GAZELENS FOR EFFICIENT EGOCENTRIC VIDEO QUESTION ANSWERING WITH SIMPLE GAZE-DRIVEN ADAPTIVE TOKEN REDUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Smart glasses can help users understand their activities and surroundings, but small vision-language models must preserve useful evidence within limited computation. This is difficult because redundant frames consume the visual budget, while aggressive spatial compression can discard relevant details. Although gaze identifies the wearer's focus, the question may require evidence from the surrounding scene. To address these constraints, we propose GazeLens, which combines frame selection with gaze- and question-guided token reduction. First, HOG transitions and image sharpness select frames before visual encoding. For each selected frame, Gaussian pooling then forms gaze-focused and complementary summaries. Because their relevance depends on the question, a lightweight gate blends them into a primary token. Together with a global scene token, this retains both focused and contextual summaries within two tokens per frame. On StreamGaze, GazeLens achieves higher macro and micro accuracy than the evaluated full-grid and token-reduction baselines, with lower input-forward FLOPs than the shared-backbone token-reduction baselines.

1 INTRODUCTION

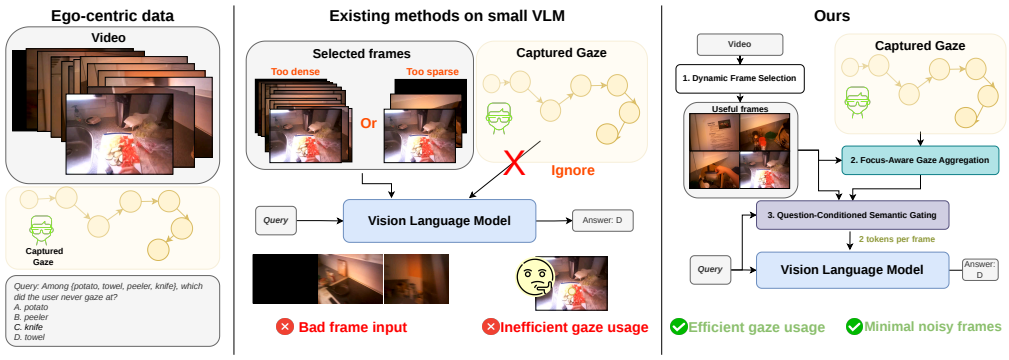


Figure 1: **Teaser of our gaze-driven adaptive token reduction framework for egocentric video QA.** *In the left panel,* egocentric video and gaze data are inherently noisy and redundant, with dense frames and unstable gaze signals. *In the middle panel,* existing VLM approaches suffer from inefficient frame selection (too dense or too sparse) and often ignore gaze or rely on computationally expensive adaptive attention mechanisms. *In the right panel,* our approach combines (1) **Dynamic Frame Selection**, (2) **Focus-Aware Gaze Aggregation**, and (3) **Question-Conditioned Semantic Gating**. Together, these modules produce two visual tokens per frame. This design reduces the visual token budget and model computation while improving performance under strict token budgets.

Smart glasses can help a wearer recall earlier actions, identify objects, or receive alerts about unfolding events (Lee et al., 2025). However, small vision-language models must retain useful evidence within limited compute and memory (Yao et al., 2024; Chu et al., 2024). Egocentric video makes this

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.

Table 1: Coverage of the design requirements. Adaptive frames denotes content-dependent selection before learned visual encoding. The other columns concern spatial compression. A check marks explicit support within the reduction mechanism, including whether compression is conditioned on the question.

Method	Adaptive frames	Gaze guidance	Peripheral context	Question guidance
GazeLLM (Rekimoto, 2025)	×	✓	×	×
LLaMA-VID (Li et al., 2024)	×	×	✓	✓
Voila-A (Yan et al., 2024)	×	✓	✓	×
GazeVLM (Yoon et al., 2025)	×	✓	✓	×
GazeLens (Ours)	✓	✓	✓	✓

difficult because repeated views consume the visual budget, while useful evidence may occupy only a small region or a brief interval.

Reducing visual tokens addresses only part of this cost. FastV and PyramidDrop discard tokens inside the language model (Chen et al., 2024b; Xing et al., 2024), after visual encoding and some language-model computation have already occurred. LLaMA-VID compresses each encoded frame to two tokens (Li et al., 2024), but still encodes every sampled frame. LongVU reduces both temporal and spatial redundancy (Shen et al., 2024), using DINOv2 features computed before frame removal. These choices save downstream computation while retaining the cost of extracting the features used for reduction. This motivates selecting frames before learned visual encoding when inexpensive image statistics can identify redundancy.

Deciding what to preserve within a frame introduces a second difficulty. Gaze provides a spatial prior for the wearer's focus (Li et al., 2021; Zhang et al., 2017; Lai et al., 2024). However, a question about a fixated object and a question about the surrounding scene require different evidence. A compact representation must therefore preserve both local and contextual information, with a balance that depends on the question.

Table 1 compares how reduction mechanisms address these requirements. LLaMA-VID conditions aggregation on the question without a gaze prior. GazeLLM crops around gaze (Rekimoto, 2025), removing off-gaze evidence. GazeVLM preserves global and local image views, then selects gaze-near and surrounding tokens with a preset allocation (Yoon et al., 2025). However, this allocation does not depend on the question. Voila-A retains gaze-conditioned features through a learned Perceiver (Yan et al., 2024), adding attention operations after encoding without reducing encoder inputs. These distinctions motivate combining frame selection before encoding with question-conditioned spatial compression that preserves scene context.

GazeLens implements these choices through three modules (Figure 1). To avoid encoding redundant or blurred candidates, **Dynamic Frame Selection** uses HOG transitions and Laplacian sharpness to select frames without learned parameters. To keep evidence beyond the fixation available, **Focus-Aware Gaze Aggregation** forms soft gaze-focused and complementary summaries. **Question-Conditioned Semantic Gating** then controls their mixture according to the question. This mixture forms the primary token, while a second, global token retains a scene summary regardless of the gate. Together, the modules reduce how many frames are encoded and control which information survives within two tokens per selected frame.

We make three contributions.

- A parameter-free frame selector reduces encoder inputs using HOG transitions and image sharpness.
- Gaze- and question-conditioned pooling preserves focused, complementary, and global summaries within two visual tokens per selected frame.
- Our 500M configuration improves StreamGaze Macro and Micro accuracy over the evaluated baselines while reducing model-forward FLOPs. EgoGazeVQA transfer retains the frame-efficiency benefit.

