GAZESIEVE: GAZE-GUIDED KEYFRAME SELECTION FOR EGOCENTRIC VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Egocentric video question answering often depends on a few evidence moments: an attended object, a brief state change, or an earlier observation that explains the query. The bottleneck is therefore selecting the few frames that contain sufficient evidence under a small visual budget, yet downstream answer accuracy alone cannot reveal whether selection succeeded. Recent video-language agents can perform query-conditioned search, but remain gaze-free and may miss moments humans naturally attend to. We present **GazeKeyBench**, a direct keyframe-selection benchmark that annotates acceptable evidence windows, required evidence groups, and alternative sufficient evidence sets, judging whether selected frames *cover* the evidence rather than land near one timestamp. We also present GazeSieve, a training-free gaze-guided selector that turns noisy gaze into fixation proposals, extracts query-relevant concepts, searches high-resolution gaze-centered crops under a gaze-informed temporal prior, and returns a relation-aware evidence set to a frozen MLLM. Under a *unified controlled protocol* that freezes the answerer, decoding, prompt, rendering, and frame budget, replacing uniform sampling with GazeSieve raises StreamGaze Macro-8 from 58.0 to **60.8** ($+2.8$, $p < 0.01$) on a frozen 4,700-unit roster. On GazeKeyBench's 2,607 human-annotated cases from 99 videos, GazeSieve reaches **56.0%** Evidence Recall@4, outperforming a strong gaze-free Agent(VLM) baseline at 52.0%. A paired timing control further shows that preserving measured gaze timestamps adds 1.9 points (p=0.0247). These results suggest that measured gaze provides complementary evidence-selection signal for egocentric video QA beyond query-conditioned video search alone.

1 INTRODUCTION

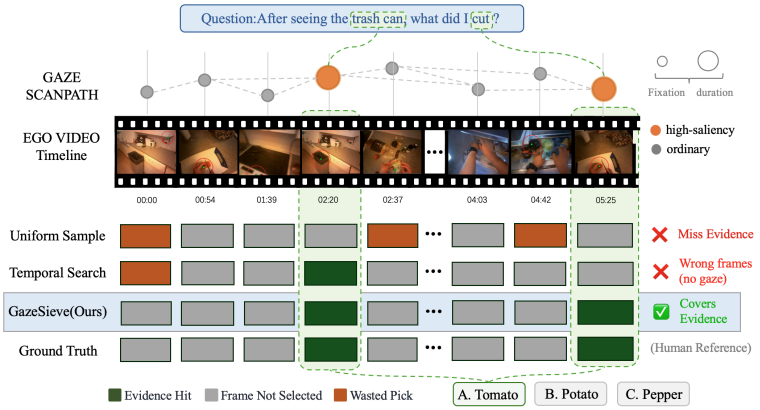


Figure 1: **From attended moments to temporal evidence.** Sparse sampling can miss the brief observations a query needs. GazeSieve turns gaze into fixation proposals, searches foveal crops around attended regions, and returns a compact relation-aware frame set to a frozen MLLM.

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.

Egocentric video question answering is becoming an important interface for wearable AR assistants, robots, and embodied agents. Gaze-QA benchmarks such as EgoGazeVQA Peng et al. (2025) and StreamGaze Lee et al. (2025) pose questions grounded in the wearer's attention: what is happening now, what happened earlier, and what to surface next. On these devices gaze comes almost for free, since eye tracking is standard on AR glasses.

Across these benchmarks, which frames the answerer sees is critical. A multimodal LLM can ingest only a limited number of frames per call, whereas an egocentric video spans hundreds to thousands. Under a fixed answerer, prompt, and frame budget, better selections raise answer quality and poor selections hide necessary evidence. We formalize choosing those frames as *pre-answer evidence selection*: given an egocentric video, a synchronized gaze trace, and a query, a selector must return, under a fixed frame budget, a set of frames sufficient to answer the query. Gaze is well suited to this role: it is a temporal index into the video, pointing to where an answer's evidence likely lies and narrowing the region a selector must search before any frame is decoded.

Yet almost no selector is designed for gaze, and this nearly free signal stays largely unused. When used at all, gaze is overlaid as a cue on frames chosen by other means, shaping the answer while playing no part in which frames are chosen. Existing frame selectors—fixed-rate sampling, CLIP-similarity retrieval, shot segmentation, and coarse-to-fine temporal search Wu et al. (2019); Yu et al. (2023); Wang et al. (2024; 2025); Ye et al. (2025)—transfer imperfectly to egocentric QA: they neither exploit gaze to localize attended moments nor cope with the small, frequently occluded, rapidly shifting targets typical of first-person video. In our direct evaluation these gaze-free selectors reach only 36.1% (colour/HSV similarity), 31.4% (Shot), and 49.0% (VLM temporal search) evidence recall, far below what a gaze-aware design attains. Even a stronger gaze-free Agent(VLM), which performs query-conditioned search over video, lacks access to measured human attention. The central question is therefore not whether gaze beats weak heuristics, but whether measured gaze adds evidence-selection signal beyond capable query-conditioned video search.

A second obstacle is evaluation. Frame selection is often judged through downstream answer accuracy, which hides selector-specific failures: a model may answer correctly from imperfect frames, fail despite seeing the evidence, or waste budget on the same moment. We therefore propose GazeKeyBench, which annotates each question as *minimal sufficient evidence sets*: each set contains required groups, frames within a group are equivalent, and different sets are alternative routes. Scoring asks whether selected frames cover any sufficient route using Evidence Recall@K, All-Group Coverage@K, and Precision@K, separating evidence sufficiency from timestamp exactness.

