EGOTRACE: CUE-CONDITIONED CONTEXTUAL REINSTATEMENT FOR LONG-HORIZON EGOCENTRIC VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon video question answering often begins with an incomplete cue: a question names part of an experience without specifying the evidence needed to answer it. Retrieving isolated records by similarity is fragile here, because similar encounters recur and the record that matches a question may not be the one that answers it. Inspired by contextual reinstatement in episodic memory, we introduce EgoTrace, which keeps original captions and aligned speech unchanged and links them to a *scene trace*: source-linked descriptions of how the surrounding situation develops. A Controller iteratively chooses among searching for records, reinstating the context around a retrieved record, and opening original evidence; a separate Reader answers from the opened sources and their linked context. Because the trace guides navigation rather than defining retrieval units, the extent of recovered context adapts to each question instead of being fixed by a window or pre-segmented events, and opened evidence can turn an incomplete match into a more precise cue. The trace is built in a single chronological pass and can be extended as footage arrives. EgoTrace achieves the highest reported scores of 74.6% on EgoLifeQA and 33.4% on MM-Lifelong Month validation, exceeding prior results by 3.4 and 8.9 percentage points, respectively. Ablations show that trace descriptions add value beyond temporal access alone, while direct source retrieval outperforms tested graph and semantic-memory additions.

Keywords: Egocentric video QA; long-horizon memory; contextual reinstatement.

1 INTRODUCTION

An assistant that remembers recorded experience often receives only a fragment of what it needs to recall. A question may name a conversation, an object, or a familiar activity, while leaving the answer-bearing detail unspecified. Across the week- and month-long histories of EgoLife, Ego-R1, and MM-Lifelong (Yang et al., 2025; Tian et al., 2025; Chen et al., 2026a), retrieval-based systems index short caption and speech records and retrieve those that best match the question. Two properties of long egocentric recordings make such matching insufficient.

First, similar encounters recur. Over a week, many records mention the same people, objects, or activities; without knowing which activity each belongs to, near-duplicates are hard to tell apart, and a strict top-k cutoff may discard the decisive one (Figure 1a). Candidates therefore need context that situates them within their activity, and the model needs a way to look beyond the initial ranking.

Second, an event can span many records while the answer lies in only one. A record of leaving a cafe may match a question about that departure, but the reason appears minutes earlier in a phone call (Figure 1b). A fixed temporal window (Choi et al., 2026) cannot resolve this: a narrow one misses the call, a wide one fills the context with unrelated records. Pre-segmented events do not remove this choice; they fix the boundary at construction time, before the question is known. Exploration must be directed by the situation around the hit, adapting its extent to the question.

Human episodic memory suggests a remedy. A partial cue can reactivate other elements of an episode, and reinstated temporal context shapes what is recalled next (Horner et al., 2015; Manning et al., 2011; Kragel et al., 2021; Polyn et al., 2009). Long-term video-memory systems commonly

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



Figure 1: **Limits of matching isolated records.** (a) Recurring near-duplicates are hard to rank. (b) The search hit (leaving) and the answer-bearing call belong to one visit. A fixed window around the hit misses the call; reinstating the visit reaches it, and the call supplies a new cue that finds the earlier repair. Examples are synthetic.

organize experience through entities, relations, consolidated facts, or narrative summaries (Long et al., 2025; Yeo et al., 2026; Li et al., 2026; Chen et al., 2026b), which must anticipate before any question is asked which part of an event will matter. We instead *preserve detailed evidence, and organize its context for reinstatement at retrieval time*.

We introduce **EgoTrace**, a memory framework for **cue-conditioned contextual reinstatement**. EgoTrace keeps original caption/ASR records unchanged and links them to a *scene trace*: chronological descriptions of the developing situation. Attached to search results, these descriptions show which activity each candidate belongs to, so a Controller can compare near-duplicate matches instead of trusting a strict ranking. Through REINSTATE, the Controller follows the trace from a hit to earlier or later records, and the recovered context tells it whether to continue, open a source, or search elsewhere, so the extent of exploration adapts to each question. The trace guides this navigation but does not define retrieval units: SEARCH ranks original records, and REINSTATE can continue across segment boundaries. OPEN exposes original evidence that grounds refined queries, and a separate Reader answers from the admitted sources and linked context.

Our contributions are threefold. First, we identify recurring near-duplicate encounters and long, sparsely informative events as obstacles to long-horizon recall, and introduce contextual reinstatement as a retrieval principle that addresses both. Second, we construct a source-linked scene trace that preserves activity continuity across sources, keeps original records directly accessible, and can be extended incrementally as footage arrives. Third, we develop a framework in which recovered context directs exploration and opened evidence grounds refined queries and the answer.

2 RELATED WORK

2.1 LONG-HORIZON VIDEO UNDERSTANDING AND MEMORY

Ego4D and EgoSchema establish benchmarks for understanding and remembering first-person experience (Grauman et al., 2021; Mangalam et al., 2023); EgoLife, Ego-R1, and MM-Lifelong extend evaluation to week- and month-long histories (Yang et al., 2025; Tian et al., 2025; Chen et al., 2026a). At these horizons, memory representations determine which parts of an experience remain accessible. M3-Agent, EgoGraph, and EGAgent organize encounters through entities, timestamped relations, and multimodal search (Long et al., 2025; Sun et al., 2026; Rege et al., 2026). LangRepo, WorldMM, and VideoRAG use textual hierarchies, complementary memory types, or multimodal graphs (Kahatapitiya et al., 2024; Yeo et al., 2026; Ren et al., 2025).

