# WHEN TO TRUST THE EYES: LEARNING ADAPTIVE GAZE RELIANCE IN EGOCENTRIC VIDEO

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Gaze recorded by wearable devices reveals where the wearer attended, and prior work exploits it for egocentric video understanding under the implicit assumption that it is useful regardless of the task. However, we show that not all gaze is useful for question answering, the core function of a wearable assistant. Rather, its usefulness depends on the question and the scene context, so the same fixation can be useful for one question yet noise for another. Accordingly, how much to rely on gaze should be decided for each question and frame, trusting gaze where it is useful and suppressing it where it acts as noise. To this end, we propose **Re**liance on **Gaze** (ReGaze), a method that learns this reliance. Specifically, we design the Adaptive Gaze Reliance Estimation (AGRE) module, a lightweight network trained without any label of gaze usefulness, using a ranking loss that favors reliance on gaze where it helps localize question-relevant frames and a noise suppression loss that penalizes it where it hurts. To train the AGRE module, we further build a Gaze-aware QA Curation (GQC) pipeline that generates questions paired with their relevant intervals from gaze-recorded videos. On three egocentric QA benchmarks, ReGaze outperforms state-of-the-art frame samplers and naive uses of gaze. Beyond performance, ReGaze captures the underlying behavior of gaze without being trained on it: its reliance decreases as the gaze shifts rapidly or lies in the periphery. It also excels on WearerQA, our evaluation set of questions written by the wearers themselves. The code and dataset will be released upon acceptance.

## 1 INTRODUCTION

Egocentric video understanding aims to understand a person's activities and interactions with surrounding environments from a first-person perspective. With the advent of wearable devices such as Augmented Reality (AR) glasses and Virtual Reality (VR) headsets (Engel et al., 2023; Varjo, 2025), this field has become increasingly vital for developing wearable assistants that continuously perceive the wearer's experiences and answer questions about the events. Along with these advances, many of these wearables have integrated eye trackers (Apple Inc., 2024; Pico Interactive, 2024), enabling seamless capture of the wearer's *gaze* as a direct visual cue for human attention (Li et al., 2021).

In this regard, many studies have explored gaze across various aspects of egocentric video understanding. Specifically, gaze has been used to assess a wearer's skill (Wu et al., 2026), to improve inference efficiency by allocating computation to attended visual regions (Rekimoto, 2025; Shi et al., 2026; Chen & Qi, 2025), and to enhance model effectiveness by regularizing model attention toward human gaze (Pani & Yang, 2026a; Yan et al., 2024). Furthermore, recent benchmarks incorporate gaze to better reflect wearable-assistant scenarios, in which wearers typically ask about what they have previously looked at (Tang et al., 2026). Overall, these studies leverage the raw gaze signal as a proxy for human attention, implicitly assuming that it is useful regardless of the task at hand.

However, *this assumption does not necessarily hold in question answering, the core function of a wearable assistant*. In particular, we argue that not all gaze is useful for answering the wearer's questions; rather, its usefulness depends on the given question. This is because gaze is driven by the wearer's ongoing action, not by the question. Hence, the information a question requires is not necessarily tied to where the wearer attended, so even a fixation on the correct object can



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.




Figure 1: (a) The same gaze fixation near the keys is useful for Q1 but noisy for Q2, whose evidence lies outside the gazed region, while the fixation in the last frame is only partially useful for Q3. (b) Accuracy on EgoExoLearn (Huang et al., 2024) as the reliance on gaze $\alpha$ increases.

provide evidence for one question while acting as noise for another[1]. For example, as shown in Fig. 1(a), the same gaze fixation near the keys is useful for answering `Where did I leave my keys?`, but provides no information for `Do you remember the location of my white letter?` (noisy), since the answer evidence lies outside the gazed region. Moreover, upon closer inspection, the fixation in the last frame is only partially useful for `Did I leave the kettle on the stove?`: it helps locate the region the wearer is engaged with, but deriving the answer requires additional evidence about the surrounding scene. In short, the usefulness of gaze is determined not only by the gaze signal itself but also by its relation to the given question and scene context of each frame, and gaze acts as noise when this relationship does not hold.

To investigate how this usefulness affects question answering, we need a setting in which the contribution of gaze can be isolated. We therefore focus on *frame sampling*, i.e., identifying the moments that contain question-relevant evidence in a long video, since it determines which visual evidence a Video Large Language Model (Video LLM) sees before answering. Within this process, we take a closer look at how the reliance on gaze affects the resulting accuracy. Specifically, we extract from each frame a *gaze feature* by Gaussian pooling around the gaze coordinate (Pani & Yang, 2026a; Yan et al., 2024) (see Sec. 4.1) and a *frame feature* that summarizes the entire frame (e.g., the CLS token), and take their convex combination with weight $\alpha \in [0, 1]$ on the gaze feature (see Eq. 2). We then select the 32 frames with the highest similarity to the question and feed them into the Video LLM (Wang et al., 2025a) to generate the answer. As shown in Fig. 1(b), a moderate reliance on gaze improves accuracy by locally capturing the wearer's attention, whereas relying solely on gaze ($\alpha$=1) leads to a sharp drop, as cases such as Q2 in Fig. 1(a) are missed. Taken together with Fig. 1(a), this reveals a key challenge: the appropriate degree of reliance on gaze varies not only with the given question but also from frame to frame with the scene context, which requires adaptively relying on gaze while suppressing its noise.

To address this challenge, we propose ReGaze, which learns the reliance on gaze for each frame given the question, in place of a fixed $\alpha$. Specifically, we first design an Adaptive Gaze Reliance Estimation (AGRE) module, a lightweight network that estimates, for a given question, how much to rely on gaze when selecting each frame, while keeping the frame sampler frozen. The AGRE module is trained with a ranking loss (Bradley & Terry, 1952), which implicitly encourages reliance on gaze wherever it helps localize question-relevant frames. To complement this, we introduce a noise suppression loss that penalizes reliance on gaze wherever it hurts, so that the AGRE module learns not only when to rely on gaze but also when to suppress it. Both losses require only the interval in which each question is answerable, so the AGRE module learns when to trust gaze without any label of gaze usefulness. These intervals, however, must come with questions whose relation to gaze varies, from what the wearer looked at to what merely happened around them, and such data is not readily available. We therefore build a Gaze-aware QA Curation (GQC) pipeline that generates such questions, paired with their relevant intervals, from gaze-recorded videos.

To validate the effectiveness of ReGaze, we evaluate it on three egocentric QA benchmarks, EgoExoLearn (Huang et al., 2024), EgoEverything (Tang et al., 2026), and EgoGazeVQA (Peng et al., 2026), comparing it with state-of-the-art frame samplers and naive uses of gaze, and ReGaze outperforms them. Beyond accuracy, although ReGaze is not trained on gaze speed or position, it

[1]In this paper, we define gaze as *noisy* when it provides little or no evidence for answering the given question, even if the gaze itself is correctly observed.

captures the underlying behavior of gaze, lowering its reliance on gaze that shifts rapidly or lies in the periphery. Finally, we evaluate ReGaze in the real-world scenario in which the wearer asks about their own experience. To this end, we collect WearerQA, in which participants write questions from memory about their own recordings, and ReGaze excels on it as well.

We summarize our main contributions as follows: 1) We identify that the usefulness of gaze for question answering depends on the question and scene context of each frame, so that neither relying on it fully nor any fixed reliance suffices. 2) We propose ReGaze, which learns the reliance on gaze through the AGRE module, trained with a ranking loss and a noise suppression loss on data curated by the GQC pipeline. 3) On three egocentric QA benchmarks and the wearer-authored WearerQA, ReGaze outperforms state-of-the-art frame samplers and naive uses of gaze.

## 2 Related Works

**Gaze as Human Attention.** Gaze is a direct visual cue of human attention, and eye trackers on wearable devices capture it seamlessly alongside egocentric video. Prior work has exploited this signal in several ways: 1) Gaze reveals the wearer's state, as fixation patterns distinguish experts from novices (Wu et al., 2026) and expose mistakes in skilled activities without supervision (Mazzamuto et al., 2025). 2) Gaze improves efficiency by telling the model where to compute, e.g., allocating visual tokens only to attended regions (Shi et al., 2026; Chen & Qi, 2025; Rekimoto, 2025) or retaining only fixation keyframes from a video (Mardanbegi et al., 2025). 3) Gaze enhances effectiveness by regularizing model attention toward human gaze, in Video LLM (Pani & Yang, 2026a; Yan et al., 2024), streaming video understanding (Pham et al., 2026), robotic manipulation (Pani & Yang, 2026b), and medical imaging (Li et al., 2026b). 4) Gaze has been incorporated into benchmarks for egocentric QA (Peng et al., 2026; Tang et al., 2026) and streaming settings (Lee et al., 2025). In all of these, the gaze signal is used as recorded, and its informativeness is taken for granted. By contrast, a few works do address the inherent noise of the gaze signal (e.g., saccades, jitter, or ambiguity) by filtering unstable gaze with stability and pupil cues (Subramanian et al., 2026), resolving ambiguous referents in gaze-facilitated interaction (Wang et al., 2025b), or keeping only fixation frames (Mardanbegi et al., 2025). While these works treat noise as a property of the gaze signal, we address noise that depends on the question and the scene context, as illustrated in Fig. 1(a): even a fixation on the right object can be noise for the question at hand.