Building on this, we propose GazeSieve (Figure 1), a gaze-adapted framework that turns noisy gaze directly into an evidence set with no task-specific parameter updates. Gaze drives selection in three steps: it groups gaze into fixation proposals and extracts query-relevant concepts; runs a gaze-conditioned foveal temporal search that scans over time within high-resolution crops around the gaze point, where an open-vocabulary detector grounds these crops against the query concepts; and finally composes the verified moments into a compact, ordered set via query relations. The roles stay clean—gaze proposes, grounding verifies, query relations compose—so GazeSieve selects at the level of *sets*, keeping complementary moments and suppressing redundant ones to match the minimal-sufficient evidence structure GazeKeyBench annotates.

We evaluate at both the keyframe and full-system levels under a *unified controlled protocol* that freezes the answerer, decoding, prompt, rendering, and frame budget, so the only variable is the selected frames and each gain is attributable to selection. Published end-to-end StreamGaze systems run under different inputs, budgets, and decoding, so we report them only as external context. In direct evidence evaluation, we compare against a strong gaze-free Agent(VLM) baseline that performs query-conditioned video search but has no access to gaze, isolating what measured gaze contributes beyond agentic search alone.

Contributions.

- **A direct benchmark for gaze-guided evidence selection.** GazeKeyBench annotates minimal sufficient evidence sets—equivalent windows, required groups, and alternative routes—and evaluates selected keyframes against this structured evidence, rather than against a single timestamp or downstream answer accuracy alone.

- **A training-free selector in which gaze drives selection.** GazeSieve promotes gaze from a cue over already chosen frames to the signal that proposes evidence moments, combining gaze-conditioned foveal temporal search, query-conditioned grounding, and relation-aware finalization.
- **Controlled validation against strong baselines.** Under a frozen-answerer protocol, GazeSieve improves StreamGaze accuracy and GazeKeyBench evidence recall over gaze-free and naive-gaze selectors, including a strong Agent(VLM) baseline; ablations and a paired timing control isolate the roles of grounding, relation composition, and measured gaze timing.

2 Related Work

Video keyframe selection. Methods that pick a few frames from a long video fall into two families: those scoring frames by learned importance and those conditioning on the query to localize informative moments. In the first, AdaFrame learns sequential selection for efficient recognition Wu et al. (2019); in the second, SeViLA chains language-aware temporal localization with video QA Yu et al. (2023), and recent selectors use multimodal encoders or lightweight MLLMs for the same scoring Hu et al. (2025). Both are judged only *indirectly* through downstream QA score, so a selector's own failures never become visible. In contrast, we score selection directly, testing whether the chosen frames contain the required temporal evidence.

Coarse-to-fine and agentic video search. Another line lets a model actively search a larger temporal workspace for evidence. VideoAgent iteratively retrieves with an LLM agent Wang et al. (2024); VideoTree builds a query-adaptive hierarchy Wang et al. (2025); T* reformulates temporal search as spatial search over frame grids Ye et al. (2025); GeLM localizes scattered intervals for multi-hop egocentric QA Chen et al. (2024). These systems are strong gaze-free baselines: they can search video conditioned on the query, but do not use measured gaze to narrow the temporal search space. Our setting is complementary, returning only a few frames and measuring directly whether they *cover* the evidence.

Gaze-guided video understanding. Measured gaze has long served recognition and prediction in first-person video rather than frame selection: localizing manipulated objects, recognizing actions, and forecasting human–object interaction Fathi et al. (2012); Li et al. (2015); Liu et al. (2020); task-dependent transitions describe how attention moves between objects Huang et al. (2018), and uncertainty-aware integration treats gaze as informative but fallible Min & Corso (2021); more recently, GazeVQA uses gaze to study collaborative intent Ilaslan et al. (2023), and StreamGaze extends gaze-guided reasoning to past, present, and proactive tasks Lee et al. (2025). These works overlay gaze on a fixed input to shape the answer rather than to select frames. In contrast, we use gaze as a *pre-answer temporal evidence index*: the output is a set of timestamps whose evidence quality is evaluated directly.

3 Method

3.1 Problem Formulation

We formulate gaze-guided keyframe selection as *pre-answering evidence search*. The selector has one job: under a small visual budget, pick the few frames that let the answerer answer the question.

Let I_q be the legal video interval for query q, containing frames $V_{I_q} = \{f_1, \ldots, f_L\}$ and a synchronized gaze stream sampled at the same L frame times,

$$G_{I_q} = \{(t_j, x_j, y_j, b_j)\}_{j=1}^{L},$$

where (x_j, y_j) is the normalized gaze coordinate at frame f_j and b_j is the gaze-state label when available. Given candidates C and an output budget B_o, the selector returns an ordered keyframe set $\mathcal{K} = \mathcal{S}(V_{I_q}, G_{I_q}, q, C; B_o)$ with $|\mathcal{K}| \leq B_o$, and a frozen MLLM predicts $\hat{y} = M(q, C, \mathcal{K})$. The selector sees the question and candidates; the gold answer, downstream correctness, and post-hoc feedback are all closed to it.