Among the closest approaches, MAGIC-Video combines graph retrieval with offline narrative chains, while EM^{2}Mem aligns evidence in event-indexed cells with temporal and semantic retrieval (Li et al., 2026; Chen et al., 2026b). Event-Causal RAG, R4DSG, and Semantic Event Graphs represent state transitions, persistent objects, or human–object interactions (Yan et al., 2026; Ma et al., 2026; Dixit & Liang, 2026). These representations organize retrieval around units fixed at construction, such as event cells, state transitions, or interaction subgraphs. EgoTrace instead retrieves original caption and speech records and uses source-linked trace segments only as navigable context: their boundaries do not limit what REINSTATE can recover, and their descriptions do not assert relations. This distinction between evidence and context underlies its interface: the model locates an observation, recovers its surroundings, and returns to the source to verify details.

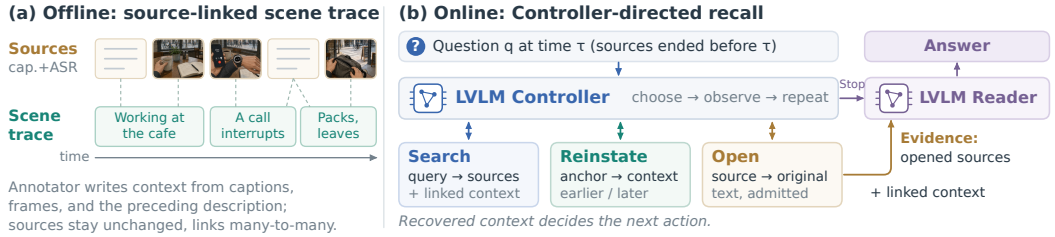

Figure 2: **EgoTrace overview.** (a) Unchanged caption/ASR sources are linked to a chronological scene trace. (b) An LVLM Controller iterates SEARCH, REINSTATE, and OPEN; a separate LVLM Reader answers from opened sources and linked context. A worked example is in Appendix A.

2.2 CONTEXT REINSTATEMENT IN EPISODIC RECALL

Naturalistic memory studies show that continuous experience is organized into events and that boundaries are associated with hippocampal responses and reinstatement (Baldassano et al., 2017; Ben-Yakov & Henson, 2018; Sols et al., 2017). At retrieval, partial cues can recover other elements of an episode (Horner et al., 2015), while reinstated temporal context helps explain the order of subsequent recall (Manning et al., 2011; Kragel et al., 2021). Retrieved-context models formalize how recalled content updates the cues for later recall (Polyn et al., 2009). EgoTrace adopts this functional relationship between recovered context and subsequent retrieval as a design principle. The surrounding situation is useful both for interpreting a retrieved observation and for discovering what else to recall. This motivates retaining source-linked contextual descriptions that can be recovered from a local cue and inspected in response to the question.

2.3 ADAPTIVE AND ASSOCIATIVE RETRIEVAL

VideoTree and DrVideo adapt retrieved context to the question (Wang et al., 2024; Ma et al., 2024); HippoMM combines content-adaptive segmentation with hierarchical recall and temporal expansion (Lin et al., 2025). MERIT uses multi-key clip retrieval and iterative query refinement, expanding a fixed-radius temporal neighborhood around each retrieved clip and filtering its captions into query-relevant information (Choi et al., 2026). EgoTrace makes temporal continuation a separate Controller decision. Its REINSTATE operation retrieves scene-trace descriptions and neighboring sources, from which the Controller can choose an earlier or later anchor for another inspection. The stored descriptions help it interpret an unfinished activity and decide where further evidence may lie. OPEN then exposes original source text to ground subsequent queries and the Reader's answer.

GraphRAG, HippoRAG, and LightRAG support semantic association (Edge et al., 2024; Gutiérrez et al., 2024; Guo et al., 2024), while CompassMem, REMem, and RippleMem enable iterative exploration of connected memories (Hu et al., 2026; Shu et al., 2026; Ji et al., 2026). More broadly, iterative retrieval agents in the ReAct and Toolformer paradigm select retrieval actions conditioned on accumulated observations (Yao et al., 2023; Schick et al., 2023). EgoTrace shares this iterative structure but adds a navigable layer: the scene trace provides source-linked context that supports lateral exploration, not only query-driven entry points. Its focus is the interplay between preserved evidence, recoverable context, and model-directed exploration: a match provides an entry point, and interpreting its context informs the next retrieval decision.

3 METHOD

EgoTrace combines preserved observations with source-linked contextual descriptions. We operationalize *contextual reinstatement* as recovering the context around a retrieved observation to guide subsequent retrieval. Figure 2 shows two model roles over a source retriever: a *Controller* chooses how to gather evidence, and a *Reader* produces the answer. At each round, the Controller can SEARCH for an entry point, REINSTATE the surrounding context, or OPEN an original source. The operations return information for the next decision and can be revisited in any supported order. In the default configuration, Controller and Reader are large vision–language models (LVLMs) operating on textual evidence; frames are used for offline context construction.

Unlike iterative retrieval agents that issue each step as an independent query conditioned on a scratchpad (Yao et al., 2023; Schick et al., 2023), EgoTrace navigates a structured evidence topology. Many-to-many provenance links between trace descriptions and source records let REINSTATE move laterally across sources, following an activity forward or back to its beginning, instead of reformulating a semantic query.

3.1 SOURCE EVIDENCE AND SCENE TRACE

Let $\mathcal{D} = \{s_j\}_{j=1}^{N}$ denote the source records, with $s_j = (I_j, x_j)$, interval $I_j = [a_j, b_j]$, and original text x_j. We use the approximately 30-second records underlying the video memory of MAGIC-Video (Li et al., 2026), whose captions integrate visual descriptions and aligned automatic speech recognition (ASR) content. These records remain unchanged and directly accessible. Given question q at time τ, only completed sources with $b_j \leq \tau$ are available.

Alongside $\mathcal{D}$, we store source-linked contextual descriptions in a *scene trace* $\mathcal{T} = (t_1, \ldots, t_M)$. These descriptions capture the surrounding situation, helping the model understand where an observation fits and what may clarify it. Each segment contains an interval, a contextual description, and links to its supporting sources. Chronological order preserves the progression of the experience; source identifiers and recorded spans establish a many-to-many correspondence between sources and trace segments. These links support both entry into the trace from a retrieved source and a return to original evidence after contextual exploration.