For additional related work on egocentric video understanding, please refer to Appendix A.

## 3 Problem Formulation

Let $V = \{(f_t, g_t)\}_{t=1}^{T}$ be an egocentric video of $T$ frames captured by the wearer, where $g_t$ is the 2D gaze coordinate recorded for frame $f_t$. Given a question $q$ about the video, the task is to answer it from $V$ (Grauman et al., 2022), and since the Video LLM that generates the answer can consume only $K$ frames out of $T$ ($K \ll T$), it is crucial to identify the moments relevant to $q$. To this end, prior work (Tang et al., 2025) identifies question-relevant moments with a vision-language encoder (Sun et al., 2023), where a visual encoder extracts a frame feature $\mathbf{v}_t$ of frame $f_t$ (e.g., CLS token), a text encoder extracts a question feature $\mathbf{q}$, and the relevance of each frame to the question is measured by their cosine similarity, from which the top-$K$ frames are selected:

$$\mathcal{S} = \underset{t\in\{1,\dots,T\}}{\text{Top-}K}\ (s_t), \qquad s_t = \cos(\mathbf{v}_t, \mathbf{q}), \tag{1}$$

where $\mathcal{S}$ is the set of frames fed to the Video LLM. Beyond $\mathbf{v}_t$, which summarizes the scene context of frame $f_t$ as a whole, gaze offers a complementary view: a gaze feature $\mathbf{z}_t$ that captures what the wearer was attending to in the frame (Sec. 4.1 details how it is obtained). We incorporate this view by balancing the two features,

$$s_t = \cos\big((1-\alpha)\,\mathbf{v}_t + \alpha\,\mathbf{z}_t,\ \mathbf{q}\big), \tag{2}$$

where $\alpha \in [0, 1]$ controls the reliance on gaze: $\alpha$=0 recovers Eq. (1), which we refer to as the *frame-only sampler*, and $\alpha$=1 judges the frame by its gazed region alone, which we refer to as the *gaze-only sampler*. However, as discussed in Sec. 1, neither extreme nor any fixed value in between suffices, since the usefulness of gaze depends on the question and varies from frame to frame. Therefore, our goal is to adaptively adjust the reliance on gaze for each question and frame, i.e., to learn a reliance $\alpha_t = f_\theta(\mathbf{q}, \mathbf{v}_t, \mathbf{z}_t)$ with a learnable function $f_\theta$ in place of a fixed $\alpha$, so that the selected frames deliver more evidence to the Video LLM.

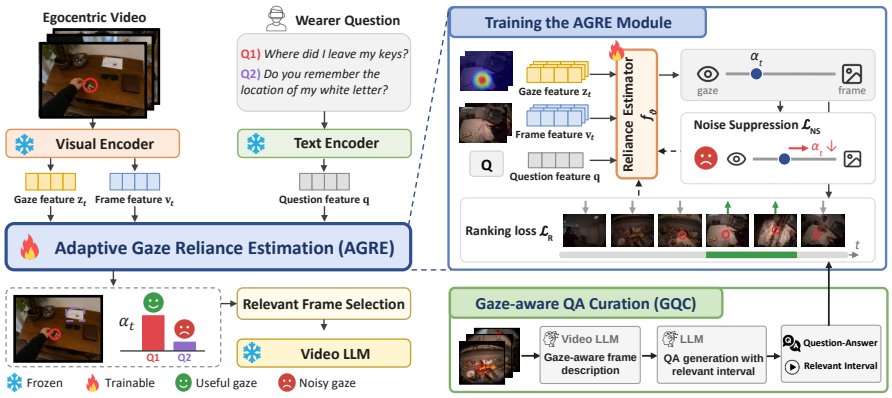


Figure 2: Overall framework of ReGaze. Given a question, the AGRE module estimates how much to rely on the wearer's gaze for each frame, so that the same gaze is trusted for one question (Q1) and suppressed for another (Q2) when selecting the frames for the Video LLM. The AGRE module is trained with a ranking loss and a noise suppression loss on the questions and relevant intervals curated by the GQC pipeline.

# 4 METHOD

Building on Sec. 3, we propose ReGaze, which learns the reliance on gaze $\alpha_t$ for each question and frame in place of a fixed $\alpha$. Specifically, we first design an Adaptive Gaze Reliance Estimation (AGRE) module that estimates $\alpha_t$ from the question and the frame (Sec. 4.1). Since the usefulness of gaze is impractical to annotate, the AGRE module is trained with a ranking loss and a noise suppression loss that require only the interval in which each question is answerable, from which it learns to rely on gaze where doing so helps. As such data is not readily available, we build the Gaze-aware QA Curation (GQC) pipeline that generates questions paired with such intervals (Sec. 4.2). Finally, we describe a wearer-authored evaluation set that reflects how a wearer actually asks about their own experience (Sec. 4.3). The overall framework is shown in Fig. 2.

## 4.1 ADAPTIVE GAZE RELIANCE ESTIMATION

Recall that, as discussed in Sec. 1, the usefulness of gaze varies with the question and with the scene context of each frame, so the AGRE module estimates the reliance $\alpha_t$ for each question and frame rather than sharing a single value across them. To this end, the AGRE module consists of three components: a gaze feature $\mathbf{z}_t$ pooled from the visual encoder around the gaze coordinate, a lightweight network $f_\theta$ that estimates $\alpha_t$ from $\mathbf{q}$, $\mathbf{v}_t$, and $\mathbf{z}_t$, and two losses that train $f_\theta$ from the relevant interval provided with each training question (see Sec. 4.2), where a ranking loss raises reliance on gaze wherever it helps localize the interval and a noise suppression loss lowers it wherever it hurts. In what follows, we describe each component in turn, as shown in Fig. 2 (top right). Only $f_\theta$ is trained; the visual and text encoders (EVA-CLIP (Sun et al., 2023)) are kept frozen.

**Gaze feature.** We begin with the gaze feature $\mathbf{z}_t$ in Eq. (2), which should describe what the wearer attended to rather than the frame as a whole. We obtain it from the patch features of the visual encoder by Gaussian-weighted pooling around the gaze coordinate (Yan et al., 2024):

$$\mathbf{z}_t = \frac{\sum_p w_{t,p}\,\mathbf{e}_{t,p}}{\sum_p w_{t,p}}, \qquad w_{t,p} = \exp\Big(-\frac{\|c_p - g_t\|^2}{2\sigma^2}\Big), \tag{3}$$

where $\mathbf{e}_{t,p}$ is the feature of patch $p$ of frame $f_t$, $c_p$ is the center of that patch, and $\sigma$ is the Gaussian width that controls the extent of the attended region. The pooled feature then passes through the same projection and normalization as the CLS token that yields $\mathbf{v}_t$, so that $\mathbf{z}_t$, $\mathbf{v}_t$, and $\mathbf{q}$ lie in the same embedding space and Eq. (2) is well defined. When gaze is not recorded for a frame, e.g., during blinks or tracking loss, we fix $\alpha_t = 0$ so that the frame is scored by $\mathbf{v}_t$ alone.

**Reliance estimator.** Given $\mathbf{q}$, $\mathbf{v}_t$, and $\mathbf{z}_t$, the estimator $f_\theta$ outputs $\alpha_t$. Because the usefulness of gaze at a frame can also depend on what happened just before it, e.g., whether the wearer keeps attending to the same region or has just shifted to a new one, $f_\theta$ looks at the frame together with its

previous context. Although all preceding frames could be used, we adopt the local-context window strategy (Di et al., 2025), which restricts the context to the most recent frames and thus keeps the cost per frame constant while preserving parallel computation through an attention mask[2]. Specifically, a shared MLP maps the concatenation of $\mathbf{q}$, $\mathbf{v}_t$, and $\mathbf{z}_t$ of each frame in the window to a hidden feature, a lightweight transformer encoder with learnable positional embeddings attends over these hidden features, and a second MLP maps the output at the current frame to a scalar, which passes through $\tanh$ to give $\alpha_t$ (see Appendix B.7 for the effect of the local window).

Herein, two design choices are essential. First, the output layer of the second MLP is initialized to zero, so that $\alpha_t \equiv 0$ at the start of training: the AGRE module starts from the *frame-only* sampler of Eq. (1), the safer of the two extremes in Fig. 1(b), and raises reliance on gaze where training finds it useful. Without this initialization, as shown in Fig. 3[3], the losses introduced below drive $\alpha_t$ toward 1 regardless of the question and frame, and the AGRE module degenerates into the *gaze-only sampler*, which is undesirable. Second, we use $\tanh$ rather than a sigmoid because $\tanh$ is exactly zero at a zero-initialized output, which is what makes the *frame-only* start possible; its negative range, which would move the mixed feature away from the gaze feature, is not needed in practice, and the learned $\alpha_t$ eventually lies in $[0, 1]$, so Eq. (2) remains a convex combination of the two features. For brevity, we write $s_t(\alpha)$ for the relevance of Eq. (2) evaluated at a given reliance, so that $s_t(0)$ is the frame-only relevance and $s_t(\alpha_t)$ the one AGRE produces.

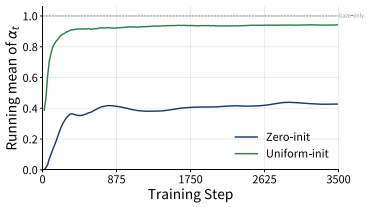

Figure 3: $\alpha$ over training steps.