2 RELATED WORKS

Gaze-guided VLMs. GazeLLM uses gaze-centered cropping (Rekimoto, 2025), while Voila-A learns gaze-conditioned visual representations (Yan et al., 2024). GazeVLM combines global/local image preprocessing with gaze-guided token selection after encoding (Yoon et al., 2025). It preserves surrounding evidence through sampled context tokens, but its image-based reduction does not select video frames or condition token allocation on the question. GazeLens instead selects frames before encoding and blends focused and complementary features according to the question within two tokens per frame. StreamGaze (Lee et al., 2025) evaluates this balance through Past, Present, and Proactive questions.

Token reduction and frame selection. FastV and PyramidDrop prune visual tokens within the language model (Chen et al., 2024b; Xing et al., 2024), whereas LLaMA-VID produces a compact representation before language-model input (Li et al., 2024). For video, LongVU and FrameFusion also exploit temporal redundancy in visual features (Shen et al., 2024; Fu et al., 2024). Frame-Voyager and M-LLM Frame Selector learn to choose frames using question relevance and model-derived supervision (Yu et al., 2024; Hu et al., 2025). Across these approaches, the stage of reduction determines which computation can be avoided. We therefore place frame selection before learned encoding and use gaze to allocate the retained spatial-token budget.

Streaming MLLMs. VideoLLM-online, ViSpeak, Dispider, Flash-VStream, and LiveVLM address interaction with arriving video (Chen et al., 2024a; Fu et al., 2025; Qian et al., 2025; Zhang et al., 2024; Ning et al., 2025), with StreamingBench evaluating streaming comprehension (Lin et al., 2024). Our study concerns the visual input budget for each QA example. Here, frame selection controls the amount of encoded video, and gaze-conditioned pooling controls its spatial representation. Selected frames are encoded independently.

3 METHOD

Each module in Figure 2 is motivated by a distinct failure mode of prior work. Uniform temporal sampling wastes the token budget on redundant or blurry frames before they ever reach the encoder, so Section 3.1 filters the candidate pool first. Hard gaze crops are brittle to sensor noise and silently drop peripheral context, so Section 3.2 instead builds complementary soft focus and unfocus tokens via Gaussian pooling. Because the right balance between gaze focus and peripheral evidence varies with the question, Section 3.3 introduces a scalar gate conditioned on question semantics to blend these tokens per sample. Section 3.4 details the final token assembly and training objective.

3.1 DYNAMIC FRAME SELECTION

A fixed frame count imposes a single global trade-off. Too few frames miss rare action moments, and too many inflate the LM-side token budget with redundant content. We instead use a variable count $t_{\min} \leq T \leq t_{\max}$, spending the budget only on frames where visual content actually changes. This flexibility favors frames that are both visually distinct from their neighbors and free of motion blur. We operationalize these two criteria with HOG transition cues and Laplacian sharpness scores, which are parameter-free, fast to compute, and require no training data.

HOG transition cue. Sharpness alone cannot distinguish a novel sharp frame from a sharp duplicate, so HOG transition cues are added to suppress consecutive frames with near-identical visual content. Let (I_{i-1}, I_i) denote consecutive frames after striding. Both are resized to a fixed HOG window $H_r \times W_r$ and converted to grayscale. Let $\phi(I_i) \in \mathbb{R}^d$ be the single-window HOG descriptor of I_i. We record the L_2 distance between consecutive descriptors at the later index i,

$$\delta_i \;=\; \big\|\phi(I_i) - \phi(I_{i-1})\big\|_2. \tag{1}$$

Large δ_i indicate strong layout or motion change between successive kept frames. HOG encodes gradient orientations, making δ_i robust to the illumination shifts and auto-exposure changes common in egocentric footage.

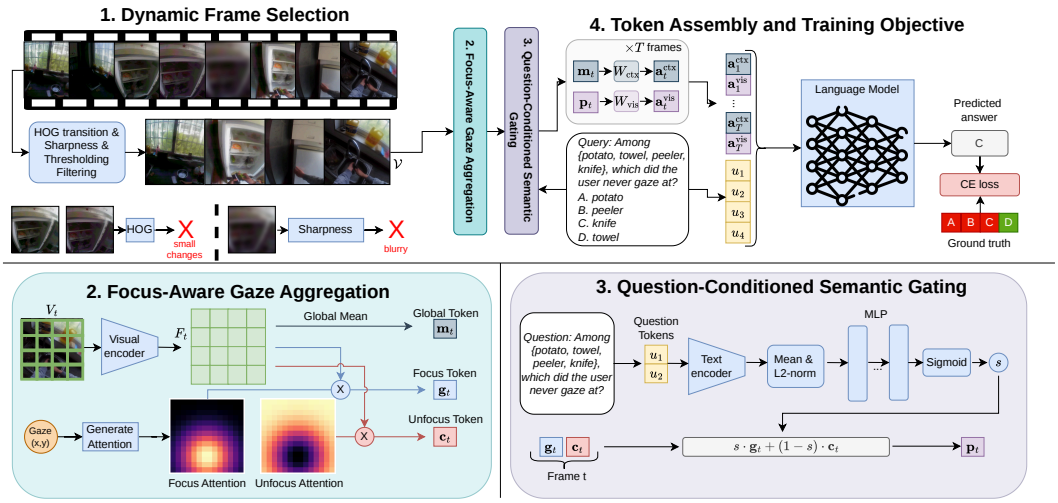


Figure 2: **Overview of the proposed focus-aware, question-conditioned token reduction framework.** *(1) Dynamic Frame Selection.* A continuous video stream is filtered using HOG transition and sharpness cues to retain informative frames while discarding small-change and blurry frames. This filtering process produces a dynamic input sequence $\mathcal{V}$. *(2) Focus-Aware Gaze Aggregation.* Each selected frame is encoded into spatial features and combined with gaze (x_t, y_t) to produce a focus token $\mathbf{g}_t$, an unfocus token $\mathbf{c}_t$, and a global token $\mathbf{m}_t$. *(3) Question-Conditioned Semantic Gating.* Question token embeddings are mean-pooled and L2-normalized to $\mathbf{t}$. A two-layer MLP and sigmoid then yield a scalar gate $s \in [0, 1]$. The gate modulates gaze-focused versus contextual information as $\mathbf{p}_t = s \cdot \mathbf{g}_t + (1 - s) \cdot \mathbf{c}_t$. *(4) Token Assembly and Training Objective.* The per-frame tokens $\mathbf{p}_t$ and $\mathbf{m}_t$ are linearly projected, concatenated with the question, and fed into a language model to predict the answer, optimized with a cross-entropy (CE) loss.

