to gaze-attended regions such as "liquid" and "red cap" in Caption A tend to be more supported (green), while unrelated terms like "garden" in Caption C are less supported (red).

In the same appendix section, we also include two SlideVQA bad cases in Figure 13, where reasonable captions receive high VEGAS scores. These examples help explain the weaker quantitative gains on SlideVQA: unlike AEA, where captions often depend on grounding concrete objects, slide captioning frequently requires conceptual abstraction. Gaze indicates where a viewer looked, but may not fully determine how the viewer interprets chart trends, causal relationships, or high-level concepts.

Is VEGAS simply selecting shorter captions? Because VEGAS is computed as a token-level sum, a natural concern is that it may simply prefer shorter or more generic captions rather than captions that better align with gaze. We therefore run a post-hoc diagnostic over the full VLM candidate pools used for rejection sampling. We correlate VEGAS with caption length, measured by word count, and caption specificity, approximated by POS-tag counts such as nouns, verbs, adjectives, numeric/entity-like tokens, and content words. As shown in Table 3, both correlations are near zero in AEA and SlideVQA, providing no evidence that VEGAS is driven by brevity or genericness. This can also be seen in Table 2 where raw VEGAS improves SBERT over random selection ($+0.013$, $p = 0.040$), while length-normalized VEGAS shows a smaller, non-significant gain ($+0.009$, $p = 0.183$) in AEA, suggests that improvements are not explained by caption length alone.

Table 3: **Caption length and specificity vs. VEGAS.** Pearson correlations over raw VLM candidates show near-zero association with VEGAS in both domains.

Domain	n	Word count r (p)	Specificity r (p)
AEA	1814	0.001 (0.979)	0.026 (0.275)
SlideVQA	1578	−0.014 (0.580)	−0.014 (0.579)

Is VEGAS sensitive to gaze corruption? To test whether VEGAS is sensitive to the correct gaze signal, we compare each AEA caption's score under true gaze against two corrupted gaze conditions: randomly mismatched gaze from another clip and a center-bias gaze surrogate. As shown in Table 4, random gaze increases VEGAS by +1.786 points on average relative to true gaze, with a 95% bootstrap confidence interval of $[+0.365, +3.289]$ and a paired Wilcoxon test of $p = 0.040$. This indicates that mismatching a caption with another clip's gaze measurably worsens gaze-caption alignment. Center-bias gaze also increases the mean score by +1.412 points, but the confidence interval crosses zero ($[-0.181, +3.019]$) and the Wilcoxon test is marginal ($p = 0.058$). Overall, true AEA gaze contains measurable information beyond a randomly mismatched gaze baseline, while the weaker separation from center-bias gaze suggests partial overlap between true gaze and coarse central salience in egocentric video.

Table 4: **Gaze corruption ablation on AEA.** Lower VEGAS is better. Δ denotes the paired score difference relative to true gaze, $\mathrm{VEGAS}(c, G_{\mathrm{condition}}) - \mathrm{VEGAS}(c, G_{\mathrm{true}})$. Positive Δ therefore indicates worse gaze-caption alignment than the true gaze condition.

Condition	n	Mean VEGAS	Mean Δ	95% CI for Δ	Wilcoxon p
True gaze	327	47.632	–	–	–
Random gaze	327	49.419	+1.786	$[+0.365, +3.289]$	0.040
Center-bias gaze	327	49.044	+1.412	$[-0.181, +3.019]$	0.058

Does VEGAS improve downstream caption-to-video retrieval? To evaluate whether VEGAS turns wearer attention into a better semantic index, we perform caption-to-video retrieval on AEA. Each query is a human-written caption describing what the camera wearer attended to in a clip, and the retrieval index is constructed from captions automatically generated from the candidate videos. The candidate captions are pooled across three VLMs (`Gemini-3.1-Pro-Preview`, `GPT-5.5`, and `Qwen3.6-35B-A3B`) and two decoding temperatures ($T = 0$ and $T = 1$).