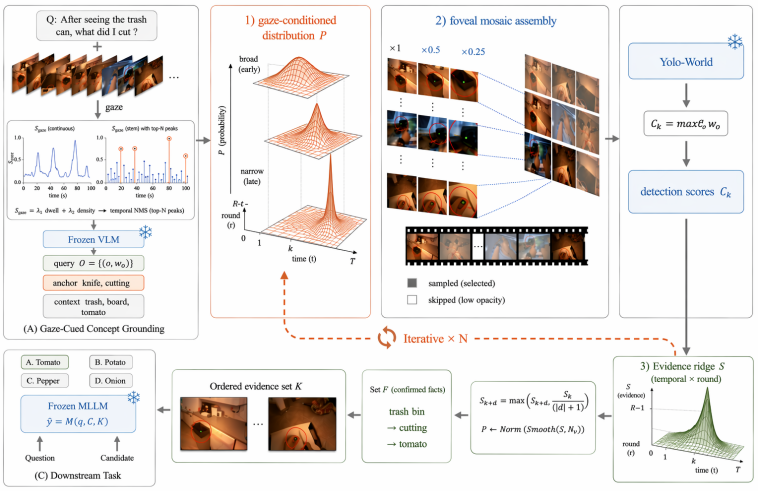


Figure 2: **GazeSieve pipeline.** Phase 1 turns gaze into fixation proposals and a query-relevant concept vocabulary. Phase 2 replaces uniform sampling with a gaze prior, searches high-resolution foveal crops, grounds the concepts, and steers sampling toward confirmed evidence; a relation-aware finalizer returns an ordered set.

The central difficulty is that gaze is ambiguous on its own: a fixation marks only where attention lands, which may be a relevant object, a distractor, a transition, or hesitation. GazeSieve therefore treats gaze as a temporal and spatial *proposal*; the query decides which proposals become evidence, and the output is an evidence *set* whose frames answer the query together.

3.2 Overview: Two-Phase Gaze-to-Evidence Search

GazeSieve searches in two phases. Phase 1 answers *what visual concepts should be searched for?* It compresses raw gaze into reliable fixation hotspots, selects a small set of temporally dispersed peaks, and uses a VLM to parse query-relevant object and scene clues from gaze-marked frames, producing a search vocabulary. Phase 2 answers *where in the legal interval is the evidence?* Two choices set it apart: where a standard searcher starts from a uniform temporal prior, we initialize from gaze saliency, pushing sampling toward attended moments; where it compresses whole frames into a low-resolution grid, we crop high-resolution foveal neighborhoods around gaze, spending pixels on small targets. An open-vocabulary detector then grounds the Phase 1 concepts; after each round the temporal distribution contracts toward confirmed evidence while retaining exploration over unvisited fixations. A relation-aware finalizer composes candidates into the evidence form the question requires.

3.3 Phase 1: Fixation Proposal and Query Clue Extraction

Fixation filtering. Raw gaze contains saccades, missing samples, jitter, and repeated observations. During saccades and other unreliable samples the gaze position drifts fast, making a poor evidence anchor. We retain fixation candidates $V_{\text{fix}} \subseteq V_{I_q}$ and discard unreliable samples from the proposal pool; when explicit gaze-state labels are absent, we approximate the same effect by grouping low-displacement samples and suppressing high-velocity transitions. A discarded sample is only one whose *position* is unreliable as an anchor; its pixels may still be useful.

Gaze saliency scoring. For each candidate $f_i \in V_{\text{fix}}$ we combine local dwell duration $\tilde{d}_i$ (how long attention remains in a temporal neighborhood) with spatial concentration $\rho(x_i, y_i)$ from a 2-D Gaussian kernel density over gaze positions:

$$S_{\text{gaze}}(f_i) = \lambda_1 \tilde{d}_i + \lambda_2 \rho(x_i, y_i). \tag{1}$$

The score only proposes candidates; whether one counts as evidence is decided later by grounding.

Temporal NMS. A long fixation creates many adjacent high-scoring frames; keeping all of them wastes budget. We apply temporal non-maximum suppression over S_{gaze} and keep the top N dispersed peaks $V_N^{\text{gaze}} = \text{TEMPORALNMS}(V_{\text{fix}}, S_{\text{gaze}}, N)$, giving a compact set of representative attended moments with coverage across the interval.

VLM-assisted concept extraction. The peak frames are rendered with a lightweight gaze marker (Gaussian heatmap or crosshair at the measured point). With the question and symmetrically ordered candidates, they are passed to a VLM instructed to extract visible object categories and scene cues relevant to the question, describing what is there and leaving the answer aside. The output is a weighted concept list $\mathcal{O} = \{(o, w_o)\}$, where core targets receive higher weight and contextual cues lower weight. Once fixed, this list stays fixed; later correctness signals cannot revise it.

3.4 PHASE 2: GAZE-CONDITIONED FOVEAL TEMPORAL SEARCH

Gaze prior initialization. Uniform temporal priors waste budget in first-person video, where evidence clusters near attended moments. GazeSieve instead pushes the prior toward gaze: $P_0(f_i) \propto S_{\text{gaze}}(f_i)$ for $f_i \in V_{\text{fix}}$ and $P_0(f_i) \propto \epsilon$ otherwise, a floor ϵ keeping a path into unattended regions. We also initialize a global score $S \leftarrow \mathbf{0}^L$, unvisited mask $N_v \leftarrow \mathbf{1}^L$, confirmed set $F \leftarrow \emptyset$, and relation requirements R_q from the query type.