Sharpness cue. Independently, each frame is resized to $H_r \times W_r$ grayscale and scored by the variance of its 3×3 Laplacian,

$$\ell_i \;=\; \mathrm{Var}\big(\mathrm{Lap}_{3\times3}(I_i)\big). \tag{2}$$

Higher ℓ_i correspond to sharper, less motion-blurred content. Motion blur suppresses high-frequency edges and collapses the Laplacian variance to near zero, giving a calibration-free score that drops precisely on frames with no recoverable spatial detail.

Scoring and selection. Transition cues alone tend to peak on the blurred interstitial frame between two sharp moments, while sharpness scores alone favour static frames that carry no new information, so combining them selects frames that are both informationally novel and spatially recoverable. For each cue $c \in \{\delta, \ell\}$, let $\Omega_c = \{i : c_i\}$, $N_c = |\Omega_c|$, and define the per-clip mean $\mu_c = \frac{1}{N_c}\sum_{i\in\Omega_c} c_i$, and variance $\sigma_c^2 = \frac{1}{N_c}\sum_{i\in\Omega_c}(c_i - \mu_c)^2$. Per-clip normalized scores are $\tilde{c}_i = \dfrac{c_i - \mu_c}{\sigma_c}$. We combine the two cues with fixed weights w_{hog} and w_{sh} to obtain

$$S_i \;=\; \sum_{c\in\{\delta,\ell\}} w_c\, \tilde{c}_i. \tag{3}$$

We retain frame candidates with $S_i > z_{\mathrm{thresh}}$ and sort them by S_i. If more than $t_{\max}$ indices remain, we truncate the set to the top-S_i entries to enforce a hard cap on encoder cost and the downstream token budget.

Each retained index is then replaced by $\arg\max_{j:|j-i|\le\tau} \ell_j$, where τ denotes the scan window. This adjustment is motivated by the observation that pairwise change cues often peak on motion-blurred transition frames, whereas larger ℓ_j values typically occur on nearby stable frames.

If fewer than $t_{\min}$ indices remain, e.g., in nearly static clips, we supplement the set with uniformly spaced frames to ensure the VLM still receives a minimum level of temporal information.

Both z_{thresh} and τ are hyperparameters selected based on training set analysis. The selected frames define our dynamic input sequence $\mathcal{V} = \{V_t\}_{t=0}^{T-1}$. Both cues are parameter-free image statistics.

Temporal reduction occurs before visual encoding. Each selected frame is then encoded independently, and tokens are not reused between frames.

3.2 Focus-Aware Gaze Aggregation

For the selected T frames, let $\mathcal{G} = \{(x_t, y_t)\}_{t=0}^{T-1}$ denote the gaze points normalized to $[0,1]^2$. We then encode each frame t to obtain spatial patch features $F_t \in \mathbb{R}^{H\times W\times D}$, which provide the input to gaze-guided pooling.

Hard-cropping at the raw gaze point (x_t, y_t) is brittle, as sensor jitter of a few pixels selects an entirely different patch and the crop discards all peripheral context. Instead, we generate an isotropic Gaussian map over the $H \times W$ feature grid, which smoothly down-weights patches away from the gaze point, preserving peripheral context with graduated weights.

$$(\widetilde{h}_t)_{i,j} = \exp\!\left(-\frac{(i - y_t(H-1))^2 + (j - x_t(W-1))^2}{2\rho^2}\right), \qquad \rho = \max\!\left(\tfrac{\max(H,W)}{4},\, 0.5\right). \quad (4)$$

Subsequently, we normalize the Focus and Unfocus branches using $h_t = \widetilde{h}_t / \sum \widetilde{h}_t$ for Focus and $\bar{h}_t = (\mathbf{1} - \widetilde{h}_t)/\sum(\mathbf{1} - \widetilde{h}_t)$ for Unfocus. This yields the two pooled vectors on the patch features

$$\mathbf{g}_t \;=\; \sum_{i,j} (h_t)_{i,j}\,(F_t)_{i,j} \;\in\; \mathbb{R}^D, \qquad \mathbf{c}_t \;=\; \sum_{i,j} (\bar{h}_t)_{i,j}\,(F_t)_{i,j} \;\in\; \mathbb{R}^D. \quad (5)$$

Here, $\mathbf{g}_t$ aggregates gaze-focused evidence, whereas $\mathbf{c}_t$ emphasizes its complement. Thus, peripheral evidence remains available to the question gate through $\mathbf{c}_t$.

In addition, we compute a Global token $\mathbf{m}_t = \frac{1}{HW}\sum_{i,j}(F_t)_{i,j}$, which provides a gaze-independent scene summary. If gaze coordinates are absent or erratic for a frame, the LM retains an unbiased holistic representation without depending on noisy (x_t, y_t) values.

3.3 Question-Conditioned Semantic Gating

Although the gaze heatmap highlights the user's focus, some questions also require peripheral evidence. We therefore condition a semantic gate on the question to control the balance between these two sources.

We represent the corresponding question as $\mathbf{u} \in \mathbb{Z}^{L_u}$. The question $\mathbf{u}$ is tokenized and embedded using the VLM's text encoder E. Mean-pooling collapses variable-length questions into a fixed-size vector without additional parameters, and L2-normalization removes the scale difference between short and long questions so gate outputs are comparable across samples.

$$\bar{\mathbf{e}} \;=\; \frac{1}{L_u}\sum_{k=1}^{L_u} E(u_k), \qquad \mathbf{t} \;=\; \frac{\bar{\mathbf{e}}}{\|\bar{\mathbf{e}}\|_2 + \varepsilon} \;\in\; \mathbb{R}^D, \qquad \varepsilon = 10^{-6}. \quad (6)$$

We then pass the normalized vector $\mathbf{t}$ through a lightweight two-layer Multi-Layer Perceptron (MLP), $\phi_\theta : \mathbb{R}^D \to \mathbb{R}$, followed by a sigmoid to obtain the gating scalar

$$s \;=\; \sigma\big(\phi_\theta(\mathbf{t})\big) \;\in\; [0,1]. \quad (7)$$

A scalar directly controls the balance between focused and complementary evidence. Because it encodes a preference from the question, we share this coefficient across the clip's frames. We then form the primary contextual token by blending the two pooled representations.

$$\mathbf{p}_t \;=\; s\cdot\mathbf{g}_t \;+\; (1-s)\cdot\mathbf{c}_t. \quad (8)$$

When the gate $s \to 1$, the primary token concentrates on the gaze region. When $s \to 0$, it is drawn from the complementary visual evidence.