We compare three index-construction strategies. The first uses a randomly selected VLM caption from the pooled captions. The second, VEGAS (Ours), selects the lowest-VEGAS caption from the same pool. The third is a human pairwise reference: one human caption is used as the query, and a different human caption for the same video is used as the single human-authored index entry. Results are averaged over all ordered pairs of distinct annotators, measuring human-human retrieval agreement.

As shown in Table 5, VEGAS (Ours) consistently improves over random VLM captions across all reported precision, recall, and mAP metrics. The gains are modest at rank 1 (+1.14% for Precision/Recall/mAP@1) but become more pronounced at larger retrieval depths, improving Recall@5 by +4.16%, Recall@10 by +4.10%, mAP@5 by +2.48%, and mAP@10 by +2.46%. Measured as random-to-human gap closure, VEGAS recovers 15.0% of the Rank@1 gap, but a substantially larger share at broader retrieval depths: 40.8% for Recall@5, 42.3% for Recall@10, 27.8% for mAP@5, and 27.8% for mAP@10. This pattern suggests that VEGAS most strongly improves the semantic neighborhood of the retrieved results: selected captions often move the correct video into the top few candidates, even when they are not yet discriminative enough to make it the top-ranked result. These results suggest that lower-VEGAS captions are better aligned with the descriptions humans naturally use to search for videos. The human pairwise row shows the remaining gap between machine-selected and human-written index captions.

Table 5: **Human-query retrieval results on AEA.** We encode human annotations with SBERT and average retrieval metrics over human-query runs. Candidate indices contain one caption per video: a random VLM caption, the lowest-VEGAS caption selected from pooled VLM captions (`Gemini-3.1-Pro-Preview`, `GPT-5.5`, and `Qwen3.6-35B-A3B`), or a human caption in an ordered pairwise setup where annotator i queries annotator j with $i \neq j$. VEGAS (Ours) improves over random VLM captions across metrics.

Candidate Source	Precision@K			Recall@K			mAP@K		
	@1	@5	@10	@1	@5	@10	@1	@5	@10
Random VLM	22.41%	9.67%	6.03%	22.41%	48.37%	60.30%	22.41%	31.73%	33.31%
VEGAS (Ours)	**23.55%** (+1.14%)	**10.51%** (+0.84%)	**6.44%** (+0.41%)	**23.55%** (+1.14%)	**52.53%** (+4.16%)	**64.40%** (+4.10%)	**23.55%** (+1.14%)	**34.21%** (+2.48%)	**35.77%** (+2.46%)
Human Pairwise	30.02%	11.71%	7.00%	30.02%	58.57%	69.98%	30.02%	40.64%	42.19%

8 Limitations

VEGAS inherits the limitations of the underlying VLMs, including hallucinations, perceptual errors, and imperfect token likelihood calibration. If a VLM produces similar hallucinated interpretations with and without gaze conditioning, VEGAS may assign overly favorable scores. A second limitation is that gaze reveals where a viewer attends, but not always how they interpret what they see. This matters for domains where captioning may require conceptual abstraction, information aggregation, or reasoning across visual elements. Thus, VEGAS may be most effective when attention is closely tied to concrete objects, actions, or entities. Finally, VEGAS requires gaze annotations. Although webcam-based eye tracking reduces collection cost, it still requires controlled data acquisition and participant involvement. In deployed systems, explicit gaze may be available in smart glasses, while implicit attention proxies may be more realistic for web and video platforms.

9 Conclusion

Existing VLMs generate video captions that largely ignore individual viewer attention, despite the fact that human video descriptions are inherently diverse and closely tied to where people look. Motivated by this observation, we introduce VEGAS, a cross-modal, information-theoretic metric that measures how well a video caption aligns with a viewer's gaze, along with a VEGAS-based rejection sampling method that uses gaze as a proxy for viewer intent to select attention-aware captions from candidate VLM outputs. This approach operates entirely at inference time and requires no retraining of the underlying vision-language model. By re-ranking and selecting among candidate captions generated by existing VLMs, VEGAS turns viewer attention into a better semantic index for downstream retrieval and attention-aware interaction.