**Ranking loss.** To let $f_\theta$ learn the reliance on gaze by itself, we supervise it implicitly, through the relevant interval $\mathcal{R} \subset \{1, \ldots, T\}$ of each training question, i.e., the set of frames during which the question is answerable (Sec. 4.2): frames inside $\mathcal{R}$ are positives and should be scored above frames outside it, and how much to rely on gaze in doing so is left to $f_\theta$. To this end, we adopt the Bradley–Terry pairwise ranking loss (Bradley & Terry, 1952) between a positive frame $i \in \mathcal{R}$ and a negative frame $j \notin \mathcal{R}$,

$$\mathcal{L}_{\mathrm{R}} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \log\Big(1 + \exp\big(-(s_i(\alpha_i) - s_j(\alpha_j))\big)\Big), \tag{4}$$

where $j$ is sampled for each positive frame from the frames outside $\mathcal{R}$. Since $s_i$ and $s_j$ depend on $\alpha_i$ and $\alpha_j$ through Eq. (2), the gradient of this loss raises reliance on gaze wherever gaze moves a positive frame up or a negative frame down, i.e., wherever gaze helps localize $\mathcal{R}$.

**Noise Suppression loss.** The ranking loss alone, however, does not teach $f_\theta$ when gaze is noisy. It raises $\alpha_t$ wherever gaze happens to correlate with the question, without asking whether gaze helps beyond the frame feature, which is precisely how Q2 in Fig. 1(a) fails. To address this, we introduce the noise suppression loss $\mathcal{L}_{\mathrm{NS}}$, which penalizes reliance on gaze in proportion to the harm it causes on positive frames,

$$\mathcal{L}_{\mathrm{NS}} = \frac{1}{|\mathcal{R}|} \sum_{t \in \mathcal{R}} \max\big(0,\ s_t(0) - s_t(\alpha_t)\big) \cdot |\alpha_t|, \tag{5}$$

where $s_t(0)$ is treated as a constant. Intuitively, the frame-only relevance $s_t(0)$ serves as a reference for what a positive frame earns without gaze: if mixing in the gaze feature with $\alpha_t$ lowers the relevance below this reference, the gaze is regarded as noise, and $\alpha_t$ is pushed down in proportion to the drop; otherwise, the loss is zero and the ranking loss is left free to raise $\alpha_t$. Note that this reference costs nothing, since the encoder is frozen and $s_t(0)$ is simply Eq. (2) at $\alpha_t{=}0$. Without $\mathcal{L}_{\mathrm{NS}}$, the learned reliance $\alpha_t$ tends to grow indiscriminately toward the gaze-only regime, and accuracy drops accordingly, which is described in Sec. 5.3.

**Training objective and inference.** The AGRE module is trained with $\mathcal{L} = \mathcal{L}_{\mathrm{R}} + \lambda\, \mathcal{L}_{\mathrm{NS}}$, where $\lambda$ balances the two terms. Note that both losses depend only on $\mathcal{R}$ and the frozen features: neither a label of gaze usefulness nor any measurement of gaze quality, such as gaze speed or pupil size, is used, so whatever the AGRE module learns about noisy gaze is learned from $\mathcal{R}$. At inference, the

[2]In this paper, the window consists of the current frame and its four preceding frames.

[3]We plot the running mean of $\alpha_t$ over the training to show its overall trajectory. Gaussian initialization (Glorot & Bengio, 2010) follows the similar trajectory as uniform initialization and is therefore omitted.

AGRE module computes $\alpha_t$ for all $T$ frames in a single pass, scores them by Eq. (2), and passes the top-$K$ frames to the Video LLM, so that the only overhead over frame-only sampling is one forward pass of $f_\theta$, whose cost per frame stays constant thanks to the local-context window.

### 4.2 GAZE-AWARE QA CURATION

As discussed in Sec. 4.1, the AGRE module requires, for each training question, the relevant interval during which it is answerable, on videos that record the wearer's gaze. A common way to obtain such data is to describe the video in text and let an LLM generate questions together with the moments on which their answers rest (Yao et al., 2025; Hu et al., 2025). In our setting, however, the questions should reflect what a wearer asks about their own experience, ranging from what they looked at to what merely happened around them, so that gaze is useful for some questions and incidental for others. Since no such data exists, we build the Gaze-aware QA Curation (GQC) pipeline on the gaze-recorded videos, which proceeds in three steps (see Fig. 2, bottom right): frames are described with the gaze taken into account, questions are generated from the descriptions together with their relevant intervals, and the results are filtered.

**Gaze-aware frame description.** We first extract frames at 1 fps and describe each frame in text, inspired by Yao et al. (2025), so that the description records not only what is in the scene but also what the wearer is looking at. To make each description gaze-aware, we overlay the gaze coordinate $g_t$ on frame $f_t$ as a marker, following the visual prompting of Tang et al. (2026), and instruct the Video LLM (Qwen Team, 2026b) to describe the frame with the marker as the wearer's point of attention. Furthermore, to keep the descriptions compact and temporally informative, each frame is described *differentially* (Chen et al., 2024), i.e., conditioned on the previous description, so that it records what changed rather than restating the scene, and consecutive descriptions with no new change are merged into a single *event* with its time span.

**Question generation with relevant intervals.** Given these time-stamped descriptions, an LLM (OpenAI, 2026) generates question–answer pairs and, for each question, reports the events on which the answer rests; snapped to event boundaries, this span becomes the relevant interval, along with a few hard-negative spans that resemble it but do not answer the question[4]. To have the AGRE module generalize across the diverse questions a wearer may ask, the generated questions span what the wearer looked at, what they did, and how events unfolded. Importantly, questions never mention timestamps, so that the relevant interval has to be found from the frames themselves. While the answers are not used to train the AGRE module, they serve to examine how gaze affects the Video LLM itself (see Sec. 5.3).

**Quality control.** Finally, the generated samples are filtered so that the supervision is reliable: we remove samples whose span cannot be snapped, whose question leaks the answer or over-specifies the moment, or whose answer is not reproduced upon regeneration. About 40% of the candidates are removed, leaving 40K questions over 230 sessions in which every question is answerable within its relevant interval, which is what the supervision in Sec. 4.1 assumes.

### 4.3 TOWARD REAL-WORLD EVALUATION

While the GQC pipeline supplies the training data, we further address evaluation in the real-world wearable-assistant scenario, in which a wearer asks about their own experience. Existing benchmarks generate their questions through third-party annotators or agents (Tang et al., 2026; Ragusa et al., 2026), so the questions reflect what is answerable from a video rather than what a wearer would ask or remember, leaving a gap between these benchmarks and the real-world scenario. To close this gap, we collect WearerQA, a wearer-authored evaluation set. Specifically, five participants recorded their daily activities (e.g., tidying up and eating) with gaze-tracking glasses (Baumann & Dierkes, 2023), and then, without rewatching the recordings, wrote questions about their own actions and surroundings that they would genuinely ask an assistant. Because recall is shaped by what one attended to, the resulting questions are naturally tied to the wearer's gaze, as in the real-world scenario. WearerQA contains 498 questions over 30 recordings, and we detail its collection protocol and statistics in Appendices C.1 and C.2, respectively.

[4]These hard-negative spans serve as the negatives in Eq. (4).

# 5 EXPERIMENTS

## 5.1 EXPERIMENTAL SETTING

**Datasets.** We evaluate on three egocentric QA benchmarks that record the wearer's gaze, EgoExoLearn (Huang et al., 2024), EgoEverything (Tang et al., 2026), and EgoGazeVQA (Peng et al., 2026), and additionally on our wearer-authored evaluation set, WearerQA (Sec. 4.3); we report accuracy throughout. To avoid data leakage, the AGRE module is trained only on the different source videos of HD-EPIC (Perrett et al., 2025) and SuperMemory-VQA (Alam et al., 2026), curated in Sec. 4.2, which share no video with the evaluation benchmarks. Regarding the additional details of datasets, please refer to Appendix B.1.

**Baselines.** We compare ReGaze with three groups of samplers that all select the same number of frames for the same Video LLM, so that the differences arise only from which frames are selected. First, we include the two single-feature samplers of Sec. 3, i.e., the *frame-only sampler* ($\alpha$=0) and the *gaze-only sampler* ($\alpha$=1). Second, we include two naive ways of combining the two features. The fixed-$\alpha$ sampler (*Fixed* $\alpha$) applies a single reliance to all questions and frames in Eq. (2), using the $\alpha$ that performs best on each benchmark. The max-similarity sampler (*Maxsim*) instead takes, for each frame, the higher of the gaze and frame similarities to the question as the relevance, thereby falling back to the frame feature whenever gaze matches the question poorly. The former represents the best that a fixed reliance can achieve, and the latter the naive rule one might use to decide when to rely on gaze. Third, we compare with state-of-the-art frame samplers that use the frame feature: FOCUS (Zhu et al., 2026), AKS (Tang et al., 2025), and Nar-KFC (Fang et al., 2026). For additional details of baselines, please refer to Appendix B.2.