Original records preserve details such as speech, identity, attribution, and negation. The trace describes the developing situation, supporting decisions about whether to examine an activity's beginning, follow its continuation, or search elsewhere. Context is constructed independently of the question; its relevance is determined during retrieval.

The scene trace therefore differs from an event memory in three respects. First, segments are not retrieval units: SEARCH ranks source records, and descriptions are attached to candidates only as context. Second, segment boundaries do not delimit what is recovered: REINSTATE follows chronological links across them, so the extent of recovered context is decided per question, within a per-call budget, rather than fixed at construction. Third, descriptions record observed development without extracting entities, relations, or causes; interpreting what an earlier or later moment implies is left to the Controller at query time.

3.2 CONSTRUCTING THE SCENE TRACE

A multimodal annotator processes source records chronologically using captions, ordered frames, and the preceding unfinished description. It records how visible context, object states, spatial configuration, and ongoing activity develop across observations.

Annotation fragments are separated at meaningful changes in activity, interaction, goal, or navigation. Camera motion or a source-file boundary alone does not force a split; adjacent fragments can be joined across sources when an activity continues. Their chronological ordering records observed development without asserting causal relationships. Boundaries set annotation granularity and do not determine the extent of recall.

Every fragment retains its interval and source provenance, and recorded gaps remain explicit. A fragment becomes visible only after it and all supporting sources have ended. A partially available segment exposes only eligible fragments, preventing later content from entering an earlier query. Because each annotation depends only on the current source and preceding context, the construction is streaming-compatible; in our evaluation the full corpus is annotated in advance for convenience. Construction settings, availability rules, and annotation cost are in Appendix B.

3.3 CONTROLLER-DIRECTED RECALL

At each round r, the Controller selects an action $a_r = \pi_\theta(q, H_r)$ from SEARCH, REINSTATE, OPEN, or STOP, where π_θ is the prompted model without additional policy training. The returned observation updates the history: $H_{r+1} = H_r \parallel \mathrm{Access}(a_r; \mathcal{D}, \mathcal{T}, \tau)$. The Controller stops when it judges the evidence sufficient or reaches its interaction budget.

SEARCH: locate an entry point. The Controller formulates a query; Qwen3-Embedding-4B embeds it for cosine-similarity retrieval against stored source embeddings. BM25 provides a complementary lexical ranking, and reciprocal rank fusion combines the lists. Candidate records contain timestamps, caption previews, and source-linked trace descriptions attached after ranking. These descriptions let the Controller distinguish near-duplicate candidates that share similar keywords but belong to different activities, rather than relying on a strict relevance cutoff. A candidate becomes an *anchor* when selected for contextual exploration; it need only locate a useful part of the experience. Full ranking details are in Appendix B.

REINSTATE: recover the surrounding context. For a source anchor, provenance links locate the corresponding trace segments. An inspection follows chronological predecessors and successors, up to a per-call limit within the same day, and returns ordered descriptions with neighboring source candidates. The Controller reads this context to interpret the anchor and decide what remains unresolved. When an event spans many records and the answer-bearing moment is not the one that matched the query, the Controller can follow the trace to an earlier or later source and inspect again, adapting the extent and direction of exploration to the question. The descriptions expose a developing situation whose relevance may not be apparent from semantic similarity alone; SEARCH can then reach a related encounter elsewhere in the history.

OPEN: inspect original evidence. A source returned by either access operation can be opened directly, including sources surfaced through REINSTATE that would not have appeared in the original search ranking. OPEN exposes its unchanged caption/ASR text and associates it with eligible linked context. The Controller prompt requires subsequent global queries to be grounded in the question or opened source evidence; words from a generated trace description alone must not become a new global query.

Reader: answer from admitted evidence. Opening a source admits it to the Reader's evidence set $\mathcal{E}$, which contains opened source texts $\{(I_j, x_j) : s_j \in O\}$ and their deduplicated linked context. Source text takes precedence for exact wording and factual details; context supports interpretation and temporal relationships. SEARCH and REINSTATE expose candidate sources, while OPEN determines which original records are admitted for answering.

Cache-efficient interaction. The Controller appends each message and tool result without rewriting prior content, preserving the previous sequence as an exact prefix for backend KV-cache reuse. Table 6 reports the observed cache usage and Controller cost.

4 EXPERIMENTS

We evaluate the recall process in Figure 2: locating observations, recovering surrounding context, and inspecting source evidence. Main comparisons measure answer quality across benchmarks and backbones. Ablations distinguish the value of temporal access from the value of trace descriptions and examine additional semantic organization in SEARCH. Cost measurements and audited cases characterize how the Controller uses these operations.

4.1 EXPERIMENTAL SETUP

We evaluate on the 500-question, multiple-choice EgoLifeQA benchmark and the open-ended Month validation split of MM-Lifelong (Chen et al., 2026a), reporting accuracy and GPT-5-judged percentage scores, respectively. Main comparisons report the solvers shown in the tables and distinguish published baselines from our evaluations. EgoTrace uses Qwen3.8-27B for both Controller and Reader unless specified otherwise, with Qwen3-Embedding-4B for source embeddings. Repeated Controller runs and Reader-seed results are in Appendix C.1. Main experiments provide caption/ASR records and trace descriptions as textual evidence; the LVLM roles in Figure 2 do not involve online frame inspection. Section 4.3 begins with the preliminary visual-input study. Retrieval settings are in Appendix B.

Relationship to MAGIC-Video. MAGIC-Video (Li et al., 2026) provides the source-memory foundation and serves as our retrieval baseline. EgoTrace reuses its original caption/ASR records

Table 1: **EgoLifeQA comparison across backbones.** Accuracy (%); best in bold, second-best underlined. Baseline scores are from the original papers except *, our Qwen3.8-27B reproduction of MAGIC-Video. EgoTrace Qwen3.8-27B results average controller runs (Appendix C.1).