To evaluate this framework, we curate a multimodal dataset pairing gaze trajectories, videos, and human-written captions across egocentric and instructional slide domains. On AEA, VEGAS-selected captions significantly improve semantic alignment with human annotations and yield modest but consistent gains in downstream caption-to-video retrieval. On SlideVQA, the gains are smaller and not statistically significant, suggesting that gaze is a weaker selector when captions require conceptual abstraction rather than grounding concrete attended objects or actions. Together, these results show that test-time gaze can provide a practical mechanism for viewer-aware caption selection, while also highlighting that its benefits depend on how directly visual attention maps onto semantic interpretation.

Future work. Looking ahead, VEGAS opens several directions. Future work could improve scalability by replacing explicit gaze with saliency models, interaction traces, or other implicit attention proxies, and by extending VEGAS to longer videos and real-time personalization. Robustness could be improved through external verification, grounding constraints, or calibrated scoring models that reduce sensitivity to VLM hallucinations and likelihood miscalibration. VEGAS could also be distilled into captioning models or reformulated as a differentiable training objective, enabling applications such as personal memory retrieval, assistant question answering, and learning-support systems.

References

Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, 52(6):1–37, 2019.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pp. 382–398. Springer, 2016.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Katherine Breeden and Pat Hanrahan. Gaze data for the analysis of attention in feature films. *ACM Transactions on Applied Perception*, 2017.

Shenghui Chen, Po han Li, Sandeep P. Chinchali, and Ufuk Topcu. VIBE: Annotation-free video-to-text information bottleneck evaluation for TL;DR. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL `https://openreview.net/forum?id=C35FCYZBXp`.

Shuwen Deng, Paul Prasse, David Reich, Tobias Scheffer, and Lena Jäger. Fine-tuning pre-trained language models with gaze supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 217–224, 2024.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634, 2015.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Po han Li, Shenghui Chen, Ufuk Topcu, and Sandeep Chinchali. Visil: Unified evaluation of information loss in multimodal video captioning, 2026. URL `https://arxiv.org/abs/2601.09851`.

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

Vishakha Lall and Yisi Liu. Eyes on target: Gaze-aware object detection in egocentric video. *arXiv preprint arXiv:2511.01237*, 2025.

Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 619–635, 2018.

Yin Li, Miao Liu, and James M Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):6731–6747, 2021.

Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, Kiran Somasundaram, Luis Pesqueira, Mark Schwesinger, Omkar Parkhi, Qiao Gu, Renzo De Nardi, Shangyi Cheng, Steve Saarinen, Vijay Baiyya, Yuyang Zou, Richard Newcombe, Jakob Julian Engel, Xiaqing Pan, and Carl Ren. Aria everyday activities dataset, 2024.

Parsa Madinei, Srijita Karmakar, Russell Cohen Hoffing, Felix Gervitz, and Miguel P Eckstein. Iris: Intent resolution via inference-time saccades for open-ended vqa in large vision-language models. *arXiv preprint arXiv:2602.16138*, 2026.

Stefan Mathe and Cristian Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(7): 1408–1424, 2014.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. *arXiv preprint arXiv:2509.07447*, 2025.

Prolific. Prolific: Online participant recruitment platform, 2026. URL `https://www.prolific.com`. Online participant recruitment platform.

Qualtrics. Qualtrics xm platform, 2024. URL `https://www.qualtrics.com`. Computer software.

Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, Prince Gupta, Mingfei Yan, Richard Newcombe, Carl Ren, and Omkar M. Parkhi. Egoblur: Responsible innovation in aria, 2023. URL `https://arxiv.org/abs/2308.13093`.

RealEye. Realeye webcam eye-tracking for desktops, laptops and smartphones: Technology white paper. Technical report, RealEye, May 2025. White paper describing system specifications and validation of webcam-based eye tracking accuracy under varying conditions.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*, pp. 3982–3992, 2019.

Andreas Säuberli, Darja Jepifanova, Diego Frassinelli, and Barbara Plank. Controlling reading ease with gaze-guided text generation. *arXiv preprint arXiv:2601.17781*, 2026.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020.

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

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*, 2023.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Chi Hsuan Wu, Kumar Ashutosh, and Kristen Grauman. Skillsight: Efficient first-person skill assessment with gaze. *arXiv preprint arXiv:2511.19629*, 2025.

Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, and Gunhee Kim. Supervising neural attention models for video captioning by human gaze data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 490–498, 2017.

Appendix

A Human Study Instructions

All participants provide informed consent prior to participation and receive compensation following Prolific payment guidelines and institutional policies. The user interface is presented in

Instructions for `SlideVQA`

You will view a deck of 10 slides and summarize the content.
Before starting

- You will complete a brief gaze calibration to estimate where you are looking.
- No images or video of your face are stored—only gaze coordinates.
- Keep your head relatively still during calibration and the task.

If you are using an external monitor, open the study on the device with the camera.
During the study

- Read each slide at your normal pace.
- After each slide, write a short paragraph summarizing the main ideas and key takeaways.

Requirements

- Age 18+
- Fluent in English

Study Instruction: You will view a deck of 10 slides.

- Please read each slide carefully at your normal pace.
- After each slide, write a short paragraph summarizing the main ideas/takeaways of the slide.

Important

- Write meaningful summaries based on the slide content. Submissions may be rejected if responses are nonsensical or repetitive.

Instructions after each slide:
Write one clear, natural paragraph summarizing the main ideas/takeaways of the last slide.

Instructions for `Aria Everyday Activities`

You are invited to take part in a research study on first-person activity recognition. You will watch short videos recorded from smart glasses and describe the wearer's actions.
What You'll Do

- Watch 27-35 short video clips (∼5 minutes total)
- Write one sentence per clip describing the action

Requirements

- Age 18+
- Fluent in English

Purpose
You are invited to participate in a research study on first-person activity recognition. You will watch short video clips recorded from smart glasses and write captions describing the wearer's actions.
What You'll Do Watch 33 short video clips (about 4.7 minutes total), and write a caption for each clip. For each clip:

- A red dot indicates the wearer's gaze.
- Write one sentence describing what the wearer is doing (from their perspective), e.g., "I pick up my phone from the table...".
- Each caption must be at least 30 characters long.

Risks, Confidentiality, and Voluntary Participation

- No known risks are associated with this study.
- Your responses are confidential, not linked to your identity, and securely stored.
- Participation is voluntary. You may stop at any time by closing the survey.

Instructions after each video clip: Then, imagine you are the wearer (the red dot shows your gaze) and write one sentence describing what you are doing.

- At least 30 characters
- Use first-person perspective (e.g., "I open the fridge and take out a Coke.")
- Describe the main action and object clearly and specifically

B User Study Interfaces

We present the interfaces used in the user studies for annotating the Aria Everyday Activities (AEA) and SlideVQA datasets.

Aria Everyday Activities (AEA). The user study recruitment details page and the main annotation interface are shown in Figure 7 and Figure 8, respectively.

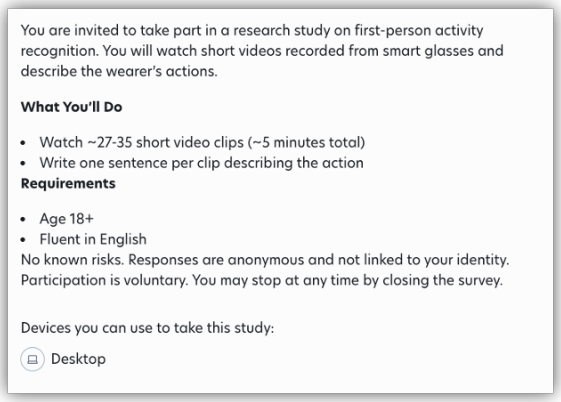


Figure 7: **Prolific landing page** for the Aria Everyday Activities (AEA) study.

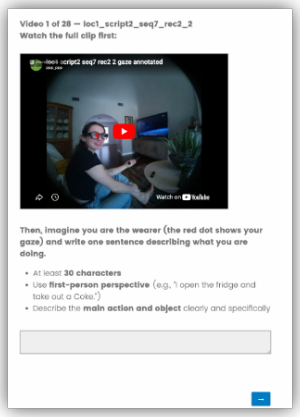


Figure 8: **Qualtrics annotation interface** for AEA, displaying a video clip with the wearer's gaze indicated by a red dot and a text box for first-person descriptions.