3.4 Token Assembly and Training Objective

To retain both question-specific and scene-level evidence, we pair the gated token $\mathbf{p}_t$ with the global token $\mathbf{m}_t$. The former carries the question-conditioned mixture of gaze and complementary features, while the latter provides a scene summary independent of gaze. We then map both tokens into the language-model representation using trainable linear projectors.

$$\mathbf{a}_t^{\text{ctx}} = W_{\text{ctx}}\,\mathbf{p}_t, \qquad \mathbf{a}_t^{\text{vis}} = W_{\text{vis}}\,\mathbf{m}_t, \qquad W_{\text{ctx}}, W_{\text{vis}} \in \mathbb{R}^{D\times D}. \tag{9}$$

The ordering for each frame is established as $[\mathbf{a}_t^{\text{ctx}}, \mathbf{a}_t^{\text{vis}}]$, acting as a visual prefix for the text. Thus, the total visual budget on the LM side is strictly limited to $2T$.

We frame the egocentric video QA task as a multiple-choice question answering (MCQA) problem. During training, we keep the backbone VLM base weights frozen and attach lightweight LoRA adapters to the attention layers, minimizing the trainable parameter footprint while preserving pre-trained reasoning capabilities. Gradients update only the LoRA adapters, the linear projectors (W_{ctx}, W_{vis}), and the semantic gate ϕ_θ.

The entire framework is optimized end-to-end using a Cross-Entropy (CE) loss over the probability distribution of the four answer options (A, B, C, D) with label smoothing.

$$\mathcal{L} = -\sum_{k\in\{A,B,C,D\}} \left((1-\alpha)y_k + \frac{\alpha}{4}\right) \log P(\text{token}_k \mid \mathcal{V}, \mathcal{G}, \mathbf{u}), \tag{10}$$

where $y_k \in \{0,1\}$ is the ground-truth binary label, α is the label smoothing factor, and $P(\text{token}_k \mid \mathcal{V}, \mathcal{G}, \mathbf{u})$ is the predicted probability generated by the LM at the final sequence position.

4 Experiments

We first compare StreamGaze accuracy and model-forward computation, then assess transfer to EgoGazeVQA. Finally, component interventions examine sensitivity to frame selection and spatial pooling. Appendix D.1 reports gaze-noise controls.

4.1 Implementation

Datasets & Metrics. We evaluate our method on StreamGaze (Lee et al., 2025), an egocentric video QA benchmark that we evaluate as three tasks, Past, Present, and Proactive, covering ten subtasks drawn from EGTEA Gaze+ (Li et al., 2021), EgoExo (Grauman et al., 2024), and HoloAssist (Wang et al., 2023). Following the official partition, we train on 5,766 QA pairs and report final results on the held-out test split of 1,399 questions, yielding 2,418 evaluation examples after expanding proactive questions into timestamped probes. We adopt macro accuracy and micro accuracy as the primary evaluation metrics. Each task score averages its constituent subtask accuracies, and Macro is the equal-weight mean of the Past, Present, and Proactive scores. Micro is accuracy pooled over all test examples. To quantify model computation, we report TFLOPs per input forward under the controlled protocol in Appendix E.2.

Baselines & Models. We compare our method against two groups of baselines. The first group consists of small VLMs evaluated without token reduction or gaze tuning, including SmolVLM2-256M and SmolVLM2-500M (Marafioti et al., 2025) (our 256M and 500M backbones), FastVLM (Vasu et al., 2025), and Mobile-VideoGPT (Shaker et al., 2025). These represent the accuracy achievable with full token grids and standard inference pipelines, none of which were evaluated in the original StreamGaze report. The second group consists of token-reduction methods re-implemented on the SmolVLM2-500M framework for a controlled comparison of visual token budget and accuracy on identical hardware, namely GazeLLM (Rekimoto, 2025), LLaMA-VID (Li et al., 2024), and Voila-A (Yan et al., 2024). We use SmolVLM2-500M as the shared backbone to control for model capacity. We select these three because they each reduce visual tokens through distinct mechanisms. GazeLLM crops each frame to a fixed patch centered on the raw gaze point, discarding all peripheral context. LLaMA-VID produces two tokens per frame by combining one global mean over the feature grid with one context token formed by text-attended averaging over the full feature grid without any gaze prior. Voila-A feeds a gaze heatmap into a Perceiver resampler and mixes the resampled output with intermediate visual features.

Table 2: StreamGaze accuracy and model cost, grouped by total parameter count. Params counts unique parameters, including vision encoders and added modules. TFLOPs are per input forward (Appendix E.2). Macro equally averages Past, Present, and Proactive, while Micro pools test examples. Both Ours variants report three-seed means. Across-seed variability is reported separately in Table 6. Bold/italic mark best/second-best accuracy. Adap. denotes dynamic selection, SVLM denotes SmolVLM2, and Mob-VGPT denotes Mobile-VideoGPT.

Method	Params (M) ↓	TFLOPs ↓	Fr.	Past ↑	Present ↑	Proactive ↑	Macro ↑	Micro ↑
Below 300M parameters								
SVLM-256M	256.48	3.966	16	0.328	0.283	0.537	0.383	0.480
Ours-256M	258.24	2.857	Adap.	0.351	*0.532*	0.643	*0.509*	*0.617*
300M+ parameters								
SVLM-500M	507.48	4.742	16	0.350	0.449	0.578	0.459	0.582
GazeLLM	508.60	3.791	16	0.309	0.233	0.636	0.393	0.502
LLaMA-VID	510.44	3.804	16	*0.378*	0.436	0.635	0.483	0.576
Ours-500M	511.37	3.055	Adap.	**0.423**	**0.551**	*0.645*	**0.539**	**0.640**
Voila-A	530.39	4.758	16	0.311	0.503	**0.650**	0.488	0.610
FastVLM	622.66	13.591	16	0.377	0.412	0.476	0.421	0.466
Mob-VGPT	656.36	3.903	16	0.350	0.370	0.507	0.409	0.492

Hyperparameters. Videos are first subsampled to 1 FPS before frame selection, normalizing the candidate pool across clips that vary widely in native frame rate. The dynamic frame selector then operates in the range $t_{min} = 4$ to $t_{max} = 16$ frames per clip, with per-video scores cached to avoid redundant computation across multiple questions from the same video. Our framework strictly limits the visual span to 2 tokens per frame. The semantic gate is a two-layer MLP with ReLU activation and sigmoid output. The gaze heatmap uses $\rho = \max(\max(H, W)/d,\ 0.5)$ with $d = 4$. We train the semantic gate, linear projectors (W_{ctx}, W_{vis}), and LoRA (Hu et al., 2021) adapters (rank $r = 8$, $\alpha = 16$, targeting q_{proj} and v_{proj}) using AdamW with cross-entropy loss and label smoothing ($\varepsilon = 0.1$). Training runs for 4 epochs on an NVIDIA A6000 GPU.