Method	Solver	EntityLog	EventRecall	HabitInsight	RelationMap	TaskMaster	Avg.
EgoGraph	GPT-4o	46.4	46.8	45.9	35.2	60.3	45.8
EGAgent	Gemini 2.5 Pro	54.4	57.1	60.3	62.4	<u>74.6</u>	57.5
WorldMM	GPT-5	62.4	64.3	75.4	62.4	71.4	65.6
EM^{2}Mem	GPT-5	60.8	61.1	63.9	72.8	<u>74.6</u>	66.0
MAGIC-Video	Qwen3.5-Flash	67.2	65.9	67.2	66.4	<u>74.6</u>	67.6
MERIT	GPT-5	67.2	70.6	73.8	74.4	71.4	71.2
MAGIC-Video*	Qwen3.8-27B	68.8	65.1	68.9	70.4	69.8	68.4
EgoTrace	Qwen3.5-Flash	68.0	70.6	70.5	<u>75.2</u>	71.4	71.2
EgoTrace	Qwen3.8-27B	**74.7**	**74.3**	76.0	74.9	71.4	<u>74.4</u>
EgoTrace	GPT-5	<u>73.6</u>	<u>71.4</u>	**82.0**	73.6	71.4	73.8
EgoTrace	DeepSeek-4.1-Flash	70.4	69.0	<u>77.0</u>	**78.4**	**84.1**	**74.6**

and adds a scene trace of source-linked contextual descriptions (Section 3.2). SEARCH uses direct dense+BM25 retrieval over those sources, and REINSTATE provides the Controller with a separate route through their temporal context. We also reproduce MAGIC-Video with Qwen3.8-27B to compare the methods using the same backbone; this result is marked with an asterisk in Table 1.

4.2 MAIN RESULTS

EgoLifeQA. Table 1 reports EgoTrace results across four backbones. EgoTrace reaches 74.6% with DeepSeek-4.1-Flash, exceeding the previous best score of 71.2% from MERIT by 3.4 percentage points. The default Qwen3.8-27B configuration achieves 74.4%, while GPT-5 and Qwen3.5-Flash achieve 73.8% and 71.2%, respectively. With GPT-5 as the solver for both methods, EgoTrace exceeds MERIT's published score by 2.6 points. This extends the comparison beyond Qwen backbones, although the methods retain their respective retrieval and evaluation pipelines.

Benefiting from stronger backbones. Across Qwen3.5-Flash, Qwen3.8-27B, GPT-5, and DeepSeek-4.1-Flash, EgoTrace remains effective (71.2–74.6%), showing that the interface transfers across base models. Because our hyperparameters and prompts were initially tuned for the Qwen family, GPT-5 and DeepSeek-4.1-Flash may be disadvantaged by limited model-specific adaptation. With other components fixed, upgrading Qwen3.5-Flash to Qwen3.8-27B improves MAGIC-Video from 67.6% to 68.4% ($+0.8$ points) and EgoTrace from 71.2% to 74.4% ($+3.2$ points), widening the method gap from 3.6 to 6.0 points. This controlled comparison changes both Controller and Reader; the cross-model results therefore indicate portability rather than a ranking of solver capabilities.

MM-Lifelong. Table 2 reports a Month validation score of 33.4% with Qwen3.8-27B as Controller and Reader, exceeding the 24.5% of MAGIC-Video with Qwen3.5-Flash by 8.9 points (36.3% relative improvement). EgoTrace improves causal reasoning (CR) from 29.4% to 43.5% and temporal reasoning (TR) from 27.1% to 41.0%. The improvements cover reasoning about both how situations develop and how observations relate over time.

Open-ended questions omit answer choices that might supply useful retrieval cues. Contextual reinstatement offers a route from an initially named encounter to details that the question leaves unspecified. This provides an interpretation of the MM-Lifelong gains: recovering the situation can help the model discover what to search for, as well as understand the evidence it finds. The aggregate comparison does not isolate this mechanism from other differences between the benchmarks.

4.3 ABLATION STUDY

We begin with the preliminary visual-input study, then examine temporal access and trace descriptions within REINSTATE, semantic additions to SEARCH, and sensitivity to retrieval budgets. Temporal context range is examined in Appendix C.2. Unless explicitly labeled otherwise, the ablations below use EgoLifeQA with Qwen3.8-27B for both the Controller and Reader; the detailed solver comparisons are reported in the main-results table.

Table 2: **MM-Lifelong Month validation.** GPT-5-judged scores (%). Baselines follow MAGIC-Video's main comparison; * marks results it cites from original papers. Subscripts give category sample counts.

Model	Cnt._{135}	ER_{108}	CR_{107}	TR_{72}	EvR_{54}	LCR_{46}	HD_{44}	AR_{26}	SI_{23}	SC_{5}	ET_{3}	Avg.
Long Video MLLMs												
VideoLLaMA3-7B	6.3	4.6	4.2	11.8	6.5	5.4	3.4	15.4	4.3	0.0	0.0	6.3
InternVideo2.5-8B	12.6	7.4	7.0	6.9	15.7	8.7	9.1	7.7	4.3	0.0	0.0	9.1
LongVA-7B	3.0	4.2	6.1	13.9	11.1	9.8	6.8	7.7	6.5	0.0	0.0	6.7
VideoChat-Flash-7B	8.5	8.3	7.9	13.9	13.0	9.8	6.8	7.7	4.3	0.0	0.0	9.1
Agentic Video MLLMs												
VideoMind-7B*	–	–	–	–	–	–	–	–	–	–	–	8.4
LongVT-7B*	–	–	–	–	–	–	–	–	–	–	–	7.5
DeepVideoDiscovery-7B*	–	–	–	–	–	–	–	–	–	–	–	10.6
ReMA-GPT-5*	–	–	–	–	–	–	–	–	–	–	–	18.6
MAGIC-Video (Qwen3.5-Flash)	7.4	31.0	29.4	27.1	36.1	40.2	15.9	32.7	17.4	10.0	0.0	24.5
EgoTrace (Qwen3.8-27B)	12.6	35.2	43.5	41.0	38.9	48.9	31.8	46.2	26.1	30.0	0.0	**33.4**

Cnt.: Counting; ER: Entity Recognition; CR: Causal Reasoning; TR: Temporal Reasoning; EvR: Event Recognition; LCR: Language Content Recall; HD: Hallucination Detection; AR: Attribute Recognition; SI: Social Interaction; SC: State Change; ET: Event Tracking.