SlideVQA. The Prolific landing page and the RealEye eye-tracking and response collection interfaces for the SlideVQA study are shown in Figures 9 to 11.

C Human Annotator Demographics

Aria Everyday Activities. We recruited 53 participants via Prolific to provide activity captions for the video clips. All participants provided informed consent prior to participation. The cohort has a mean age of 44.3 years (SD = 13.1, median = 42). With respect to gender, 56.6% of participants identified as female, 37.7% as male, 3.8% as non-binary or third gender, and 1.9% preferred not to say.

SlideVQA. We recruited 150 participants via Prolific to serve as eye-tracking annotators. All participants viewed stimuli on desktop browsers using RealEye's webcam-based eye tracker. The mean eye-tracking quality grade was 4.88 out of 6, with 87.4% of sessions rated as high quality (grade ≥ 4). Participants and their demographic information (age, gender) were not collected as part of the anonymous Prolific study protocol.

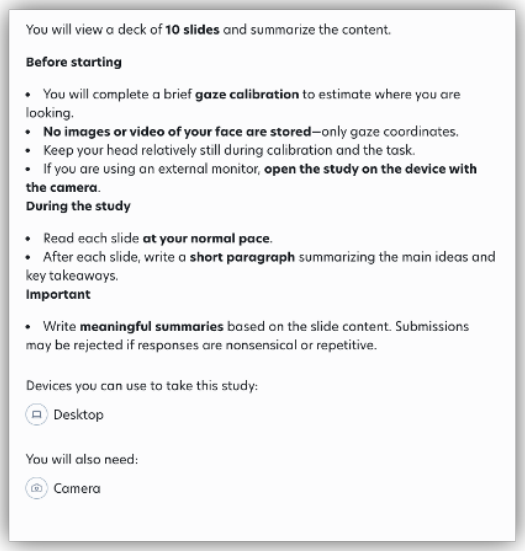


Figure 9: **Prolific study landing page** for the SlideVQA study, outlining eye-tracking calibration and response requirements.

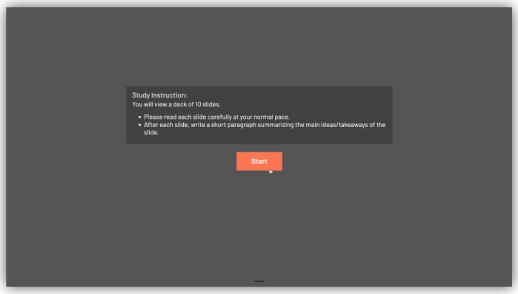


Figure 10: **RealEye study instruction screen** after eye-tracking calibration.

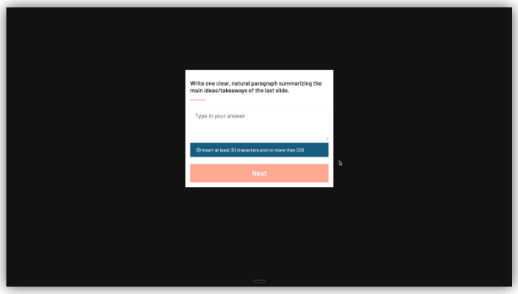


Figure 11: **RealEye annotation interface** after showing a slide to a participant.

D Prompts Used

We use the following prompts for VLM candidate caption generation and VEGAS scoring.

Candidate Caption Generation Prompt

```
You are a video content annotator.  Generate one concise, objective paragraph describing the provided
video in a natural, human-annotator style.

Guidelines:
    • Describe the setting, main subjects or objects, and key actions in chronological order.
    • Do NOT use meta phrases such as "the video begins with," "the clip shows," or "the camera
      shows."
    • Start sentences directly with the observable subject or action, e.g., "A woman in a blue dress
      walks..."
    • Include only essential visual details when relevant.
Constraints:
    • Output only the descriptive paragraph.
    • No interpretation, speculation, or non-observable information.
    • Maintain a neutral, factual tone.
```