**Implementation details.** We use EVA-CLIP (Sun et al., 2023) as the vision-language encoder of the frame sampler, InternVL3.5-8B (Wang et al., 2025a) and Qwen3.5-9B (Qwen Team, 2026a) as the Video LLM backbones, and $K$=32 frames throughout. For all datasets, the Gaussian width $\sigma$ of the gaze feature, shown in Eq. 3, is set to 1.5 and $\lambda$ to 1.0, and frames are processed at 1 fps. In the GQC pipeline, the Video LLM for gaze-aware frame description is Qwen3.6-27B (Qwen Team, 2026b) and the LLM for question generation is GPT-5.4 (OpenAI, 2026). WearerQA is recorded with Pupil Labs Neon glasses (Baumann & Dierkes, 2023) at a resolution of $1600 \times 1200$. Further details are provided in Appendix B.3.

## 5.2 QUANTITATIVE RESULTS

In Tab. 1, we compare ReGaze with the baselines on three benchmarks, where the rows with the Video LLM alone use uniform sampling. The subscript indicates the number of uniformly sampled frames. We have the following observations. **1) The usefulness of gaze differs across benchmarks.** The gaze-only sampler slightly exceeds the frame-only sampler on EgoEverything, whose questions are generated around the wearer's gaze, but falls below it on the other two benchmarks, whose questions are not. This indicates that the appropriate reliance on gaze differs even across benchmarks. **2) Gaze helps only when used selectively.** The gaze-only sampler underperforms the frame-only sampler on average, whereas both Fixed $\alpha$ and Maxsim outperform it as well as the state-of-the-art frame samplers. This indicates that gaze provides information that the frame feature lacks, but this information comes with noise, so that relying on gaze indiscriminately hurts. Therefore, the key is to decide how much to rely on gaze, rather than whether to use it at all. **3) ReGaze performs best.** ReGaze achieves the highest aver-

Table 1: Accuracy on three egocentric QA benchmarks. For each backbone, the best and second-best results are in **bold** and underlined, respectively.

| Method | EgoExoLearn | EgoEverything | EgoGazeVQA | Avg. |
|---|---|---|---|---|
| EgoGPT-7B$_{frame:64}$ | 32.54 | 29.15 | 54.34 | 38.68 |
| Qwen3.5-9B$_{frame:64}$ | 43.18 | 43.37 | 56.86 | 47.80 |
| MiniCPM-V 2.6$_{frame:64}$ | 34.72 | 40.48 | 45.66 | 40.29 |
| LongVILA-7B$_{frame:128}$ | 31.11 | 40.68 | 44.06 | 38.62 |
| VideoLLaMA3-7B$_{frame:128}$ | 36.24 | 40.33 | 51.71 | 42.76 |
| InternVL3.5-8B | 37.08 | 40.41 | 53.26 | 43.58 |
| + Frame-only sampler | 38.05 | 41.23 | 53.94 | 44.41 |
| + Gaze-only sampler | 35.51 | 41.28 | 53.66 | 43.48 |
| + Fixed $\alpha$ | 38.54 | 41.37 | 54.17 | 44.69 |
| + Maxsim | 38.42 | 41.38 | 53.89 | 44.56 |
| + FOCUS | 38.12 | 41.58 | 53.49 | 44.40 |
| + AKS | 38.23 | 41.18 | 53.43 | 44.28 |
| + Nar-KFC | 36.66 | 40.78 | 53.71 | 43.72 |
| + ReGaze | **38.91** | **41.82** | **54.97** | **45.23** |
| Qwen3.5-9B | 40.58 | 43.64 | 55.49 | 46.57 |
| + Frame-only sampler | 42.36 | 43.46 | 57.09 | 47.64 |
| + Gaze-only sampler | 40.51 | 43.77 | 56.63 | 46.97 |
| + Fixed $\alpha$ | 43.23 | 43.77 | 57.26 | 48.09 |
| + Maxsim | 43.31 | **44.07** | **57.54** | 48.31 |
| + FOCUS | 42.97 | 43.71 | 56.80 | 47.83 |
| + AKS | 43.76 | 43.49 | 56.23 | 47.83 |
| + Nar-KFC | 42.19 | 42.97 | 56.80 | 47.32 |
| + ReGaze | **44.28** | 43.79 | **57.54** | **48.54** |

age accuracy on both backbones. In particular, it outperforms Fixed $\alpha$, which applies the best single reliance of each benchmark to all questions and frames, confirming that no fixed reliance suffices. It also outperforms Maxsim, which decides between the two features for each frame by a naive rule, showing that learning the reliance for each question and frame captures when to rely on gaze and when to suppress it beyond such a rule. Lastly, the state-of-the-art frame samplers bring only marginal gains over the frame-only sampler, which we discuss in Appendix B.2.

### 5.3 In-depth Analysis of Gaze Reliance

**Effect of $\sigma$ and $\lambda$.** We study two hyperparameters of ReGaze: the Gaussian width $\sigma$ of the gaze feature in Eq. (3) and the weight $\lambda$ of the noise suppression loss. As shown in Fig. 4, for $\sigma$, accuracy peaks at $\sigma$=1.5 and drops for smaller and larger values. Behind this observation, a small $\sigma$ heavily confines the gaze feature to a narrow region around the gaze, whereas a large $\sigma$ spreads it over the frame context. This indicates that the extent of the attended region needs to be neither too local nor too broad. For $\lambda$, removing the noise suppression loss ($\lambda$=0) gives the lowest accuracy, confirming that the ranking loss alone leaves noisy gaze unsuppressed, while a larger $\lambda$ beyond 1.0 also degrades accuracy, since it suppresses reliance on gaze even where gaze helps. A moderate $\lambda$ therefore balances the two losses, and we use $\sigma$=1.5 and $\lambda$=1.0 throughout (see Appendix B.4 for how the loss reshapes the distribution of $\alpha_t$).

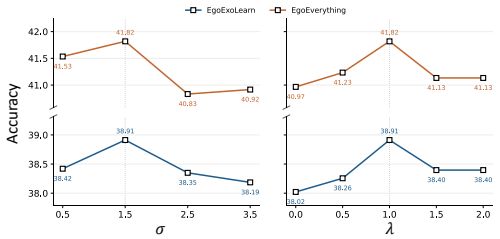


Figure 4: Accuracy of ReGaze over varying $\sigma$ and $\lambda$.

**Learned Reliance under Typical Gaze Behaviors.** Although the AGRE module is trained without any label of gaze usefulness and takes neither gaze position nor gaze speed as input, we explore the gaze behavior known to be noisy (Mardanbegi et al., 2025; Tatler, 2007), such as rapid shifts and peripheral positions, using the learned reliance $\alpha_t$. To this end, we bin the frames of EgoExoLearn by the distance of the gaze coordinate from the frame center and by gaze speed[5], and compute the mean $\alpha_t$ in each bin. As shown in Fig. 5, in both cases, $\alpha_t$ decreases monotonically: the farther the gaze lies from the center and the faster it moves, the less the AGRE module relies on it. This indicates that the AGRE module has learned to lower its reliance on rapidly shifting and peripheral gaze solely from the relevant intervals, i.e., it captures the underlying behavior of gaze beyond a surface pattern of the question.

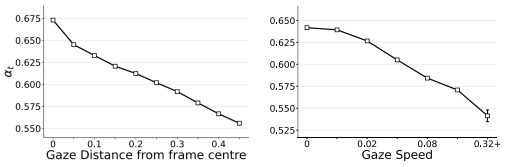


Figure 5: Average reliance on gaze $\alpha_t$ over the distance of the gaze coordinate from the frame center (left) and over gaze speed (right).

**Gaze Supervision for the Video LLM.** Beyond frame sampling, we ask whether the reliance $\alpha_t$ also identifies the gaze that is useful for the Video LLM itself. To this end, we follow Gaze-VLM (Pani & Yang, 2026a), which regularizes the attention of a Video LLM toward the wearer's gaze, and fine-tune InternVL3.5-4B (Wang et al., 2025a) on the data curated by the GQC pipeline with this regularization applied to a quarter of the frames within each relevant interval, selected either by the highest $\alpha_t$ (High-$\alpha_t$ gaze) or by the lowest $\alpha_t$ (Low-$\alpha_t$ gaze), where None denotes the model without fine-tuning. To isolate the effect of gaze supervision, all models are given the same frames selected by ReGaze at inference. As shown in Tab. 2, High-$\alpha_t$ gaze yields the best accuracy on both benchmarks, whereas Low-$\alpha_t$ gaze falls below None on EgoExoLearn, where gaze is noisier as observed in Sec. 5.2, and improves over None only marginally on EgoEverything, where gaze is broadly useful. This indicates that the same amount of gaze supervision helps or hurts depending on which gaze is selected, and that $\alpha_t$ serves as a measure for selecting the gaze that helps, which extends its use beyond frame sampling.

Table 2: Performance over different gaze supervision selected by $\alpha_t$.

| Gaze supervision | EgoExoLearn | EgoEverything |
|---|---|---|
| None | 36.29 | 39.43 |
| Low-$\alpha_t$ gaze | 35.91 | 40.16 |
| **High-$\alpha_t$ gaze** | **38.00** | **40.85** |

[5]Gaze coordinates are normalized to $[0, 1]$ by the frame width and height. The distance is measured from the frame center $(0.5, 0.5)$, and the speed is the Euclidean displacement between the gaze coordinates of consecutive frames.

### 5.4 Qualitative Results

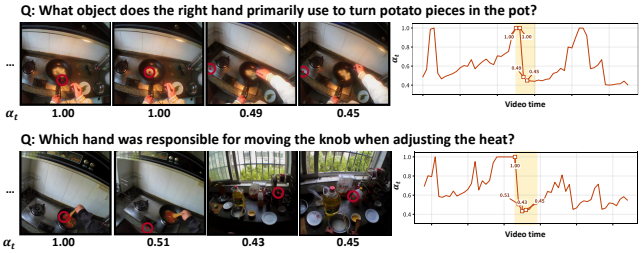


Figure 6: Qualitative results of the learned reliance $\alpha_t$.