Table 3: **Ablation on visual input (preliminary study).** Accuracy comes from the earlier single-seed EgoLifeQA evaluation with a Qwen3.5-Flash Reader; costs are per-question means from a separate three-question profile. Input/output tokens cover the full answering pipeline.

	Earlier evaluation	Cost profiling	
Input strategy	Acc. (%)	Input tokens	Output tokens
Broad frames	70.0	224.7k	19.4k
Sparse vision	**71.8**	103.8k	11.7k
No visual channel	70.8	98.7k	9.9k

Preliminary study: visual input. The earlier single-seed study with a Qwen3.5-Flash Reader finds a 1.0-point gain from sparse visual input over text-only evidence (Table 3). A separate three-question cost profile reports 224.7k/19.4k input/output tokens for broad frames and 98.7k/9.9k without a visual channel. These exploratory comparisons support using textual evidence to study the retrieval design; visual information remains available through captions and context construction.

Access to surrounding experience. Table 4 starts from SEARCH and OPEN without the scene-trace module (69.6%), which reduces EgoTrace to an iterative retrieval agent operating over the same source records with semantic search and direct evidence inspection. Retaining REINSTATE's temporal access while withholding its generated descriptions raises accuracy to 71.8%. This variant keeps neighboring source candidates, continuation through new anchors, and original captions. The 2.2-point gain supports a temporal route to evidence beyond the initial semantic matches.

Context beyond reachability. Table 4 further shows that restoring the trace descriptions raises accuracy from 71.8% to 74.4%, with the same temporal access operations available. The additional 2.6 points support the value of representing the surrounding situation beyond merely making its observations reachable. Trace descriptions help the Controller distinguish near-duplicate candidates that belong to different activities and identify which nearby records are worth inspecting, addressing the two retrieval challenges outlined in Section 1. The ablation removes these descriptions from every Controller and Reader view, including search previews and linked evidence context. The 2.6-point gain therefore measures their combined contribution to exploration and answering; it does not isolate continuity from other descriptive content.

Preserving direct access to evidence. Table 5 tests whether additional semantic organization improves the full system. We independently add Personalized PageRank (PPR) graph retrieval or precomputed semantic memory, following MAGIC-Video (Li et al., 2026), to the dense+BM25 implementation of SEARCH. Direct source retrieval achieves 74.4%, compared with 70.4% with PPR and 73.0% with semantic memory. These results favor keeping original observations directly ac-

Table 4: **Ablation on contextual reinstatement.** Temporal access and trace descriptions are added successively to dense+BM25 retrieval. EgoLifeQA accuracy (%).

Variant	Acc. (%)
Semantic search only	69.6
+ temporal access	71.8
+ trace descriptions (full)	**74.4**

Table 5: **Ablation on semantic organization.** PPR and semantic memory are separate additions to the full system's locator. EgoLifeQA accuracy (%).

Retrieval strategy	Acc. (%)
Dense+BM25	**74.4**
Dense+BM25 + PPR	70.4
Dense+BM25 + semantic memory	73.0

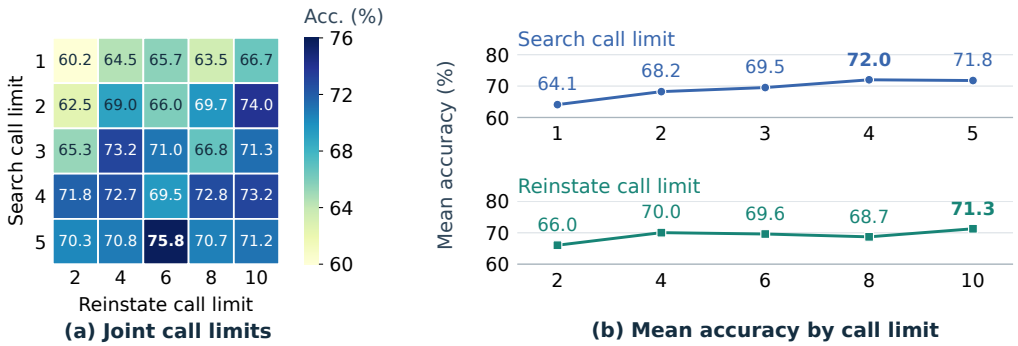

Figure 3: **Retrieval budget sensitivity.** EgoLifeQA accuracy (%) on 200 questions: (a) joint call limits; (b) mean accuracy versus each call limit, averaged over the five settings of the other limit.

cessible for question-dependent interpretation, a principle also emphasized by MERIT (Choi et al., 2026). Additional associations may alter source rankings or compete for the evidence budget, although the experiment does not isolate these mechanisms. Together with Table 4, the comparison supports the distinction between retaining context that helps interpret an encounter and adding precomputed semantic associations to source retrieval.

Sensitivity to retrieval budgets. Figure 3 suggests that sufficient source search is important for recall, with diminishing returns as the SEARCH budget grows. Increasing the REINSTATE budget also helps relative to its most restrictive setting, but the gains are less consistent. These results support iterative search and contextual exploration, with budgets tuned jointly rather than assuming larger limits always improve accuracy. Protocol details are in Appendix C.3.