Foveal crop mosaic. Whole-frame mosaics are compact but destroy evidence in egocentric video: small attended objects (tools, buttons, medicine, manipulated parts) may occupy a tiny fraction of a frame and vanish under grid compression. Each iteration we sample g^2 indices from $P \odot N_v$; for a sampled frame f_k with gaze (x_k, y_k) we map to raw coordinates $(X_k, Y_k) = (x_k W_{\text{raw}},\, y_k H_{\text{raw}})$ and extract a fixed-size crop

$$\text{Crop}_k = f_k\big[Y_k - \tfrac{h}{2} : Y_k + \tfrac{h}{2},\; X_k - \tfrac{w}{2} : X_k + \tfrac{w}{2}\big]\,, \tag{2}$$

with mirror padding at boundaries and a center-crop fallback when gaze is missing. The g^2 crops are stitched into a mosaic G_{gaze} that spends pixels on attended neighborhoods, preserving resolution for small or occluded evidence.

Open-vocabulary grounding. G_{gaze} and $\mathcal{O}$ pass to an open-vocabulary detector. For the cell of frame f_k we compute a weighted grounding score $C_k = \max_{o \in \mathcal{D}_k} c_o w_o$, with detector confidence c_o and Phase 1 weight w_o. Once a detection passes threshold and matches an unsatisfied requirement, the original frame joins F and R_q updates, keeping every grounding tied to the evidence contract.

Distribution update. Sampled frames are marked visited and their grounding scores written to S. Since egocentric evidence is temporally coherent (fixate, reach, manipulate, observe), a high score should raise nearby priority, so we diffuse scores locally, $S_{k+\delta} \leftarrow \max(S_{k+\delta},\, S_k/(|\delta| + 1))$ for $\delta \in [-w_t, w_t]$, and renormalize a smoothed interpolation over visited scores, $P \leftarrow \text{NORMALIZE}(\text{SMOOTH}(S, N_v))$. This concentrates later samples on coherent evidence while keeping mass over unvisited fixations, moving the search from proposals to confirmed evidence.

3.5 RELATION-AWARE EVIDENCE FINALIZATION

The search yields scored candidates and grounded concepts, but the answer often needs a *structure* of several frames together. The finalizer keys on the evidence structure the query requires, set by the relations in the query itself: a local state near query time, an attended anchor followed by a successor, an ordered chain of fixations, coverage across candidates, contrast between an attended moment and prior context, or a checkpoint-centered observation. Within each structure it selects the highest-scoring non-redundant frames satisfying the required groups, deduplicates timestamps, respects B_o, restores chronological order, and returns $\mathcal{K}$. Unlike plain top-K, which picks by single-frame score and tends to draw in neighboring frames of the same moment, GazeSieve picks by evidence relation and returns a compact set that covers the required groups and stays complementary rather than repetitive.

Evidence boundaries. All selection stays inside the legal interval I_q: past questions use history up to t_q, present questions use the prescribed query-time window, and proactive tasks are evaluated

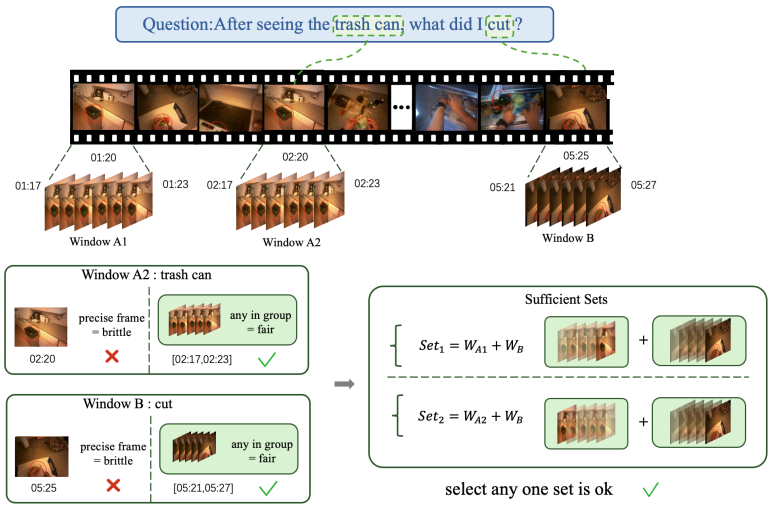


Figure 3: **GazeKeyBench annotation and scoring.** Each case has alternative sufficient sets, required groups, and equivalent windows; scoring tests whether the selected frames cover the evidence *structure*, not proximity to one timestamp.

at public checkpoints. The interval is fixed before selection and preserved in provenance, so the selector cannot look beyond it for evidence.

4 GazeKeyBench

GazeKeyBench measures one thing: whether the few frames a selector picks under a small budget suffice to answer the question. This might seem answerable with a single *gold keyframe*, but two common situations defeat a single-frame label: a question may have several equivalent evidence routes that each suffice, and it may require several moments together.

We therefore annotate each question's evidence as a three-level structure (Figure 3). At the bottom is a *window*, a closed interval within which any frame is interchangeable. In the middle is a *required group*, one piece of evidence that must be obtained, satisfied by any one of its equivalent windows. At the top is a *sufficient set*, a complete collection of required groups that together answer the question; a question can have several sufficient sets, i.e. alternative routes. Scoring asks only whether the selected frames cover every group of some sufficient set.