4.2 Results

Table 2 compares accuracy, total parameters, and counted input-forward computation. Across three seeds, our 256M configuration achieves 50.9% Macro and $61.66 \pm 0.50\%$ Micro. With the larger backbone, our 500M configuration achieves $53.94 \pm 1.54\%$ macro and $64.01 \pm 0.49\%$ micro accuracy (mean $\pm$ sample standard deviation). At the same time, Ours-500M uses 3.055 TFLOPs per input, compared with 3.791 for GazeLLM, 3.804 for LLaMA-VID, and 4.758 for Voila-A under the shared-backbone protocol.

The methods differ in how they allocate the compact visual budget. Soft Gaussian pooling retains peripheral evidence that GazeLLM's crop discards, while the gate controls its contribution to the primary token. LLaMA-VID also uses two tokens per frame, but obtains its context token through text-conditioned attention over the full grid. GazeLens adds the wearer's gaze as a spatial prior and preserves a global summary alongside it.

However, Voila-A achieves slightly higher Proactive accuracy than Ours-500M (0.650 vs. 0.645), with its advantage concentrated in gaze-triggered alerts, whereas ours performs better on object-appearance alerts. This advantage comes with higher counted input-forward computation of 4.758 versus 3.055 TFLOPs per input, about 56% more under the RTX 3080 profiling protocol (Figure 3). Voila-A processes 16 frames with a gaze-conditioned Perceiver, whereas ours averages 12.72 selected frames and uses two visual tokens per frame. The Perceiver may retain spatial cues useful for alert triggering. Temporal sampling and spatial representation both differ between these methods. Overall, ours retains higher Macro and Micro accuracy.

Beyond StreamGaze, Section 4.3 evaluates transfer to EgoGazeVQA. Appendices B, C, and D examine frame selection, spatial representations, and gaze robustness, respectively. Appendix E.2 reports the controlled input-forward FLOP protocol.

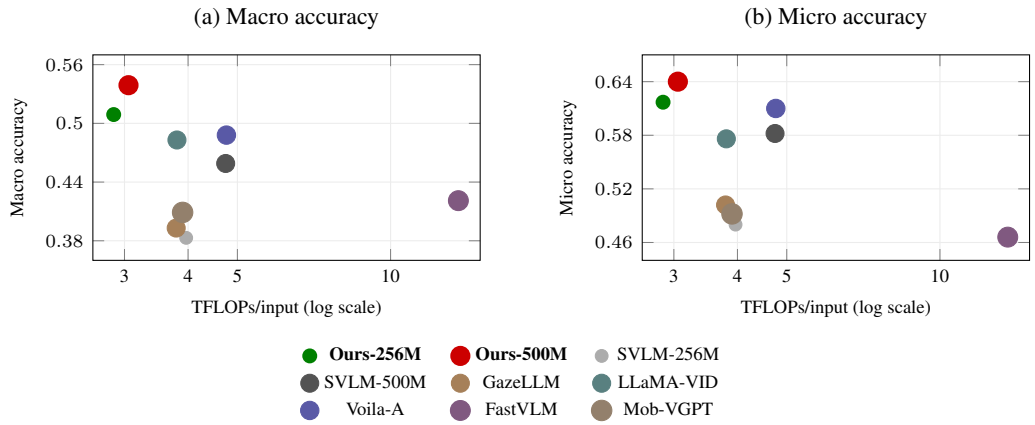


Figure 3: Accuracy versus input-forward computation at each model's native image resolution. Muted colors identify baselines, while green and red highlight Ours-256M and Ours-500M. Linear marker size scales with the square root of total parameters. Higher and further left is better. The horizontal axis is logarithmic. FLOPs exclude frame selection, preprocessing, and answer generation (Appendix E.2).

4.3 Cross-Dataset Generalization

To assess transfer beyond StreamGaze, we train for four epochs on the EGTEA source of EgoGazeVQA and evaluate on held-out Ego4D and EgoExo examples. Figure 4 reports accuracy pooled over evaluation examples. Uniform budgets of seven and eight frames provide comparisons at frame budgets close to the dynamic configuration.

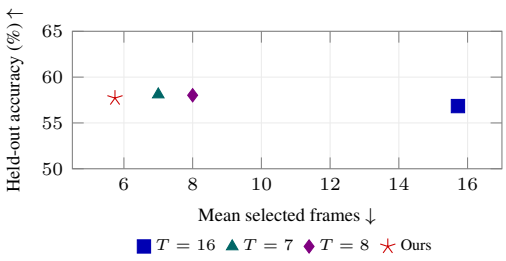


Figure 4: EgoGazeVQA transfer on 1,265 held-out examples. T denotes uniform sampling, whereas ours is dynamic. Higher and further left is better.

Dynamic selection uses the fewest frames and exceeds Uniform $T = 16$ in accuracy, while Uniform $T = 7$ and $T = 8$ achieve slightly higher accuracy. These results show that adaptive selection retains competitive accuracy with a smaller frame budget on a second gaze-VQA benchmark, supporting the transfer of its frame-efficiency benefit beyond StreamGaze. Appendix A.1 provides the source-specific breakdown.

4.4 Ablation Studies

Protocol. To measure sensitivity to frame selection and the mixing coefficient, we change one component at inference while keeping trained weights fixed. The full-method reference uses three Ours-500M checkpoints selected by validation micro accuracy. As in the main comparison, Macro equally averages Past, Present, and Proactive. Mean and sample SD summarize the full-method reference across checkpoints.

Frame selection. The picker selects 12.72 frames per test example on average (SD 3.86 across examples, median 14, range 4–16). Frame counts are identical across the three checkpoints because selection is independent of learned weights. To compare temporal policies at a fixed spatial-token budget, Table 3 replaces the HOG-and-sharpness selector with uniform sampling of up to 16 frames from the same 1-FPS candidate sequence. The dynamic configuration retains between 4 and 16 frames. Both settings use the same gaze pooling, learned gate, and two visual tokens per selected frame. Figure 5 illustrates the selected frames.