Cost of additional interaction. Table 6 reports Controller-only cost per question. EgoTrace's prefix-preserving interaction achieves an 89.23% cache-hit rate with 28.0% fewer uncached input tokens, at 1.55× the baseline equivalent cost. Scene-trace annotation cost and its comparison with MAGIC-Video's multi-stage construction are in Appendix B.

4.4 QUALITATIVE ANALYSIS

Figure 4 examines how recovered context changes subsequent retrieval in EgoLifeQA Case 38. It aligns tool inputs, output excerpts, and information gained; the frames illustrate source intervals, while retrieval and answering use text. Calls 6–7 are grouped, and calls 10–12 are collapsed because they add no new evidence about the purchasing role.

Context resolves a purchasing role. The question asks what Shure, Katrina, and Lucia mainly purchase. The initial source mentions rum, fruit, and grapes without establishing the group's responsibility. REINSTATE exposes the preceding allocation of shopping tasks, and OPEN verifies references to "the drinks group," buying alcohol, and "you three know about drinks." Read together, these exchanges shift the interpretation from individual preferences to the group's assignment. The phrase "drinks group," absent from the question and the initial search results, then enters a new SEARCH query. The Controller opens a newly returned source stating that alcohol should have its

Table 6: **Controller cost.** EgoLifeQA means per question using GPT-5-equivalent pricing; Reader and scene-trace construction costs are excluded.

Method	Uncached input tokens	Cache hit (%)	Cost ($)
MAGIC-Video	54,961	13.77	0.1547
EgoTrace	39,589	89.23	0.2393

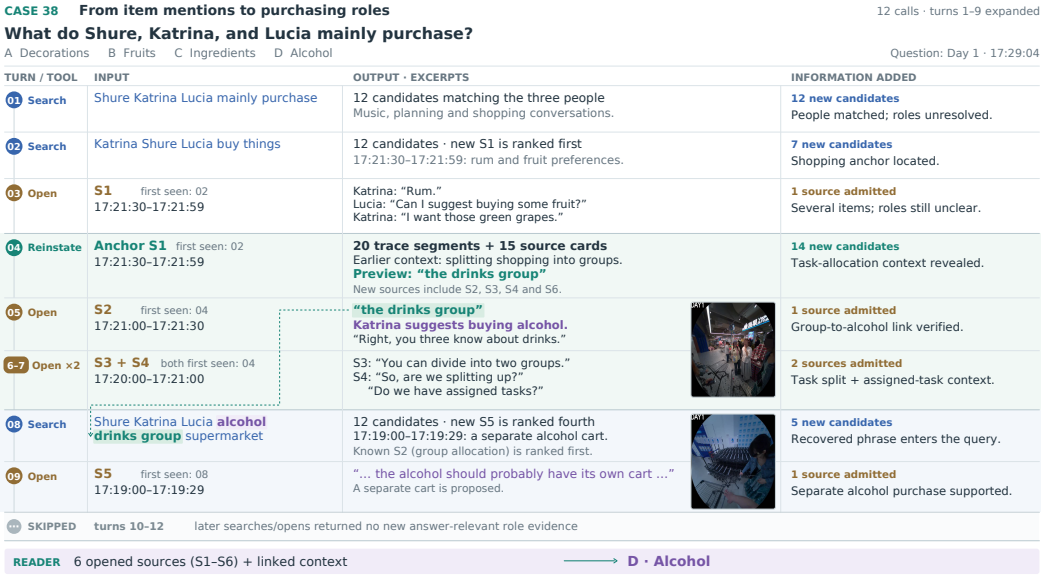

Figure 4: **Context guides retrieval.** Recovered task allocation supplies the "drinks group" cue for a subsequent search. Green: contextual cues; purple: answer evidence.

own cart, corroborating the grouping evidence already obtained. This trace illustrates how recovered context supplies a concrete query cue and directs further evidence gathering. Historical semantic-only runs also found relevant evidence; the run lacking the explicit grouping exchange selected fruit, whereas runs containing it predominantly selected alcohol. This pattern is consistent with the value of task context, without establishing that any individual call was necessary.

5 Conclusion

EgoTrace preserves detailed source evidence and stores source-linked contextual descriptions in a scene trace for reinstatement during recall. A Controller selects among locating an entry point, recovering its surrounding context, and opening original evidence; a Reader answers from admitted sources and linked context. This interaction lets recovered context guide what to retrieve next. Because each annotation depends only on the current source and its preceding context, the scene trace can be extended incrementally as new footage arrives, without revisiting earlier content. This contrasts with memory structures that require a global pass over the complete video (Li et al., 2026). EgoTrace achieves the highest reported scores of 74.6% on EgoLifeQA and 33.4% on MM-Lifelong Month validation. Ablations support the value of trace descriptions beyond temporal access and favor direct source retrieval over the tested semantic additions.

Limitations and future work. The scene trace's quality depends on source records and annotator. The current system operates on textual evidence at query time, with visual information preserved indirectly through captions and context construction. This focuses the study on retrieval mechanisms but limits answers that require fine-grained visual details beyond captions. Online visual retrieval, such as keyframes linked to reinstated context, is a natural extension. Both benchmarks evaluate first-person lifelogging, whose continuity the scene trace exploits; adapting it to less temporally coherent domains such as third-person video remains open.

AI USE STATEMENT

In this work, we used generative AI tools, namely Claude Code, Codex, and a DeepSeek-based agent harness, to assist with literature search, code implementation, and polishing of the manuscript text. The synthetic scene illustrations in Figures 1, 2, and 5 were produced with an image-generation model and are labeled as synthetic in their captions. The authors reviewed and verified all AI-assisted literature findings, code, and text. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W. Pillow, Uri Hasson, and Kenneth A. Norman. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721.e5, 2017. doi: 10.1016/j.neuron.2017.06.041. URL https://pubmed.ncbi.nlm.nih.gov/28772125/.