VEGAS Scoring Prompt

```
Given the video frames and the masked caption of a video:  {masked_caption}.  Guess all [MASK] words
originally representing any words describing the video, e.g., first_guess second_guess.  Return only
the answers, without any explanation.  Do not use quotes or commas; separate tokens with a single
space.
```

E Caption Post-processing

Because the LLM gateway serving pipeline did not reliably strip Qwen's thinking-mode content from returned captions, we filtered Qwen outputs containing reasoning artifacts such as intermediate analysis, drafting language, or phrases like "the user wants," "identify," and "let's refine." This reduced the Qwen SlideVQA captions from 300 to 214 examples at temperature 0.0 and from 300 to 164 at temperature 1.0; the remaining captions were pooled with Gemini- and GPT-generated captions for downstream SBERT analyses.

F Compute Resources Used

Gemini and GPT captions were generated through their official provider APIs, while Qwen captions were generated using an on-premise deployment served through an LLM gateway. VEGAS scores were generated using a locally deployed Gemma 4 model on a Primus inference deployment with 4× AMD Instinct MI300X GPUs (192 GB HBM per GPU), 32 CPU cores (Intel Xeon Platinum 8570), and 512 GB RAM. Data processing and downstream analyses were performed on a separate workstation equipped with 4× NVIDIA RTX 6000 Ada GPUs (48 GB VRAM per GPU), an Intel Xeon Gold 6346 CPU, and 251 GB RAM.

G Qualitative Example

This section provides three additional qualitative examples. Figure 12 shows an AEA case where the gaze-conditioned frame supports a caption about the attended television screen. Figure 13 shows SlideVQA cases where VEGAS is less effective: the captions are reasonable, but receive high scores because they summarize global chart trends or causal relationships rather than only the locally attended regions.

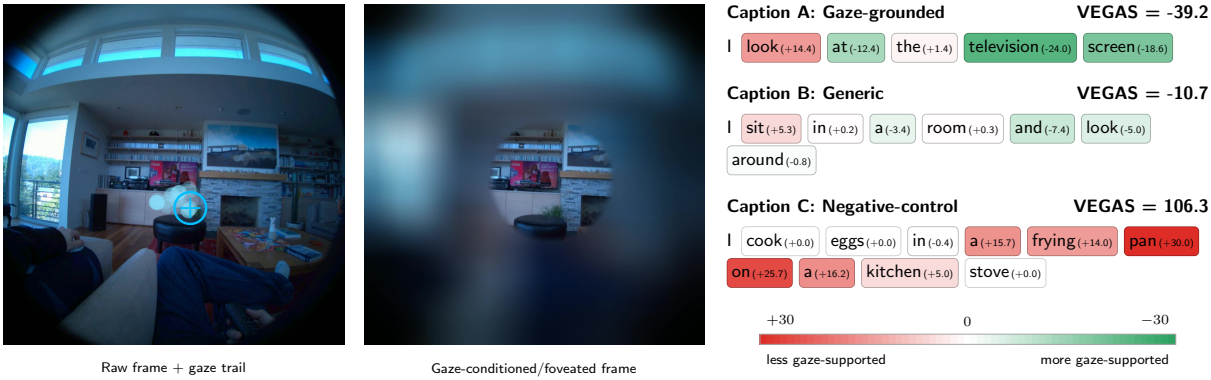
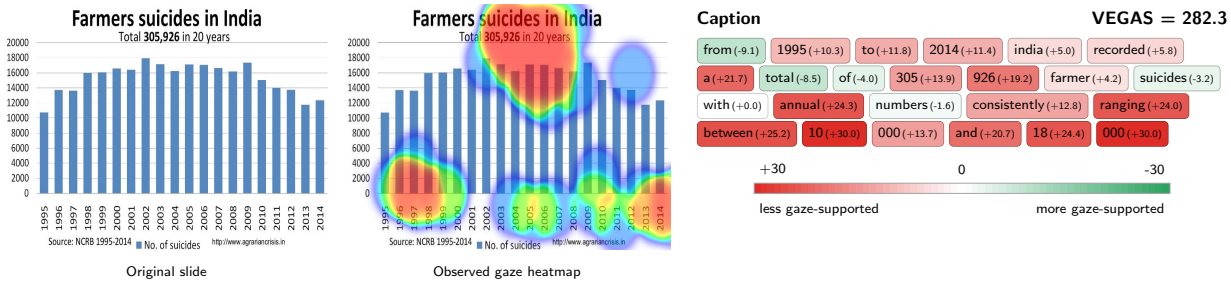
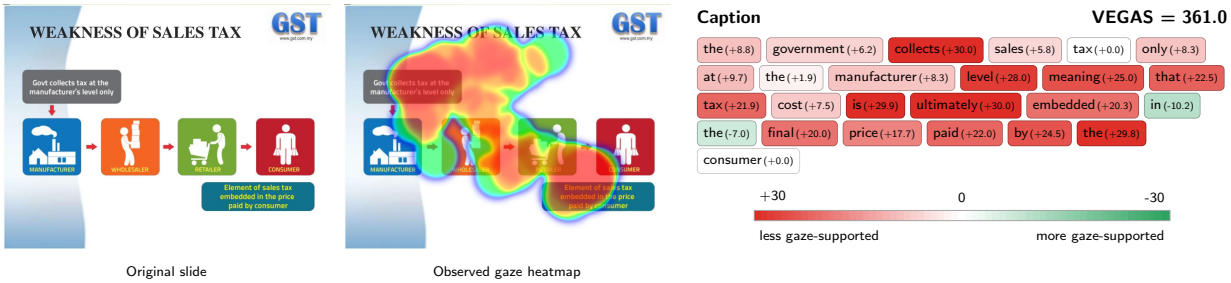