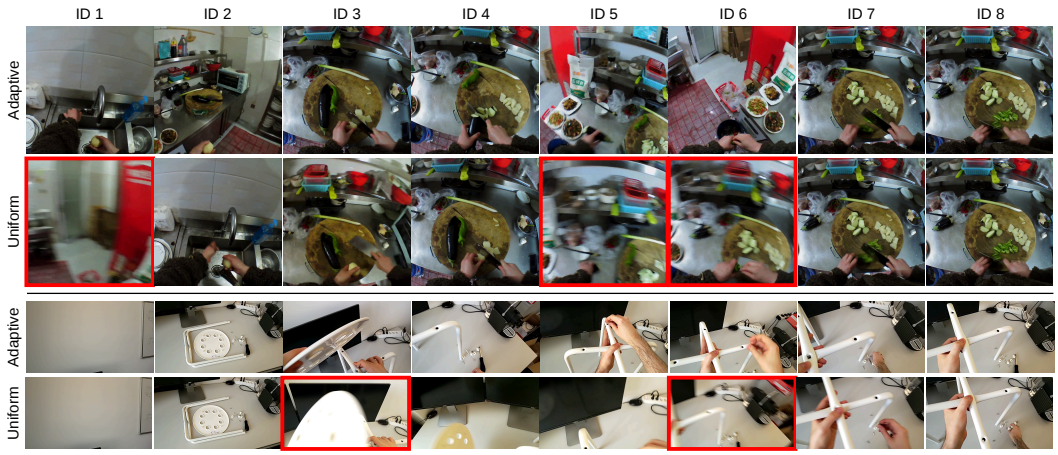


Figure 5: Adaptive versus uniform frame selection for example video sequences. Red boxes mark redundant or motion-blurred frames in the uniform selections. Frame IDs run from 1 to 8.

Table 3: Frame-selection ablation. Ours reports mean Macro across three checkpoints, and $\bar{T}$ is mean frames per evaluation example.

Selector	$\bar{T}$	Macro
Uniform $T = 16$	16.00	0.525
Dynamic	12.72	0.539

Table 4: Inference-time gate ablation. Fixed gates replace the learned mixing coefficient.

Gate	Macro
Gaze only ($s = 1$)	0.524
Equal blend ($s = 0.5$)	0.529
Complement only ($s = 0$)	0.520
Learned gate	0.539

Learned gate. To examine the role of the learned mixture, Table 4 overrides the learned mixing coefficient with $s = 1$, $s = 0.5$, or $s = 0$ during evaluation. These settings use gaze-focused pooling, an equal mixture, or complementary pooling in the primary token, respectively. The global token and selected frames remain unchanged. Appendix B gives additional component comparisons.

Gate behavior. To examine how the learned gate allocates attention, Figure 6 shows one evaluated Ours-500M checkpoint on the StreamGaze test set. Larger s favors gaze-focused pooling, while the global token remains present. Across examples, the mean is 0.6509 and the within-run standard deviation is 0.0253. Present has the highest median gaze weight. Each value averages visual-block gate outputs, and evaluation rows receive equal weight within each group.

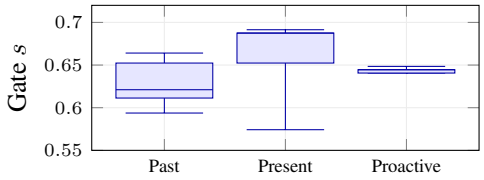


Figure 6: Gate distributions from one evaluated checkpoint. Boxes show quartiles and medians, while whiskers span min–max. Groups pool evaluation rows, including repeated question templates.

5 Conclusion

In this paper, we proposed GazeLens for compact egocentric video question answering. First, adaptive frame selection uses HOG transitions and image sharpness to reduce temporal redundancy before visual encoding. Next, gaze-focused and complementary pooling preserve evidence within and beyond fixation, while a question-conditioned gate adjusts their contribution. A global summary then retains scene context alongside the primary token, yielding two visual tokens per selected frame. Together, these modules address the need to preserve relevant evidence within a limited visual budget. Future work could reuse visual tokens across frames with little change to reduce repeated computation during longer interactions. In addition, adapting the spatial token budget to the question and scene complexity could preserve finer details when needed.

AI USE STATEMENT

In this work, we used generative AI tools (i.e. ChatGPT) for revising passages, and correcting grammar.

All research ideas, methods, experiments, analyses, and conclusions are the authors' own, and the authors reviewed and take full responsibility for all content of this paper.

ETHICS STATEMENT

This work studies egocentric video and gaze data from existing benchmarks. Such data can expose private activities, bystanders, and patterns of attention. Potential applications include assistive visual interfaces, but gaze-conditioned systems could also enable intrusive monitoring. Access to gaze coordinates does not itself establish consent. Any deployment should obtain informed consent from affected users, limit data collection and retention, protect access to video and gaze records, and respect the source datasets' usage conditions. The reported benchmark experiments do not establish safety for deployment or energy savings on wearable hardware.

REPRODUCIBILITY STATEMENT

Section 3 describes the method, and Section 4.1 specifies the data split, metrics, and training settings. Section 4.3 describes transfer evaluation. Appendices B–D specify the component comparisons and report across-seed variability for both GazeLens backbones. Appendix E documents parameter counts and the FLOP measurement protocol.

REFERENCES

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18407–18418, 2024a.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. *arXiv preprint arXiv:2403.06764*, 2024b.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Yi-Xing Peng, Kun-Yu Lin, Xihan Wei, Jian-Fang Hu, Xiaohua Xie, and Wei-Shi Zheng. Vispeak: Visual instruction feedback in streaming videos. pp. 21778–21788, 2025.

Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. Framefusion: Combining similarity and importance for video token reduction on large visual language models. *arXiv preprint arXiv:2501.01986*, 2024.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: low-rank adaptation of large language models. arxiv preprint. *arXiv preprint arXiv:2106.09685*, 2021.

Kai Hu et al. M-llm based video frame selection for efficient video understanding. *arXiv preprint arXiv:2502.19680*, 2025.

Bolin Lai, Fiona Ryan, Wenqi Jia, Miao Liu, and James M Rehg. Listen to look into the future: Audio-visual egocentric gaze anticipation. In *European Conference on Computer Vision*, pp. 192–210. Springer, 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.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision (ECCV)*, 2024.

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.

Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. *arXiv preprint arXiv:2411.03628*, 2024.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Zhenyu Ning, Guangda Liu, Qihao Jin, Wenchao Ding, Minyi Guo, and Jieru Zhao. Livevlm: Efficient online video understanding via streaming-oriented kv cache and retrieval. *arXiv preprint arXiv:2505.15269*, 2025.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24045–24055, 2025.

Jun Rekimoto. Gazellm: Multimodal llms incorporating human visual attention, 2025. URL https://arxiv.org/abs/2504.00221.