Aya Ben-Yakov and Richard N. Henson. The hippocampal film editor: Sensitivity and specificity to event boundaries in continuous experience. *Journal of Neuroscience*, 38(47):10057–10068, 2018. doi: 10.1523/JNEUROSCI.0524-18.2018. URL https://pubmed.ncbi.nlm.nih.gov/30301758/.

Guo Chen, Lidong Lu, Yicheng Liu, Liangrui Dong, Lidong Zou, Jixin Lv, Zhenquan Li, Xinyi Mao, Baoqi Pei, Shihao Wang, Zhiqi Li, Karan Sapra, Fuxiao Liu, Yin-Dong Zheng, Yifei Huang, Limin Wang, Zhiding Yu, Andrew Tao, Guilin Liu, and Tong Lu. Towards multimodal lifelong understanding: A dataset and agentic baseline, 2026a. URL https://arxiv.org/abs/2603.05484.

Yijun Chen, Yaqi Zheng, Yanya Li, Boyi Xiao, Buqiang Xu, Shuofei Qiao, Jizhan Fang, Xinle Deng, Yunzhi Yao, Xuehai Wang, Liuxin Zhang, Hui Li, Huajun Chen, and Shumin Deng. EM^{2}Mem: Event-centric multimodal memory for large language models, 2026b. URL https://arxiv.org/abs/2609.00551.

Yeeun Choi, Youngbeom Yoo, Joon-Young Lee, Hyolim Kang, and Seon Joo Kim. Keep it simple: Multi-key episodic memory retrieval for ultra-long video understanding, 2026. URL https://arxiv.org/abs/2608.07663.

Aradhya Dixit and Tianxi Liang. Semantic Event Graphs for Long-Form Video Question Answering, 2026. URL https://arxiv.org/abs/2601.06097.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From Local to Global: A Graph RAG Approach to Query-Focused Summarization, 2024. URL https://arxiv.org/abs/2404.16130.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the World in 3,000 Hours of Egocentric Video, 2021. URL https://arxiv.org/abs/2110.07058.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. LightRAG: Simple and Fast Retrieval-Augmented Generation, 2024. URL https://arxiv.org/abs/2410.05779.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models, 2024. URL https://arxiv.org/abs/2405.14831.

Aidan J. Horner, James A. Bisby, Daniel Bush, Wen-Jing Lin, and Neil Burgess. Evidence for holistic episodic recollection via hippocampal pattern completion. *Nature Communications*, 6: 7462, 2015. doi: 10.1038/ncomms8462. URL https://pubmed.ncbi.nlm.nih.gov/26136141/.

Yuyang Hu, Jiongnan Liu, Jiejun Tan, Yutao Zhu, and Zhicheng Dou. Memory matters more: Event-centric memory as a logic map for agent searching and reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 22389–22407. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1123. URL https://aclanthology.org/2026.findings-acl.1123/.

Jingbo Ji, Lingyi Li, Xilong Cheng, Yuhao Zhou, Wenji Zhang, Yuting Tan, and Yunxiao Qin. RippleMem: From Isolated Retrieval to Associative Recollection for Long-Term Agent Memory, 2026. URL https://arxiv.org/abs/2608.13334.

Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S. Ryoo. Language Repository for Long Video Understanding, 2024. URL https://arxiv.org/abs/2403.14622.

James E. Kragel, Youssef Ezzyat, Bradley C. Lega, Michael R. Sperling, Gregory A. Worrell, Robert E. Gross, Barbara C. Jobst, Sameer A. Sheth, Kareem A. Zaghloul, Joel M. Stein, and Michael J. Kahana. Distinct cortical systems reinstate the content and context of episodic memories. *Nature Communications*, 12(1):4444, 2021. doi: 10.1038/s41467-021-24393-1. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC8295370/.

Jiazheng Li, Chi-Hao Wu, Yunze Liu, Kaize Ding, Jundong Li, and Chuxu Zhang. Bridging modalities, spanning time: Structured memory for ultra-long agentic video reasoning, 2026. URL https://arxiv.org/abs/2605.08271.

Yueqian Lin, Jingyang Zhang, Qinsi Wang, Hancheng Ye, Yuzhe Fu, Yudong Liu, Hai "Helen" Li, and Yiran Chen. HippoMM: Hippocampal-inspired Multimodal Memory for Long Audiovisual Event Understanding, 2025. URL https://arxiv.org/abs/2504.10739.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, Listening, Remembering, and Reasoning: A Multimodal Agent with Long-Term Memory, 2025. URL https://arxiv.org/abs/2508.09736.

Ke Ma, Yamin Mao, Weiming Li, Shuai Tan, Yijie Zhong, Hao Chen, Haofen Wang, and Meng Wang. R4DSG: Relative 4D Scene Graph Memory for Object-Centric Question Answering in Long Egocentric Video, 2026. URL https://arxiv.org/abs/2608.11017.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. DrVideo: Document Retrieval Based Long Video Understanding, 2024. URL https://arxiv.org/abs/2406.12846.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding, 2023. URL https://arxiv.org/abs/2308.09126.

Jeremy R. Manning, Sean M. Polyn, Gordon H. Baltuch, Brian Litt, and Michael J. Kahana. Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences*, 108(31):12893–12897, 2011. doi: 10.1073/pnas.1015174108. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3150951/.

Sean M. Polyn, Kenneth A. Norman, and Michael J. Kahana. A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, 116(1):129–156, 2009. doi: 10.1037/a0014420. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2630591/.

Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. Agentic Very Long Video Understanding, 2026. URL https://arxiv.org/abs/2601.18157.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. VideoRAG: Retrieval-Augmented Generation with Extreme Long-Context Videos, 2025. URL https://arxiv.org/abs/2502.01549.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2302.04761.

Yiheng Shu, Saisri Padmaja Jonnalagedda, Xiang Gao, Bernal Jiménez Gutiérrez, Weijian Qi, Kamalika Das, Huan Sun, and Yu Su. REMem: Reasoning with Episodic Memory in Language Agent, 2026. URL https://arxiv.org/abs/2602.13530.