Figure 12: **Additional AEA qualitative example.** The gaze-conditioned evidence supports the caption describing looking at the television screen, while generic and negative-control captions are less gaze-supported.



(a) Chart aggregation. The caption compresses information from the full bar chart, including years, counts, and the total, while the gaze heatmap only partially covers those supporting regions.



(b) Causal interpretation. The caption describes the overall tax-flow mechanism, whose support is spread across multiple diagram nodes and arrows rather than a single attended region.

Figure 13: **Challenging cases in SlideVQA.** Both have reasonable captions with high VEGAS scores.

H Ablation over top-K and fallback log-probability λ

VEGAS uses the scoring VLM's returned top-K token log probabilities to approximate Equation (3). If a target token is not in the returned set, we assign it a fallback log-probability λ. We ablate both implementation choices by selecting one caption per AEA clip from the same fixed candidate pool and measuring the selected captions with median SBERT similarity to the human-caption centroid and human-query Recall@5. We also report a random-candidate baseline that samples one caption uniformly from the same candidate pool, averaged over 1000 seeds. We use `Qwen2.5-VL-32B-Instruct` as an additional scorer because it is a strong open-weight VLM of comparable scale that can be deployed locally.

Table 6: **Ablation over top-K and fallback log-probability λ in AEA.** Each cell reports median SBERT / Recall@5 for VEGAS-selected captions under a top-K and fallback log-probability setting. The random-candidate baseline samples one caption uniformly from the same candidate pool and is independent of scorer, top-K, and λ.

		Fallback log-probability λ		
Scorer	Top-K	-10	-20	-30
`Gemma-4-31B-IT`	5	0.736 / 0.509	0.735 / 0.511	0.737 / 0.517
	10	0.736 / 0.514	0.739 / 0.522	0.738 / 0.523
	20	0.739 / 0.515	0.737 / 0.508	0.741 / 0.523
`Qwen2.5-VL-32B-Instruct`	5	0.736 / 0.524	0.736 / 0.519	0.736 / 0.521
	10	0.736 / 0.505	0.736 / 0.508	0.736 / 0.506
	20	0.740 / 0.513	0.725 / 0.511	0.725 / 0.509
Random candidate caption, 1000 seeds			0.727 / 0.484	

Table 6 shows that downstream performance changes little across this grid. All settings are above the random-candidate baseline (0.727 median SBERT and 0.484 Recall@5), indicating that VEGAS improves over chance selection from the candidate pool. Thus, while K and λ change the absolute VEGAS scale, the semantic and retrieval quality of the selected captions is not sensitive to these implementation details.