Abdelrahman Shaker, Muhammad Maaz, Chenhui Gou, Hamid Rezatofighi, Salman Khan, and Fahad Shahbaz Khan. Mobile-videogpt: Fast and accurate video understanding language model. *arxiv*, 2025. URL https://arxiv.org/abs/2503.21782.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokul Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, and Hadi Pouransari. Fastvlm: Efficient vision encoding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20270–20281, 2023.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

Kun Yan, Zeyu Wang, Lei Ji, Yuntao Wang, Nan Duan, and Shuai Ma. Voila-a: Aligning vision-language models with user's gaze attention, 2024.

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.

JinYi Yoon, Kazi Hasan Ibn Arif, and Bo Ji. GazeVLM: Gaze-guided vision-language models for efficient and robust inference. OpenReview preprint, 2025. URL https://openreview.net/forum?id=RAUMsywTko.

Sicheng Yu et al. Frame-voyager: Learning to query frames for video large language models. *arXiv preprint arXiv:2410.03226*, 2024.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams, 2024.

Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4372–4381, 2017.

A Additional Benchmark Results

A.1 Transfer Results by Source

This experiment tests whether the frame-selection benefit extends to a second benchmark, complementing the StreamGaze ablations. Following Section 4.3, we train on the EGTEA source of EgoGazeVQA for four epochs and evaluate on held-out Ego4D and EgoExo examples. Table 5 expands Figure 4 by source and compares uniform sampling with GazeLens's HOG-and-sharpness selector. The combined accuracy pools correct predictions across both sources.

Table 5: Frame-selection comparison on EgoGazeVQA. Each row reports a selector on the indicated evaluation set. *Examples* is the number of evaluation items, and *Frames* is the mean selected frames per item. Accuracy is in percent.

Evaluation set	Selector	Examples	Frames	Accuracy
Combined	Uniform T16	1265	15.72	56.84
Combined	Uniform T7	1265	7.00	58.10
Combined	Uniform T8	1265	8.00	58.02
Combined	GazeLens selector	1265	5.74	57.71
Ego4D	Uniform T16	577	15.90	51.30
Ego4D	Uniform T7	577	7.00	51.99
Ego4D	Uniform T8	577	8.00	51.82
Ego4D	GazeLens selector	577	5.83	52.17
EgoExo	Uniform T16	688	15.58	61.48
EgoExo	Uniform T7	688	7.00	63.23
EgoExo	Uniform T8	688	8.00	63.23
EgoExo	GazeLens selector	688	5.67	62.35

GazeLens's selector exceeds Uniform T16 on both sources while retaining fewer than six frames per evaluation example on average. It also achieves the highest accuracy among the tested selectors on Ego4D, whereas T7 and T8 remain more accurate on EgoExo. Thus, the smaller frame budget transfers across both sources, while the accuracy trade-off depends on the source.

B Additional Ablations

We report across-seed variability for both backbones, then examine frame-selection cues and keyframe capacity on SmolVLM2-500M. Macro and Micro follow the definitions in Section 4.1. Full-method means average three checkpoints.

B.1 Across-Seed Variability

To examine how training variability is distributed across tasks, Table 6 reports task scores and aggregate accuracy across three seeds for each backbone. All checkpoints are evaluated on the same held-out examples.

Table 6: GazeLens accuracy across three training seeds per backbone on the same held-out examples. Smaller values following the means denote sample SDs across checkpoints. These means supply the main comparison.

Backbone	Past ↑	Present ↑	Proactive ↑	Macro ↑	Micro ↑
256M	0.351 $\pm$0.053	0.532 $\pm$0.010	0.643 $\pm$0.006	0.509 $\pm$0.017	0.617 $\pm$0.005
500M	0.423 $\pm$0.041	0.551 $\pm$0.011	0.645 $\pm$0.011	0.539 $\pm$0.015	0.640 $\pm$0.005

Past has the largest across-seed variation for both backbones, whereas Micro has an SD of 0.005 in each case. Moving to the 500M backbone improves Past more than Present or Proactive. Consequently, its higher Macro primarily reflects improved Past accuracy.

B.2 SELECTION CUES

HOG emphasizes changes between frames, while sharpness favors visually clear frames. Following the uniform-versus-dynamic comparison in Table 3, Table 7 examines whether combining these cues improves the accuracy and frame-count trade-off over using either cue alone.

Table 7: HOG vs. sharpness inside the dynamic picker. $\bar{T}$ is mean frames per evaluation example.

Picker configuration	Macro	Mean $\bar{T}$
HOG only	0.501	11.31
Sharpness only	0.508	13.23
HOG + sharpness	0.539 $\pm$0.015	12.72

The combined picker achieves the highest Macro even though it selects fewer frames than sharpness alone. This supports using the cues together to balance changes in scene content with image clarity within the frame budget.

B.3 KEYFRAME CAPACITY

To examine how much temporal capacity the selector needs, Table 8 varies t_{max} from 4 to 16 with $t_{min} = 4$. The $t_{max} = 16$ column reports the full-method configuration.

Table 8: Impact of dynamic keyframe capacity (t_{max}) on StreamGaze macro and micro accuracy. A fixed $t_{min} = 4$ is maintained.

t_{max} cap	4	8	12	16
Macro	0.505	0.526	0.536	0.539 $\pm$0.015
Micro	0.612	0.629	0.637	0.640 $\pm$0.005

Both metrics improve as the cap increases, with the largest gains occurring between four and eight frames. Raising the cap from 12 to 16 adds 0.003 to both Macro and Micro. Thus, additional temporal capacity helps most at the smallest budgets, with diminishing accuracy gains as the cap grows. The full configuration allows up to 16 frames while selecting 12.72 per evaluation example on average.

C SPATIAL REPRESENTATION

These experiments compare pooling rules, token composition, and spatial token budgets on SmolVLM2-500M. Frame selection uses HOG and sharpness as described in Section 3.1. The full-method reference reports mean $\pm$ sample SD across three checkpoints. The comparisons below distinguish changes made at evaluation time from variants trained separately under the four-epoch protocol.

In these controls, p_t denotes the primary context token and m_t the global mean token, corresponding to $\mathbf{a}_t^{\text{ctx}}$ and $\mathbf{a}_t^{\text{vis}}$ in the method description.

C.1 GAZE-WEIGHTED POOLING

While Table 4 examines how the focused and complementary pools are mixed, Table 9 tests whether the focused pool itself benefits from gaze weighting. It replaces gaze-weighted pooling in the primary branch with the spatial mean while retaining complementary pooling and the global token.

Table 9: Gaze-weighted pooling vs. average pooling. *Mean pool* sets $\mathbf{g}_t$ to the spatial mean of F_t before gating with $\mathbf{c}_t$.