To investigate whether the learned $\alpha_t$ relies on gaze where the gazed region is relevant to the question and suppresses it otherwise, we visualize frames of EgoExoLearn with the wearer's gaze and the trajectory of $\alpha_t$ over the video in Fig. 6, where the shaded region marks the visualized frames. As shown in Fig. 6, we observe that $\alpha_t$ is high while the wearer looks at the region relevant to the question, and drops as the gaze moves away from it, even when the evidence for the answer remains visible in the frame. Moreover, over the video, $\alpha_t$ neither stays at a fixed value nor collapses to either extreme of Fig. 1(b): it rises sharply only at a few moments and otherwise remains at a moderate level. This indicates that, although trained solely on the relevant intervals without any label of gaze usefulness, the AGRE module adjusts the reliance on gaze from frame to frame according to how the gazed region relates to the question, a behavior that no fixed $\alpha$ can express. Furthermore, the reliance changes with the question when the frame and gaze are held fixed (see Appendix B.5), i.e., the same fixation is trusted for one question and suppressed for another. As a result, when the gaze turns into noise, the frame is scored more by its frame feature, so that the evidence outside the gazed region remains reachable by the frame sampler.

### 5.5 Experiments on WearerQA

Finally, we evaluate ReGaze on WearerQA to assess it in the real-world wearable-assistant scenario. Since a wearer asks such questions without answer options, we evaluate them in an open-ended manner: the Video LLM (Wang et al., 2025a) answers each question from the frames selected by each sampler, and an LLM judge (OpenAI, 2026), given the question, determines whether the predicted answer matches the wearer-written one (see Appendix C.3 for the prompt). As shown in Tab. 3, ReGaze achieves the highest accuracy. Notably, the gaze-only sampler outperforms the frame-only sampler, as on EgoEverything, even though no question in WearerQA is generated around the gaze. This is consistent with how the questions are written: wearers recall their experience through what they attended to, so their questions naturally tend to concern the gazed region. Nevertheless, not every question can be answered from the gazed region, and relying solely on gaze can miss questions whose evidence lies in the surrounding scene. By relying on gaze where the gazed region relates to the question and leaning more on the frame feature otherwise, ReGaze outperforms both single-feature samplers as well as Fixed $\alpha$ and Maxsim. This suggests that, in the real-world scenario, gaze is useful for many of the questions a wearer asks, yet insufficient on its own, which makes learning how much to rely on it all the more important.

Table 3: Performance on WearerQA.

| Method | Acc. |
|---|---|
| InternVL3.5-8B | 34.34 |
| + Frame-only sampler | 37.15 |
| + Gaze-only sampler | 37.95 |
| + Fixed $\alpha$ | 38.76 |
| + Maxsim | 37.75 |
| + ReGaze | **39.56** |

## 6 Conclusion

In this paper, we revisited the implicit assumption that the gaze recorded by wearable devices is useful regardless of the task, and showed that, in question answering, its usefulness depends on the question and the frame, so that the same fixation can be evidence for one question yet noise for another. Based on this finding, we proposed ReGaze, which learns how much to rely on gaze for each question and frame through the AGRE module, trained with a ranking loss and a noise suppression loss on the relevant intervals curated by the GQC pipeline. On three egocentric QA benchmarks and WearerQA, ReGaze outperforms state-of-the-art frame samplers and naive uses of gaze, including the best fixed reliance.

Beyond performance, we believe this work offers a broader insight into how a wearable assistant should treat human signals. Gaze tells where the wearer looked, not what the wearer will ask, and its value therefore lies not only in exploiting it but, as our analysis suggests, even more in knowing when to set it aside. Moreover, this judgment can be learned without any label, only from where the answers lie, and the learned reliance extends beyond frame sampling to identify the gaze that helps a Video LLM itself. We hope this perspective encourages future work to treat gaze, and human signals in general, not as ground truth to follow but as evidence, i.e., to learn when to trust the eyes.

Regarding limitations and future work, please refer to Appendix D.

## USE OF AI ASSISTANTS

Throughout this paper, we used AI assistants (e.g., Claude, GPT) for limited assistance with code completion and language polishing of the manuscript. All research ideas, experimental design, analysis, and conclusions are the authors' own. No AI assistant was used to generate experimental results or scientific claims.

## ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. Apart from WearerQA, all datasets and pre-trained models used in our experiments are publicly available and used under their licenses. WearerQA was collected from five adult participants who gave informed consent; faces and license plates were anonymized with EgoBlur (Raina et al., 2023) (Appendix C.1). WearerQA will be released for research purposes upon acceptance.

## REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, we describe the experimental setup and implementation details in Section 5.1 and Appendix B.3, respectively.

## REFERENCES

Samiul Alam, Shakhrul Iman Siam, Michael J Proulx, James Fort, Richard Newcombe, Hyo Jin Kim, and Mi Zhang. Supermemory-vqa: An egocentric visual question-answering benchmark for long-horizon memory. *arXiv preprint arXiv:2606.00825*, 2026.

Apple Inc. Apple vision pro. `https://www.apple.com/apple-vision-pro/`, 2024. Accessed: November 28, 2024.

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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL `https://arxiv.org/abs/2502.13923`.

Chris Baumann and Kai Dierkes. Neon accuracy test report, 2023.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024.

Qinyu Chen and Jiawen Qi. Eye gaze tells you where to compute: Gaze-driven efficient vlms. *arXiv preprint arXiv:2509.16476*, 2025.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*, volume 2025, pp. 18227–18246, 2025.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: the dataset. In *European conference on computer vision*, pp. 753–771. Springer, 2018.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 42115–42127, 2025.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

Bo Fang, YuXin Song, Haoyuan Sun, Qiangqiang Wu, Wenhao Wu, and Antoni Chan. Threading keyframe with narratives: Mllms as strong long video comprehenders. In *International Conference on Learning Representations*, volume 2026, pp. 63417–63431, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

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.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-llm based video frame selection for efficient video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13702–13712. IEEE, 2025.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22072–22086. IEEE, 2024.

Daeun Lee, Subhojyoti Mukherjee, Branislav Kveton, Ryan A Rossi, Viet Dac Lai, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, and Mohit Bansal. Streamgaze: Gaze-guided temporal reasoning and proactive understanding in streaming videos. *arXiv preprint arXiv:2512.01707*, 2025.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026a.

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.

Yiwei Li, Zihao Wu, Yanjun Lv, Hanqi Jiang, Weihang You, Zhengliang Liu, Dajiang Zhu, Xiang Li, Quanzheng Li, Tianming Liu, et al. Thinking with gaze: Sequential eye-tracking as visual reasoning supervision for medical vlms. *arXiv preprint arXiv:2603.06697*, 2026b.

Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pp. 445–465. Springer, 2024.

Diako Mardanbegi, Nicholas Hylands, Neil Sarkar, and Nino Zahirovic. Gazelog: Optimizing eye-tracking with fixation keyframes & llm insights. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, pp. 1–8, 2025.

Michele Mazzamuto, Antonino Furnari, Yoichi Sato, and Giovanni Maria Farinella. Gazing into missteps: Leveraging eye-gaze for unsupervised mistake detection in egocentric videos of skilled human activities. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8310–8320. IEEE, 2025.

OpenAI. GPT-5.4 Thinking System Card. Technical report, OpenAI, March 2026. URL https://deploymentsafety.openai.com/gpt-5-4-thinking/gpt-5-4-thinking.pdf. Accessed: 2026-09-16.

Anupam Pani and Yanchao Yang. Gaze-vlm: Bridging gaze and vlms through attention regularization for egocentric understanding. *Advances in Neural Information Processing Systems*, 38: 163544–163577, 2026a.

Anupam Pani and Yanchao Yang. Gaze-regularized vision-language-action models for robotic manipulation, 2026b. URL https://arxiv.org/abs/2603.23202.

Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. *Advances in Neural Information Processing Systems*, 38, 2026.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23901–23913. IEEE, 2025.

Trong Thang Pham, Hien Nguyen, and Ngan Le. Gazeqwen: Lightweight gaze-conditioned llm modulation for streaming video understanding. *arXiv preprint arXiv:2603.25841*, 2026.

Pico Interactive. Neo3 pro / pro eye: Leading all-in-one 6DoF VR headset for enterprises. https://www.picoxr.com/nl/products/neo3-pro-eye, 2024. Accessed: November 28, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27b dense model, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Francesco Ragusa, Michele Mazzamuto, Rosario Forte, Irene D'Ambra, James Fort, Jakob Engel, Antonino Furnari, and Giovanni Maria Farinella. Ego-extra: video-language egocentric dataset for expert-trainee assistance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4438–4450, 2026.

Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, et al. Egoblur: Responsible innovation in aria. *arXiv preprint arXiv:2308.13093*, 2023.

Jun Rekimoto. Gazellm: Multimodal llms incorporating human visual attention. In *Proceedings of the Augmented Humans International Conference 2025*, pp. 302–311, 2025.