Formally, for case i the annotation $\mathcal{E}_i$ is a set of sufficient sets $\mathcal{S}_{is}$; each sufficient set contains required groups $\mathcal{G}_{isj}$; each group is a union of equivalent intervals $[l, u]$:

$$\mathcal{S}_{is} = \{\mathcal{G}_{is1}, \ldots, \mathcal{G}_{isJ_s}\}, \quad \mathcal{G}_{isj} = \{[l_{isjr}, u_{isjr}]\}. \tag{3}$$

Windows in one group are equivalent observations, distinct groups are complementary evidence, and distinct sets are alternative routes. Each window stores one representative keyframe as an audit point.

Human-annotated evidence and release. GazeKeyBench contains 2,607 human-annotated cases from 99 source videos spanning all ten StreamGaze tasks, with FAP annotated as a temporally ordered fixation sequence rather than a single detection event. Annotators mark acceptable windows, required groups, and sufficient sets while gold answers and selector outputs are hidden during evidence discovery. Each window stores a representative keyframe as an audit point, and each case retains its source-video identity for splitting and uncertainty estimation. We will release the annotation schema, evaluation protocol, scoring code, and reproducible metadata subject to the original data-access terms.

Matching and metrics. We give the selector a small budget of at most four frames, matching the four-option scale of each StreamGaze question, where each option draws on at most one frame of evidence. Scoring asks only whether these frames cover some sufficient set: each frame claims at

most one required group and each group is claimed by at most one frame (an edge exists when a frame falls inside any equivalent window of that group). Covering all groups of some sufficient set counts as a hit ($\mathrm{AGC@4} = 1$, else 0). How many groups a question needs is set by its own evidence structure and is often fewer than four, so covering them with three or two frames is equally correct; whether frames are wasted is left to precision.

Concretely, predictions are rank-sorted, interval outputs converted to midpoints, duplicates removed, and at most four unique timestamps retained. With J_s required groups in set s and matching cardinality M_{is}, $\mathrm{AGC@4}(i) = \mathbb{I}[\exists s : M_{is} = J_s]$. For a finer score, we take the most-covered canonical set $s_i^\star$, chosen by largest $r_{is} = M_{is}/J_s$, then largest M_{is}, then fewest groups, then smallest index, and on it define Evidence Recall $\mathrm{Recall@4}(i) = M_{is_i^\star}/J_{s_i^\star}$ and Selection Precision $\mathrm{Precision@4}(i) = M_{is_i^\star}/R_i$ if $R_i > 0$ (else 0), where $R_i \leq 4$ is the number of retained predictions. Covering the needed groups is thus independent of frame count, while extra useless frames lower precision. Because the 2,607 cases come from 99 videos, case-level observations are not independent; all fitting and uncertainty analyses group by source video.

5 Experiments

Setup. We evaluate full-system accuracy on StreamGaze Lee et al. (2025) under a unified controlled protocol. Controlled comparisons use a frozen 4,700-unit non-alert roster spanning eight past/present tasks and report the unweighted task mean (Macro-8) with matched per-unit frame budgets; uncertainty uses a source-video-clustered bootstrap (172 clusters) with task-stratified sign-flip tests. The published external-reference systems (Appendix E) instead use the full ten-task population (at most 2,000 instances per task, 7,391 units) and are not directly comparable. Direct keyframe evaluation uses GazeKeyBench's 2,607 cases, with category scores averaging task-level recall. GazeSieve selects up to four output timestamps, uses a frozen Qwen2.5-VL-7B-Instruct answerer, resizes marked frames to 640×480, and decodes at temperature 0. Across all controlled comparisons the answerer, decoding, candidate order, prompt policy, rendering, and output frame budget are fixed and only the selected frames change, so differences are attributable to selection. The selector's internal clue extraction and grounding use a low detector confidence threshold to preserve recall for small attended objects. All GazeSieve results require no task-specific parameter updates.

Baselines. For direct keyframe evaluation we compare across information regimes. *Agent(VLM)* is our strongest gaze-free baseline: it performs query-conditioned search over the same legal video interval and uses the same question, candidates, output budget, frozen answerer, prompt, decoding, and rendering as GazeSieve, but has no access to measured gaze. *T*(VLM)* is a gaze-free temporal-search selector with the same visual budget; *Colour-sim* samples 64 candidates and ranks by question–frame colour-histogram (HSV) similarity as a low-cost appearance retriever; *Shot* selects detected-shot representatives. For full-system context we list published StreamGaze results (GPT-4o, Qwen2.5-VL, ViSpeak, adapted ViSpeak), but these are external references rather than controlled baselines. All controlled selectors use no task-specific parameter updates.

We test four claims: (1) with a fixed answerer, gaze-guided selection improves full-system accuracy; (2) the selected frames cover human evidence windows on GazeKeyBench; (3) the gains come from semantic grounding combined with query-adaptive relations; (4) the measured *timing* of gaze contributes beyond a generic gaze-derived distribution.