Primary branch pooling	Macro
Gaze-weighted pool	0.539 $\pm$0.015
Spatial mean pool	0.520

Replacing gaze weighting with a spatial mean lowers Macro by 0.019. This comparison supports preserving a gaze-focused summary before the gate blends it with complementary context, even when a separate global token is available.

C.2 REPRESENTATION AND SPATIAL BUDGET

To assess which information deserves a limited token budget, Table 10 compares token representations and spatial budgets. Primary-only, global-only, and duplicated-primary rows modify the reference checkpoint at evaluation time, whereas the focus-plus-complement and spatial-grid variants use separate training runs. Each row specifies the tokens supplied to the language model per selected frame.

Table 10: StreamGaze representation and spatial-budget controls on the full test set. Tokens are per frame, and accuracies are percentages. Ours reports three-seed mean $\pm$ SD.

Representation	Tokens/frame	Macro	Micro
Primary $[p_t]$	1	52.0	62.5
Global $[m_t]$	1	49.5	60.5
Focus + complement $[g_t, c_t]$	2	52.5	63.0
Primary pair $[p_t, p_t]$	2	52.1	62.6
Ours $[p_t, m_t]$	2	53.94 $\pm$1.54	64.01 $\pm$0.49
2×2 grid	4	54.0	64.2
4×4 grid	16	54.3	64.5

Pairing the primary and global tokens gives the highest Macro and Micro among the one- and two-token variants, while duplicating the primary token brings little gain. The 4×4 grid improves Macro by 0.36 percentage points and Micro by 0.49 percentage points over our pair, but uses eight times as many tokens per frame. These comparisons support allocating the two-token budget to a question-conditioned primary summary and a global scene summary.

D GAZE ROBUSTNESS

To assess the gaze input, we evaluate sensitivity to coordinate noise and quantify fixation coverage. Macro and Micro use the same definitions as the main comparison. The zero-noise reference reports mean $\pm$ sample SD across three checkpoints.

D.1 ROBUSTNESS TO GAZE NOISE

To test sensitivity to imprecise gaze, Table 11 varies Gaussian noise in normalized gaze coordinates at inference time. The zero-noise column reports the full-method result.

Table 11: Model performance under synthetic spatial gaze noise at test time.

Noise σ	0.00	0.01	0.02	0.05	0.10	0.20
Macro	$0.539_{\pm 0.015}$	0.537	0.534	0.526	0.514	0.498
Micro	$0.640_{\pm 0.005}$	0.638	0.635	0.627	0.615	0.599

At $\sigma = 0.02$, both Macro and Micro are 0.005 below the zero-noise reference. At the largest tested noise level of $\sigma = 0.20$, both decrease by 0.041. Accuracy therefore changes little under small perturbations and declines as the noise grows, indicating that the compact representation tolerates modest coordinate errors but remains sensitive to gaze quality.

D.2 FIXATION COVERAGE

Fixation coverage is measured on a separate cached selection trace containing 17,950 selected-frame uses. Of these, 82.84% fall inside detected fixation intervals, 7.28% have valid coordinates outside those intervals, and 9.88% lack valid coordinates. Repeated uses of a frame across questions or probes are counted separately. These percentages describe frame coverage by fixation status and coordinate availability.

E IMPLEMENTATION AND COMPUTATION

E.1 TRAINABLE PARAMETERS

Our method trains 1.75M parameters on the 256M backbone and 3.88M on the 500M backbone, including LoRA. Gaussian pooling and the frozen question embedding introduce no new parameters. Linear layers use biases, and LoRA uses rank $r = 8$ and $\alpha = 16$ on q_{proj} and v_{proj} in the vision encoder and language model.

E.2 CONTROLLED INFERENCE COST

We count input-forward computation at batch size one on an NVIDIA RTX 3080, using synthetic frames and a fixed question-and-options prompt. The native processed frame sizes are 512×512 for SmolVLM2, 1024×1024 for FastVLM, and 224×224 for both Mobile-VideoGPT vision encoders. SmolVLM2 and FastVLM use bfloat16, whereas Mobile-VideoGPT uses float16. Full-grid backbones are pretrained, and adapted SmolVLM2 methods add randomly initialized rank-eight LoRA adapters and method modules. Parameter values do not change the counted tensor shapes.

FLOPs cover supported operators in the vision encoders, token modules, and language model, including all prompt-position logits. A multiply-add counts as two operations, and grouped-query attention uses dense-equivalent attention arithmetic, including the full causal matrix convention. These are arithmetic counts, not hardware instruction counts. Video decoding, external frame selection, preprocessing, model loading, and autoregressive answer generation are excluded. Normalization and nonlinear operations are not included in the operator-counting convention.

Baselines use 16 frames. For both adaptive backbones, we count each frame length from 4 to 16 and weight it by the 2,418-example test-set selector distribution, averaging 12.72 frames per example.

Per-example frame counts are identical across all three runs of both the 256M and 500M variants. The adaptive costs combine temporal selection and spatial compression.

Mobile-VideoGPT additionally uses an opaque selective-scan kernel. From the observed tensor dimensions, we estimate its arithmetic as $BLD(7N + 2)$ per directional scan, where B is batch size, L sequence length, D inner width, and N state size. This counts state updates, output contraction, and the skip connection, excluding exponentials and softplus. Its 64 observed scans add 0.013187 TFLOPs to 3.889336 operator-counted TFLOPs. The total combines operator counts with this analytical scan estimate.

F Limitations and Future Work

GazeLens targets devices with gaze sensors. When coordinates are missing, pooling becomes approximately uniform and loses the spatial prior that distinguishes attended content. A useful next step is to estimate gaze confidence and adapt the pooling rule when tracking is unreliable, so that uncertain fixation measurements do not determine the spatial representation.

Temporal selection uses HOG transitions and image sharpness without conditioning on the question. Consequently, these cues can suppress visually repetitive frames that still contain relevant evidence. Future work could incorporate question relevance into selection while keeping its cost below that of encoding discarded frames. Selected frames are currently encoded independently, so reusing visual tokens across unchanged regions is another route to reducing repeated computation during longer interactions. Such extensions should be evaluated with end-to-end latency and memory measurements, since the reported model-forward FLOPs exclude frame selection, decoding, and preprocessing.

Finally, two tokens per frame provide a compact input but can discard spatial detail needed for fine-grained reasoning. The spatial-budget comparisons motivate adapting token capacity to the question and scene complexity. Increasing capacity only when the task needs finer evidence could extend the present accuracy–computation trade-off, while retaining a compact representation for simpler inputs.