Baifeng Shi, Stephanie Fu, Long Lian, Hanrong Ye, David Eigen, Aaron Reite, Jan Kautz, Boyi Li, David M Chan, Trevor Darrell, et al. Attend before attention: Efficient and scalable video understanding via autoregressive gazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17022–17034, 2026.

Ajan Subramanian, Sumukh Bettadapura, and Rohan Sathish. Real eyes realize faster: Gaze stability and pupil novelty for efficient egocentric learning. *arXiv preprint arXiv:2603.04098*, 2026.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Qiance Tang, Ziqi Wang, Jieyu Lin, Ziyun Li, Barbara De Salvo, and Sai Qian Zhang. Egoeverything: A benchmark for human behavior inspired long context egocentric video understanding in ar environment. *arXiv preprint arXiv:2604.08342*, 2026.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128. IEEE, 2025.

Benjamin W Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of vision*, 7(14):4–4, 2007.

Varjo. Varjo – the world's most advanced virtual and mixed reality. `https://varjo.com/`, 2025. Accessed: 2025-09-04.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Zeyu Wang, Baiyu Chen, Kun Yan, Hongjing Piao, Hao Xue, Flora D Salim, Yuanchun Shi, and Yuntao Wang. Resolving ambiguity in gaze-facilitated visual assistant interaction paradigm. *arXiv preprint arXiv:2509.21980*, 2025b.

Chi Hsuan Wu, Ashutosh Kumar, and Kristen Grauman. Skillsight: Efficient first-person skill assessment with gaze. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38891–38903, 2026.

Kun Yan, Zeyu Wang, Lei Ji, Yuntao Wang, Nan Duan, and Shuai Ma. Voila-a: Aligning vision-language models with user's gaze attention. *Advances in neural information processing systems*, 37:1890–1918, 2024.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28885–28900. IEEE, 2025.

Linli Yao, Haoning Wu, Kun Ouyang, Yuanxing Zhang, Caiming Xiong, Bei Chen, Xu Sun, and Junnan Li. Generative frame sampler for long video understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17900–17917, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701. IEEE, 2025.

Zirui Zhu, Hailun Xu, Yang Luo, LIU YONG, Kanchan Sarkar, Zhenheng Yang, and Yang You. Focus: Efficient keyframe selection for long video understanding. In *International Conference on Learning Representations*, volume 2026, pp. 98463–98486, 2026.

# *Supplementary Material*

*- When to Trust the Eyes: Learning Adaptive Gaze Reliance in Egocentric Video -*

**A Additional Related Work** **15**

**B Experiments** **15**

B.1 Details of Datasets . . . . . 15

B.2 Additional Baseline Details . . . . . 16

B.3 Additional Implementation Details . . . . . 16

B.4 Effect of the Noise Suppression Loss on $\alpha_t$ . . . . . 17

B.5 Question-Dependent Reliance on the Same Frame . . . . . 17

B.6 Effect of Training the Encoder . . . . . 18

B.7 Effect of Temporal Context and Training Data . . . . . 18

**C Details of WearerQA** **18**

C.1 Collection Protocol . . . . . 18

C.2 Video Duration and Question Types . . . . . 19

C.3 Evaluation Protocol . . . . . 19

**D Limitations and Future Work** **21**

Table 4: Datasets used for training and evaluation. Unlike the existing benchmarks, whose questions are created by third parties from the video, WearerQA consists of questions written by the wearers themselves from memory, reflecting the real-world scenario in which a wearer asks an assistant about their own experience.

| Dataset | Scenario | Eye tracker | Videos | Avg. Dur. | Questions | Question source |
|---|---|---|---|---|---|---|
| *Training (curated by the GQC pipeline)* | | | | | | |
| HD-EPIC | Home cooking | Project Aria | 148 | 16.0 min | 19K | LLM |
| SuperMemory-VQA | Daily activities | Project Aria | 81 | 36.7 min | 22K | LLM |
| *Evaluation on existing benchmarks* | | | | | | |
| EgoExoLearn | Cooking, lab procedures | Pupil Invisible | 421 | 9.8 min | 4.2K | Third party |
| EgoEverything | Daily indoor/outdoor | Project Aria | 348 | 17.2 min | 6K | Third party |
| EgoGazeVQA | Kitchen, living room, etc. | Pupil Invisible / Project Aria / SMI | 263 | 8.3 min | 1.7K | Third party |
| *Evaluation in the real-world scenario* | | | | | | |
| WearerQA (Ours) | Daily activities | Pupil Labs Neon | 30 | 10.2 min | 498 | Wearer |

# A Additional Related Work

**Egocentric Video Understanding.** Egocentric video understanding has traditionally targeted the wearer's ongoing activity, e.g., action recognition, anticipation, and hand-object interaction (Damen et al., 2018; Grauman et al., 2022). Recently, wearable assistants have shifted the focus to question-driven understanding: the model must answer whatever the wearer asks about their experience, whether by recalling events from a long video (Grauman et al., 2022; Chandrasegaran et al., 2024; Yang et al., 2025) or by responding as the video streams in (Lee et al., 2025). In either setting, answering with a Video LLM hinges on which frames the model gets to see, since it can consume only a few dozen frames from recordings that last minutes to hours, and uniform sampling routinely misses the question-relevant evidence. To address this, question-aware frame sampling has emerged, which scores frames by their similarity to the question using a vision-language encoder (Radford et al., 2021; Sun et al., 2023). For example, AKS (Tang et al., 2025) refines the selection via adaptive temporal coverage, Nar-KFC (Fang et al., 2026) via joint relevance–diversity optimization with interleaved narratives, and FOCUS (Zhu et al., 2026) via bandit-style exploration over temporal regions. However, all of them identify question-relevant frames from frame content alone, leaving the wearer's gaze, which comes for free in wearable settings, unused. We extend question-aware frame sampling with gaze and, as discussed above, learn how much to rely on it for each question and frame.

# B Experiments

## B.1 Details of Datasets

**Training data.** For the GQC pipeline, we curate the training data from the videos of HD-EPIC (Perrett et al., 2025) and SuperMemory-VQA (Alam et al., 2026), both recorded with Project Aria glasses (Engel et al., 2023). HD-EPIC consists of 148 videos recorded by nine participants in their own kitchens over at least three days, whereas SuperMemory-VQA consists of 49.6 hours recorded by ten participants in a simulated home, including indoor and outdoor activities with other participants. We use only their videos and gaze, not the VQA benchmarks they provide.

**Evaluation benchmarks.** For EgoExoLearn (Huang et al., 2024), we use the multiple-choice questions of the video QA annotations[6] on its egocentric videos, which record participants replicating cooking and laboratory procedures. EgoEverything (Tang et al., 2026) comprises five-option questions over the videos of the Aria Everyday Activities (Lv et al., 2024) and Nymeria (Ma et al., 2024), whose target objects are sampled around the recorded gaze. EgoGazeVQA (Peng et al., 2026) comprises five-option questions on clips from Ego4D (Grauman et al., 2022), Ego-Exo4D (Grauman et al., 2024), and EGTEA Gaze+ (Li et al., 2021), generated by a Video LLM (Bai et al., 2025) from frame captions and verified by human annotators; for each question, we sample frames from its original source video. For all benchmarks, we select $K = 32$ frames at 1 fps, and all questions are used only for evaluation. Note that the training videos are newly collected by HD-EPIC and SuperMemory-VQA and share no source with any evaluation benchmark.

[6] https://github.com/OpenGVLab/EgoExoLearn/tree/main/Video_QA

**Gaze characteristics.** The datasets are recorded by different wearers with different eye trackers, i.e., Project Aria (the training data, EgoEverything, and the Ego-Exo4D part of EgoGazeVQA), Pupil Invisible (EgoExoLearn and the Ego4D part of EgoGazeVQA), SMI (the EGTEA Gaze+ part of EgoGazeVQA), and Pupil Labs Neon (WearerQA), at different sampling rates, e.g., 30 Hz for EGTEA Gaze+, 60 Hz for Project Aria, and 120 Hz for EgoExoLearn, all of which we align to the 1 fps frames.

### B.2 Additional Baseline Details

To examine whether processing more frames or specializing in egocentric videos can substitute for selecting the frames, we additionally evaluate Video LLMs that take longer inputs or are fine-tuned on egocentric videos. Specifically, we evaluate EgoGPT-7B (Yang et al., 2025), which is fine-tuned on egocentric videos, and MiniCPM-V 2.6 (Yao et al., 2024), LongVILA-7B (Chen et al., 2025), and VideoLLaMA3-7B (Zhang et al., 2025), which support long video inputs, together with Qwen3.5-9B given twice as many frames. Each model receives 64 or 128 frames uniformly sampled from the video, with the same prompt as in Sec. 5.1 and without gaze. As shown in Tab. 1, we observe that all of them fall below ReGaze on average, even though they process two to four times as many frames as the 32 frames selected by ReGaze. In particular, doubling the frames of Qwen3.5-9B improves the accuracy less than selecting 32 frames with ReGaze. This indicates that which frames the Video LLM sees matters more than how many it sees, and that the benefit of ReGaze is not replaced by fine-tuning on egocentric videos.