5.1 StreamGaze Full-System Accuracy

Our headline is a controlled, same-backbone comparison: every arm feeds the *same* frozen Qwen2.5-VL-7B answerer through the same pipeline, decoding, prompt, and rendering, differing only in how frames are chosen. On the frozen 4,700-unit non-alert roster (eight tasks, Macro-8) with a matched per-unit frame budget, swapping uniform sampling for GazeSieve's frames raises Macro-8 from **58.0 to 60.8** (Figure 4). The gain is significant and specific to selection: $+2.78$ against a gaze-weighted uniform grid (95% CI $[0.86, 4.69]$, p=0.0046) and $+2.49$ against a saliency Top-K selector ($[0.63, 4.33]$, p=0.0097). Notably, the three non-GazeSieve strategies—plain uniform, gaze-weighted uniform, and saliency Top-K—are statistically indistinguishable ($p>0.7$): using gaze naively, as a saliency reweighting of ordinary sampling, does not beat uniform selection. What

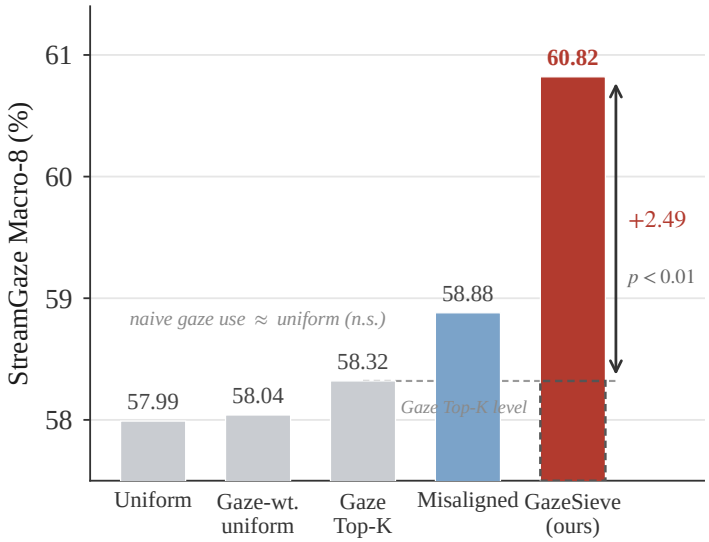


Figure 4: **StreamGaze controlled same-backbone comparison (%).** All arms share the frozen Qwen2.5-VL-7B answerer, pipeline, decoding, and budget (4,700-unit roster, Macro-8); only selection differs. Naive gaze uses are indistinguishable (p>0.7); GazeSieve beats the strongest naive-gaze baseline by +2.49 (p=0.0097).

moves the needle is GazeSieve's relation-aware composition of gaze-conditioned foveal evidence, not the mere presence of gaze.

The gains concentrate in specific task families (Appendix D). Relative to the saliency Top-K selector, GazeSieve improves most on temporal-relation tasks (GSM +10.4, OTP +4.1, SR +4.0, NFI +2.7), where the answer depends on ordering an anchor with its successor or aligning a sequence, and is roughly flat on present-state tasks (OI-E, OI-H) where a single attended frame already suffices—matching the design: gaze localizes attended evidence, and query-conditioned composition is what is needed when the answer spans successors, histories, or distributed observations.

Published end-to-end StreamGaze systems (Appendix E) run under different inputs, budgets, and decoding and are *not* directly comparable to our controlled arms; we list them only to situate task difficulty.

5.2 GazeKeyBench Direct Evidence Recall

Table 1 evaluates the selector itself on the 2,607 cases, removing a common ambiguity in downstream-only reporting: a model may answer correctly without seeing all evidence, or fail despite receiving it. The methods sit in different information regimes—GazeSieve receives noisy gaze while the VLM, CLIP, and shot baselines are gaze-free—so the question is how much temporal-evidence recovery gaze adds when available but not answer-labeled. GazeSieve reaches **56.0%** case-level recall, above a strong gaze-free Agent(VLM) at 52.0% and a VLM temporal selector at 49.0%. The gaze-free agent is itself a capable baseline: it runs query-conditioned search over video, differing primarily in that it has no access to gaze. The consistent +4.0 margin therefore isolates what measured gaze contributes rather than what agentic search alone recovers.

The gain holds uniformly across task types, adding about two points in each category (past 31.8 vs. 29.8, present 74.5 vs. 72.5, proactive 59.5 vs. 57.5), indicating that recovering the attended evidence helps whether it sits near or far from query time. Past-event tasks remain the hardest regime for every method, and gaze narrows that gap by the same margin it does elsewhere. Precision follows the same ranking rather than trading off against recall: GazeSieve attains the highest precision (22.5%), spending its compact budget on *complete* evidence groups rather than redundant frames, though absolute precision stays low across all methods because a four-frame budget is spent broadly. All values are point estimates; uncertainty is estimated with source-video clustering because cases from the same video are not independent.

Table 1: **GazeKeyBench evidence recall by information regime** (case-level %). Only GazeSieve uses measured gaze; Agent(VLM) is the strongest gaze-free query-conditioned baseline. Past, Pres., and Proact. are category recalls.

Method	Rec.	Prec.	Past	Pres.	Proact.
GazeSieve (video+gaze)	**56.0**	**22.5**	**31.8**	**74.5**	**59.5**
Agent(VLM)	52.0	16.5	29.8	72.5	57.5
$T^{\star}$(VLM)	49.0	14.5	26.0	69.8	53.0
Colour-sim (HSV)	36.1	9.5	11.5	59.2	39.0
Shot	31.4	8.9	6.5	60.1	37.0

Ablations, timing control, and compact evidence sets. Component ablations (Appendix B) show that semantic grounding gives the largest single-component drop (-4.5) and that removing relation-specific construction fails in task-aligned ways: anchor–successor for OTP, sequence alignment for GSM, option-aware coverage for NFI. The duration-only variant reaches 46.5% recall—close to the gaze-free baselines—so raw fixation density serves only as a candidate source, and stripping every component together (-9.5) costs less than the sum of the individual removals because the single-component drops overlap. A paired control that fixes the answerer, evaluation population, legal windows, and realized frame budgets (Appendix C) isolates the contribution of gaze *timing*: preserving measured timestamps improves Macro-8 from 58.9% to 60.8% ($+1.9$, source-clustered 95% CI $[0.3, 3.6]$, p=0.0247), with the largest gains on GSM and SR. Keeping up to three observations instead of only the first selected frame lifts Macro-8 from 52.5% to 55.8% on the 4,700-unit control split (Appendix F), and over the 7,391-unit population GazeSieve returns 3.86 output frames and 1.26 answer-model calls per unit.

6 Limitations

GazeSieve's compact output comes at the cost of selector-side computation: clue extraction and open-vocabulary grounding add model calls and latency over search-free baselines, and grounding can miss small or occluded objects. Our full-system gain is strongest on the co-designed Qwen2.5-VL-7B answerer ($+2.49$, p=0.0097); under the same protocol, all three tested backbones show positive point estimates, but the other two are not individually significant (Qwen2.5-VL-3B $+1.10$, InternVL3.5-8B $+1.42$; $p>0.1$). We therefore scope the strongest significance claim to the co-designed answerer; GazeKeyBench's answerer-independent results are unaffected. Finally, the benchmark contains 2,607 cases from 99 videos and assumes synchronized, user-consented gaze, so broader sources and gaze-data minimization are important for deployment.

7 Conclusion

We introduced GazeKeyBench, which scores keyframe selection directly through human-annotated evidence windows, required groups, and alternative sufficient sets, and GazeSieve, a training-free selector that turns measured gaze into fixation proposals, searches gaze-centered foveal crops under a temporal prior, and builds relation-aware evidence sets under a budget. Under a unified controlled protocol, GazeSieve improves StreamGaze accuracy and GazeKeyBench evidence recall over gaze-free and naive-gaze selectors, including a strong Agent(VLM) baseline. A paired control further confirms that gaze's value lies in its temporal registration. Overall, measured gaze acts as a compact temporal evidence index for egocentric video question answering.

AI use statement

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for [REQUIRED: fill in the tasks with required disclosure]. We have not used generative AI tools for [other tasks with required disclosure], and [the rest of the required disclosure tasks] are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoQA in long-form egocentric videos. *arXiv preprint arXiv:2408.14469*, 2024. URL https://arxiv.org/abs/2408.14469.

Alireza Fathi, Yin Li, and James M. Rehg. Learning to recognize daily actions using gaze. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 314–327, 2012.

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

Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 754–769, 2018.

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

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.

Yin Li, Zhefan Ye, and James M. Rehg. Delving into egocentric actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 287–295, 2015.

Miao Liu, Siyu Tang, Yin Li, and James M. Rehg. Forecasting human-object interaction: Joint prediction of motor attention and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 704–721, 2020.

Kyle Min and Jason J. Corso. Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1069–1078, 2021.

Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. In the eye of MLLM: Benchmarking egocentric video intent understanding with gaze-guided prompting. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. arXiv:2509.07447.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S. Davis. AdaFrame: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1278–1287, 2019.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. T*: Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Algorithm 1 GazeSieve: Gaze-Conditioned Foveal Evidence Search

Require: legal interval V_{I_q}, gaze G_{I_q}, question q, candidates C, search budget B_s, output budget B_o, grid size g
Ensure: ordered evidence frames $\mathcal{K}$

Phase 1: Fixation Proposal and Concept Extraction

$V_{\text{fix}} \leftarrow \textsc{FixationFilter}(V_{I_q}, G_{I_q})$
$S_{\text{gaze}} \leftarrow \textsc{GazeSaliency}(V_{\text{fix}}, G_{I_q})$
$V_N^{\text{gaze}} \leftarrow \textsc{TemporalNMS}(V_{\text{fix}}, S_{\text{gaze}})$
$\mathcal{O} \leftarrow \textsc{ExtractConcepts}(V_N^{\text{gaze}}, q, C)$

Phase 2: Foveal Temporal Search

$P \leftarrow \textsc{GazePrior}(S_{\text{gaze}})$
$S \leftarrow \mathbf{0}^L;\ N_v \leftarrow \mathbf{1}^L;\ F \leftarrow \emptyset$
$R_q \leftarrow \textsc{EvidenceRequirements}(q, C)$
while R_q not satisfied **and** $B_s > 0$ **do**
  $I \leftarrow \textsc{Sample}(P \odot N_v, \min(g^2, B_s))$
  $G_{\text{gaze}} \leftarrow \textsc{FovealMosaic}(V_{I_q}[I], G_{I_q}[I])$
  $C_I \leftarrow \textsc{Ground}(G_{\text{gaze}}, \mathcal{O})$
  $(S, N_v, F, R_q) \leftarrow \textsc{Update}(S, N_v, F, R_q, I, C_I)$
  $P \leftarrow \textsc{UpdateDistribution}(S, N_v)$
  $B_s \leftarrow B_s - |I|$
end while
$\mathcal{K} \leftarrow \textsc{RelationAwareFinalize}(F, S, q, C, B_o)$
return $\textsc{ChronologicalSort}(\mathcal{K})$

A Algorithm Details

Algorithm 1 gives the full procedure described in Section 3.