**Why do state-of-the-art frame sampling methods help little?** In Table 1, FOCUS, AKS, and Nar-KFC change the average accuracy of the frame-only sampler marginally, even though they refine its top-$K$ selection. We attribute this to a mismatch between their design and the questions a wearer asks. These samplers spread the frame budget beyond the frames most similar to the question, through adaptive temporal coverage (AKS), joint relevance–diversity optimization (Nar-KFC), and exploration over temporal regions (FOCUS). Such spreading helps when the evidence is distributed over the video, as in benchmarks designed to require long-range understanding (Fu et al., 2025; Zhou et al., 2025). A wearer, however, typically asks about a specific moment of their experience, e.g., where they left an object (Grauman et al., 2022). For such questions, the bottleneck is how precisely the relevance score pinpoints the moment rather than how widely the selected frames cover the video, and frames spent on coverage displace frames around the evidence. ReGaze targets exactly this bottleneck: it keeps the top-$K$ selection and sharpens the relevance score of each frame with the wearer's gaze, which reveals where the wearer was engaged at that moment.

### B.3 Additional Implementation Details

**Reliance estimator and optimization.** The transformer encoder of the reliance estimator consists of two layers with eight heads and a hidden dimension of 512. We train the AGRE module with a learning rate of $3 \times 10^{-4}$ and a weight decay of $0.01$, using a batch size of 4 with 32 gradient accumulation steps, i.e., an effective batch size of 128. All experiments are run with deterministic operations for reproducibility. The training and inference are conducted on the NVIDIA RTX A6000 GPUs.

**Gaze alignment.** Since gaze is recorded at a higher rate than the 1 fps frames (Appendix B.1), we assign to each frame the gaze sample closest to its timestamp.

**Positive and negative frames.** Since the LLM in the GQC pipeline outputs the relevant interval at the level of the events described in text, described in Sec. 4.2, the interval may include frames that do not show the evidence for the question. To obtain more reliable supervision, we use as the positive the single frame in the interval that is most similar to the question, measured by Qwen3-VL-Embedding-8B (Li et al., 2026a), i.e., $|\mathcal{R}| = 1$ in Eqs. 4 and 5. This embedding model is used only to select the positives during training, not at inference. Using all frames in the interval as positives instead lowers the accuracy with InternVL3.5-8B from 38.91 to 38.30 on EgoExoLearn and from 41.82 to 41.25 on EgoEverything, which reflects the noise in the interval. For each positive, the negative is randomly sampled from the frames of the hard-negative spans provided by the GQC pipeline.

**Video LLM and baselines.** The Video LLM receives only the selected frames and the question, without any gaze information, so that the differences across samplers arise only from which frames

are selected. For FOCUS (Zhu et al., 2026), AKS (Tang et al., 2025), and Nar-KFC (Fang et al., 2026), we use their official code with the hyperparameters reported in their papers, together with the same EVA-CLIP (Sun et al., 2023) encoder as ReGaze.

### B.4 Effect of the Noise Suppression Loss on $\alpha_t$

To examine how the noise suppression loss $\mathcal{L}_{\text{NS}}$ affects the learned reliance $\alpha_t$, we compare the distribution of $\alpha_t$ over the frames of EgoExoLearn with and without the loss. As shown in Fig. 7, without the loss, $41.93\%$ of the frames receive $\alpha_t \geq 0.95$, i.e., ReGaze behaves as the gaze-only sampler on these frames. With the loss, this share drops to $22.93\%$, and the mass shifts from full reliance toward a moderate reliance instead. This indicates that the noise suppression loss curbs excessive reliance on gaze rather than discarding it: full reliance is kept for under a quarter of the frames, in contrast to the gaze-only sampler that relies fully on gaze for every frame, while the rest are scored with a moderate reliance. This also accounts for the accuracy drop without the loss in Sec. 5.3.

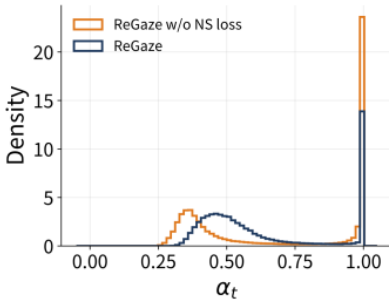


Figure 7: Distribution of the learned reliance $\alpha_t$ over the frames of EgoExoLearn with and without $\mathcal{L}_{\text{NS}}$.

Note that $\alpha_t$ rarely falls below $0.2$ with or without the loss, which is expected for three reasons. First, frames without recorded gaze are assigned $\alpha_t = 0$ by design and are excluded here. Second, the gaze feature is pooled from the same frame and is thus never entirely uninformative, as Fig. 1(b) shows that a small reliance outperforms none even as a fixed choice. Third, the noise suppression loss penalizes $\alpha_t$ in proportion to both the harm and $\alpha_t$ itself, so that its penalty vanishes as $\alpha_t$ approaches zero, consistent with its role of preventing excessive reliance rather than removing gaze. In addition, no frame receives a negative $\alpha_t$, so that Eq. 2 remains a convex combination of the two features, as discussed in Sec. 4.1.

### B.5 Question-Dependent Reliance on the Same Frame

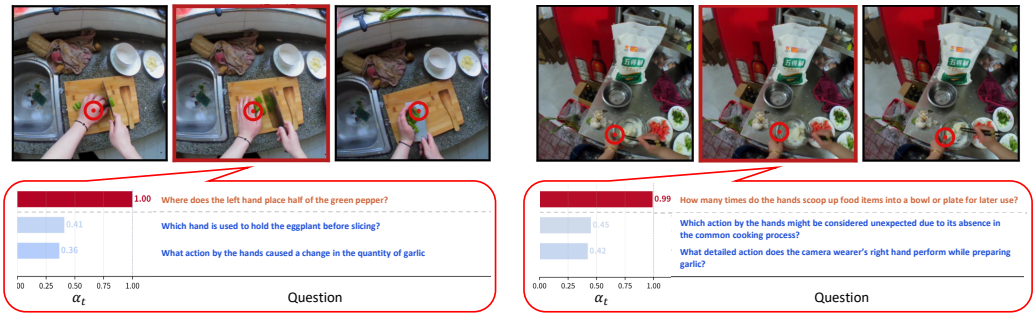


Figure 8: Learned reliance $\alpha_t$ on the same frame and gaze under different questions.

In Sec. 5.4 of the main paper, we observe that the learned reliance $\alpha_t$ changes from frame to frame for a given question. To complement this, we investigate whether the reliance also changes with the question when the frame and gaze are held fixed, i.e., whether the AGRE module captures the question dependence of gaze usefulness illustrated in Fig. 1(a). To this end, we take a single frame from each video and estimate $\alpha_t$ on it for several questions asked about the same video. As shown in Fig. 8, we observe that the same frame and gaze receive a high reliance for the question whose evidence lies in the gazed region, and a low reliance for the questions whose evidence lies outside it, whether elsewhere in the frame or at another moment of the video. This indicates that ReGaze assigns the reliance not to the frame or the gaze alone, but to their relation with the given question, so that the same fixation is trusted for one question and suppressed for another, as discussed in Sec. 1. Taken together with Sec. 5.4, the reliance thus varies along both the question and the frame, which neither a fixed $\alpha$ nor a reliance determined by the gaze signal alone can capture.

### B.6 Effect of Training the Encoder

To examine the source of the performance gain in depth, we ask: *Does the gain of ReGaze come from exposing the sampler to egocentric videos rather than from the learned reliance?* To this end, we fine-tune the vision-language encoder on the data curated by the GQC pipeline, on which the AGRE module is trained, and compare fine-tuning the encoder alone under the frame-only sampler, i.e., without AGRE, fine-tuning it jointly with AGRE, and training AGRE with the encoder frozen, i.e., ReGaze. As shown in Fig. 9, we observe that neither fine-tuning the encoder alone nor fine-tuning it jointly with AGRE improves over the frame-only sampler with the frozen encoder, whereas ReGaze with the frozen encoder achieves the highest accuracy. This indicates that the gain of ReGaze does not come from adapting the encoder to the egocentric training data, but from learning how much to rely on gaze. We hypothesize that fine-tuning the encoder on the curated data compromises the generality of its pretrained embedding space, on which the relevance of Eq. 2 is measured. Moreover, when trained jointly with AGRE, the encoder can reduce the ranking loss by reshaping the frame and gaze features themselves rather than by adjusting $\alpha_t$, and it also shifts the frame-only reference $s_t(0)$ of the noise suppression loss, which is designed to stay fixed (Sec. 4.1). These observations support keeping the encoder frozen in ReGaze.

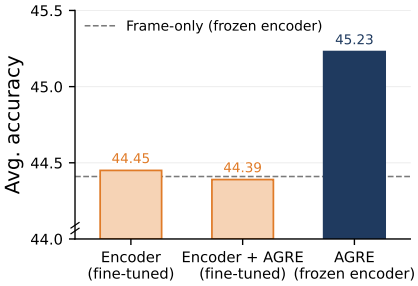


Figure 9: Average accuracy over the three benchmarks when the vision-language encoder is fine-tuned on the curated data.

### B.7 Effect of Temporal Context and Training Data

To examine whether the AGRE module needs the local-context window of Sec. 4.1, we compare it with a single-frame variant that replaces the transformer encoder with an MLP and thus estimates $\alpha_t$ from the current frame alone. We train both on each of the two curated datasets and on their union, and report the average accuracy over the three benchmarks.

As shown in Fig. 10, ReGaze with the local window AGRE outperforms Maxsim, which is described in Sec. 5.1, regardless of the training data. In contrast, the single-frame variant varies considerably with the training data: while it is competitive when trained on SuperMemory-VQA alone, it falls below Maxsim in the other two cases, and even below the frame-only sampler when trained on both datasets. Moreover, combining the two datasets improves the accuracy only with the local window, and degrades it for the single-frame variant.

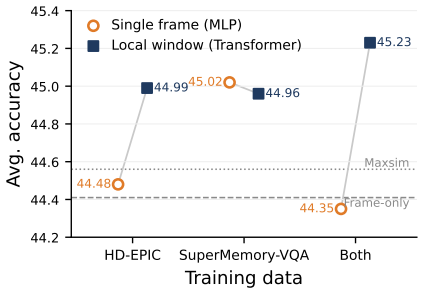


Figure 10: Average accuracy over the three benchmarks, when the reliance is estimated from a single frame (MLP) or from a local window of preceding frames (Transformer), for each training data.

This indicates that the preceding frames make the learned reliance robust to the training data. We hypothesize that this is because the two datasets are recorded by different wearers whose gaze differs. From a single frame alone, such variation is hard to separate from the usefulness of gaze, whereas the local window lets the AGRE module draw on how the gaze evolves, e.g., whether it has settled, a cue shared across wearers. As a result, only with the local window does the AGRE module benefit from more diverse training data.

## C Details of WearerQA

### C.1 Collection Protocol

WearerQA is designed to reflect what a wearer would genuinely ask an assistant about their own experience, rather than what is answerable from a video. To this end, we gave each participant the guideline in Tab. 5 and collected the questions in two steps. In the first step, participants wrote questions 15–45 minutes after recording, from memory alone and without rewatching the video, since a

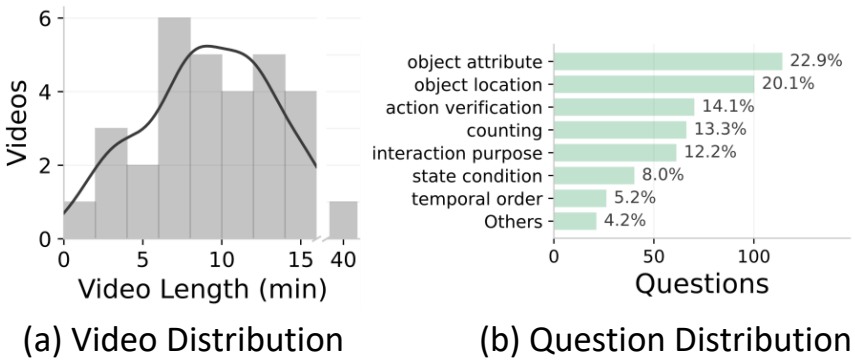


Figure 11: Statistics of WearerQA. (a) Distribution of recording durations. (b) Distribution of question types, derived from the collected questions rather than from a predefined taxonomy.

wearer asks about past experience as they remember it. We imposed no required question types, and the examples in the guideline were given only to convey the spirit of the task, so that the questions would not be steered toward particular categories. The only constraint was that each question should concern the participant's own actions or surroundings during the recording, excluding requests for advice or plans, which cannot be answered from the recording. In the second step, participants rewatched the video, marked whether each question could be answered from it, and wrote the final answer for those that could; questions whose answers could not be determined from the video, e.g., due to occlusion or low resolution, were removed. Finally, we manually checked every remaining question to confirm that it is answerable from the video.

**Privacy protection.** Before recording, participants followed a privacy checklist, e.g., removing documents and screens with personal information from view, and anyone who might appear in the video signed a consent form beforehand. After recording, we further anonymized the videos with EgoBlur (Raina et al., 2023), which blurs faces and license plates that may have been captured.

### C.2 Video Duration and Question Types

WearerQA contains 498 questions over 30 recordings, which last 10.2 minutes on average. As shown in Fig. 11(a), most videos last between 5 and 15 minutes.

Unlike benchmarks that define question types in advance and generate questions for each type, WearerQA imposes no question type on the participants, as in the real-world scenario. We therefore analyze the question types in the reverse direction: an LLM (OpenAI, 2026) derives categories from the collected questions and assigns each question to one of them, with rare categories merged into Others. The resulting distribution in Fig. 11(b) thus reflects what wearers ask rather than what a predefined taxonomy prescribes. Most questions concern objects, i.e., their attributes and locations, followed by whether an action was performed, counting, and the purpose of an interaction. This emphasis on the objects the wearer engaged with is consistent with our premise that recall is shaped by what one attended to, as discussed in Sec. 4.3. Moreover, questions that may require evidence from multiple moments, such as and temporal order, account for less than a fifth of the set, supporting our observation in Appendix B.2 that a wearer typically asks about a specific moment of their experience.

### C.3 Evaluation Protocol

Since WearerQA is evaluated in an open-ended manner, we use an LLM-as-a-judge (GPT5.4 (OpenAI, 2026)) to determine whether the prediction of the Video LLM conveys the same fact as the wearer-written answer, and report the fraction of questions judged correct. The judge receives only the question, the reference answer, and the prediction, without access to the video, and is instructed to judge agreement with the reference alone, so that it neither second-guesses the reference nor relies on knowledge of the video it does not have. We use the same judge and prompt, shown in Fig. 12, for all methods.

Table 5: **Wearer instruction for question collection.** The full guideline given to each wearer.

| Section | Instruction |
|---|---|
| **Setup** | You just finished recording your activity while wearing smart glasses. Imagine the glasses have an AI assistant that saw everything you saw.<br>15–45 minutes after recording, without rewatching the video, write down questions you would genuinely want to ask this assistant about what you did or saw during the recording. |
| **Before recording**<br>*Privacy check (1 minute)* | • Remove or cover: mail/parcel labels, ID cards, documents with names, and phone/laptop screens showing personal content.<br>• If another person may appear in the video, they must have signed the consent form beforehand.<br>• Avoid capturing your address signs, or car license plates, etc. |
| **Step 1**<br>*Write questions from memory* | No auxiliary agent is used for question generation. The wearer writes the questions. Questions are produced after a delay, from memory alone, mirroring how humans naturally recall past experience. Because human recall is driven by where attention was allocated, the resulting questions are implicitly grounded in gaze rather than in uniform frame sampling.<br>1. Rely only on your memory. Do not open the video yet.<br>2. Write at least 5 questions. There is no upper limit — write down as many questions as come to mind. We will filter them later.<br>3. There are no required types or categories. Ask whatever you are actually curious or unsure about — things you did, touched, saw, moved, or forgot.<br>4. Questions must be about your own actions or surroundings during the recording — not requests for advice, recipes, or future plans.<br>5. For each question, fill in the sheet:<br>• `question` — the question itself<br>*Examples* (these show the spirit, not a format to copy — ask what YOU actually wonder about):<br>• *Which shelf did I take my shoes from?*<br>• *Did I close the door carefully when I left?*<br>• *How many bags did I carry when I went out?* |
| **Step 2**<br>*Re-watch and verify* | After finishing Step 1, re-watch the video and complete the remaining columns for every question:<br>• `Answer Check` — can the answer be determined from the video?<br>– **Clear** — the video provides enough evidence to answer the question<br>– **Unclear** — the answer cannot be inferred from the video (e.g., low resolution, out of frame, occluded)<br>• `Final Answer` — the final correct answer based on the video. Fill this in whenever `Answer Check` is **Clear**. |

**System prompt**

```
You grade answers to questions about a first-person video you cannot
see. For each item you get the question, the REFERENCE answer (ground
truth) and a model's PREDICTION.

Mark correct=true when the prediction conveys the same fact as the
reference. Wording, extra detail, hedging and full sentences do not
matter. A near-synonym, a more or less specific form of the same
thing ('sofa' vs 'couch', 'on the kitchen island' vs 'the island'),
or the same number written differently all count as correct.

Mark correct=false when the prediction names a different object,
place, colour, count, person or outcome than the reference, when it
answers a different question, when it only says it cannot tell, or
when it hedges across both a right and a wrong option without
committing.

Judge only agreement with the reference. Never reason about whether
the reference itself is right, and never use knowledge of the video
you do not have.
```

**User prompt**

```
Grade each item.

1. QUESTION: {question}
   REFERENCE: {ground_truth}
   PREDICTION: {model_output}
2. QUESTION: ...
```

Figure 12: Prompt for the LLM judge on WearerQA. {question}, {ground_truth}, and {model_output} are replaced with the question, the wearer-written answer, and the prediction of the Video LLM, respectively.

## D Limitations and Future Work

Following the frame rate commonly used for Video LLMs (Zhang et al., 2025), we sample frames at 1 fps and assign to each frame the gaze sample closest to its timestamp. The eye trackers in our data, however, record gaze at 30–120 Hz, and eye movements such as fixations and saccades unfold on a much finer timescale. Consequently, the AGRE module observes how gaze evolves only at the scale of seconds, e.g., whether the wearer keeps attending to a region or shifts to a new one, and cannot tell a stable fixation from a transient glance within a second. Nonetheless, our analysis in Sec. 5.3 suggests that even this coarse view lets the AGRE module lower its reliance on rapidly shifting gaze. We expect that modeling gaze dynamics at the native rate of the eye tracker, e.g., by detecting fixations and saccades and aggregating the gaze within each frame interval, would further help to tell useful gaze from noise, and leave it to future work.

Regarding the scope of this paper, we study gaze in question answering over recorded videos. Extending the learned reliance to streaming settings (Lee et al., 2025), in which an assistant must decide online whether to trust the current gaze, is a natural next step.