B Component Ablations

Table 2 isolates each component as a mechanism check. Removing semantic grounding causes the largest single-component drop (-4.5), and removing relation-specific construction produces task-aligned failures: anchor–successor for OTP, sequence alignment for GSM, option-aware coverage for NFI. The duration-only variant reaches 46.5% recall—close to the gaze-free baselines—showing that raw fixation density serves only as a candidate source. The single-component drops are partly overlapping rather than additive, so stripping every component together (-9.5) costs less than the sum of the individual removals. Gaze events become useful when semantic projection and query-adaptive temporal composition are applied *together*.

Table 2: Component ablations on GazeKeyBench (recall %). Drops show that gaze saliency alone is not enough; grounding and relation-aware construction are both needed.

Variant	Recall	Δ	Most affected
Full GazeSieve	56.0	—	—
w/o grounding	51.5	-4.5	OI-H
w/o successor	53.6	-2.4	OTP
w/o sequence	53.0	-3.0	GSM
w/o coverage	54.2	-1.8	NFI
w/o scene contrast	55.0	-1.0	SR
Duration-only	46.5	-9.5	GSM

Does measured gaze timing matter? We test the temporal registration of gaze while holding the answerer and evaluation population fixed. The *aligned* arm preserves each observation's measured timestamp; the *reassigned* arm applies one deterministic salted circular temporal reassignment within the inherited legal interval. Both arms share the same non-alert units, legal windows, realized frame budgets, evaluator, prompt policy, and forbidden-fallback rule. Preserving measured timing improves Macro-8 from 58.9% to 60.8% ($+1.9$), with a source-clustered 95% CI of $[0.3, 3.6]$ and a

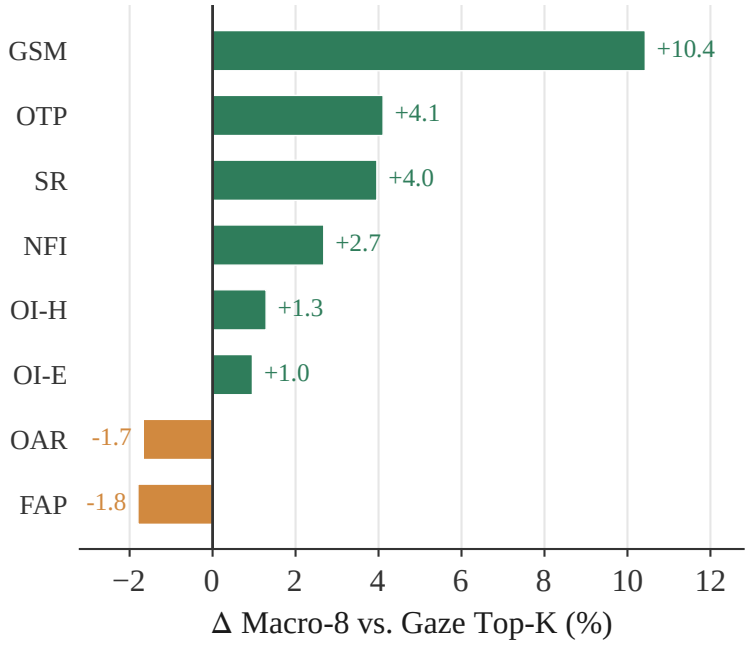


Figure 5: **Per-task gain over gaze Top-K** (Macro-8 Δ, same answerer). Gains concentrate on temporal-relation tasks (GSM, OTP, SR, NFI); present-state tasks are flat and OAR/FAP regress slightly.

Table 3: Published end-to-end StreamGaze systems, *external context only*; not comparable to our controlled arms. Macro over ten tasks.

System	Setting	Macro
GPT-4o	zero-shot, 16 frames	53.5
Qwen2.5-VL	zero-shot, adaptive	47.8
ViSpeak	zero-shot, 1 fps	46.7
Adapted ViSpeak	SFT, 1 fps	54.2

two-sided cluster sign-flip test p=0.0247. The gain is largest on GSM and SR, where answers depend on localized-in-time evidence—direct support that gaze's value lies in its temporal registration.

C Timing Control Protocol

We test the temporal registration of gaze while holding the answerer and evaluation population fixed. The *aligned* arm preserves each observation's measured timestamp; the *reassigned* arm applies one deterministic salted circular temporal reassignment within the inherited legal interval. Both arms share the same non-alert units, legal windows, realized frame budgets, evaluator, prompt policy, and forbidden-fallback rule. Preserving measured timing improves Macro-8 from 58.9% to 60.8% (+1.9), with a source-clustered 95% CI of $[0.3, 3.6]$ and a two-sided cluster sign-flip test p=0.0247. The gain is largest on GSM and SR, where answers depend on localized-in-time evidence—direct support that gaze's value lies in its temporal registration.

D Per-Task Breakdown

E External Reference Systems

F Compact Evidence Sets

Retaining up to three observations instead of only the first selected frame improves Macro-8 from 52.5% to 55.8% on the 4,700-unit control split (443 incorrect→correct vs. 291 correct→incorrect):

a small ordered evidence set is more reliable than a single frame. Over the 7,391-unit population, GazeSieve selects 3.86 output frames and makes 1.26 answer-model calls per unit, so the answering context stays compact even though selector-side clue extraction and grounding add computation.