Ignasi Sols, Sarah DuBrow, Lila Davachi, and Lluìs Fuentemilla. Event boundaries trigger rapid memory reinstatement of the prior events to promote their representation in long-term memory. *Current Biology*, 27(22):3499–3504.e4, 2017. doi: 10.1016/j.cub.2017.09.057. URL https://pubmed.ncbi.nlm.nih.gov/29129536/.

Shitong Sun, Ke Han, Yukai Huang, Weitong Cai, and Jifei Song. EgoGraph: Temporal Knowledge Graph for Egocentric Video Understanding, 2026. URL https://arxiv.org/abs/2602.23709.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-R1: Chain-of-tool-thought for ultra-long egocentric video reasoning, 2025. URL https://arxiv.org/abs/2506.13654.

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, 2024. URL https://arxiv.org/abs/2405.19209.

Peizheng Yan, Yu Zhao, Liang Xie, Juntong Qi, Mingming Wang, and Erwei Yin. Event-Causal RAG: A Retrieval-Augmented Generation Framework for Long Video Reasoning in Complex Scenarios, 2026. URL https://arxiv.org/abs/2605.06185.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. EgoLife: Towards egocentric life assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.03803.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. WorldMM: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2512.02425.

A Illustrative Walkthrough

Figure 5 traces one question through all three operations on a synthetic example that combines the two challenges of Figure 1. Search for the question returns the record of leaving the cafe, which matches the wording but not the reason. Reinstate recovers the preceding scene-trace descriptions, in which a call interrupts the visit, and exposes the call as a neighboring source. Open admits its original ASR, "It's ready. We close at five." This text is still ambiguous, but it supplies grounded terms for a refined Search that reaches an earlier, separate encounter: a bicycle left with Alex for repair. The Reader answers from both opened sources. Neither the question nor the trace descriptions mention the bicycle; the link emerges only through recovered context and opened evidence. Section 4.4 examines a real trajectory from EgoLifeQA.

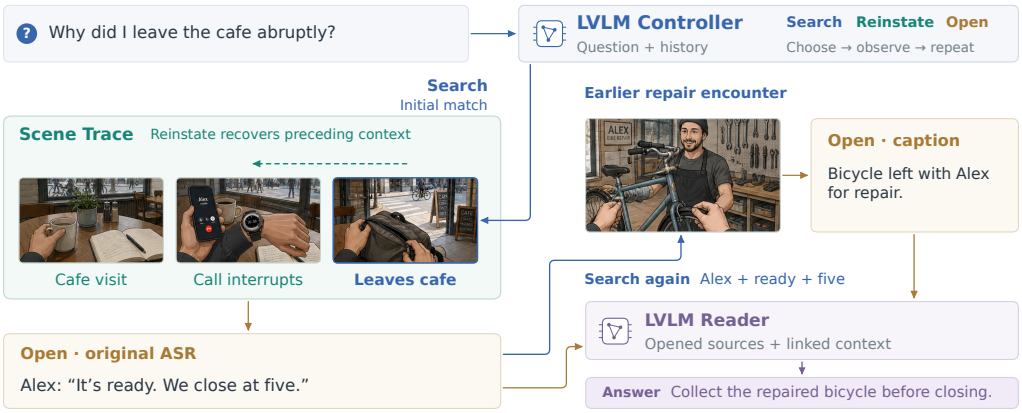

Figure 5: **Illustrative walkthrough.** A semantic match to leaving the cafe (Search) becomes informative when Reinstate recovers the surrounding scene trace and identifies the preceding call. Opening its original ASR supplies the grounded cue "Alex + ready + five" for a refined search that locates the earlier repair encounter; the Reader answers from both opened sources. Scene illustrations and excerpts are synthetic; online recall uses text.

B Retrieval Implementation Details

This appendix specifies the construction and temporal inspection of the scene trace introduced in Section 3.1. Search, Reinstate, and Open denote source retrieval, temporal inspection of surrounding context, and source admission, respectively; their implementation details are given below.

Scene-trace construction. For the annotation used here, candidate frames are sampled at 2 fps, and up to 20 are selected using temporal coverage and visual-change cues. Fragment offsets have 0.5-second resolution, with a nominal minimum duration of 2 seconds. Cross-source joining is restricted to gaps of at most 10 seconds and a maximum segment duration of 180 seconds. Each fragment retains its interval, boundary rationale, confidence, evidence-frame offsets, and exact source provenance.

The annotation used for the reported 3×3 evaluation contains 16,407 scene-trace segments and 17,865 fragments over A1_JAKE, with 2.637 segments per source caption on average and a 7.52% cross-source segment rate. The predecessor annotation contains 15,412 segments and 17,164 fragments, with 2.477 segments per caption and a 10.92% cross-source rate. The reported evaluation uses only the Qwen3.8 annotation and its matching text index; the two generations are not mixed.

Construction cost. The scene-trace annotation required 7,254 LLM calls over the 6,223 source captions (1.17 calls per source on average), comprising approximately 6,945 segmentation calls, 309 boundary rechecks, and 722 invalid-response retries. Each call receives one source caption, up to 20 selected frames, and the preceding partial scene description. Per-call token counts were

not logged in the annotation pipeline; we therefore report call counts rather than aggregate token consumption.

Because each source is annotated using only its caption and the preceding partial description, the construction requires a single chronological pass and is streaming-compatible: new segments can be appended as footage arrives without revisiting earlier content. In our evaluation, the full corpus was annotated in advance for convenience; the method does not require access to future observations. For comparison, MAGIC-Video constructs a Personalized PageRank graph over entity and event nodes, a precomputed semantic-memory bank, and narrative chain summaries, each through a separate global pass over the complete video (Li et al., 2026)—a limitation acknowledged in that work. EgoTrace replaces these multiple precomputation stages with one lightweight annotator pass, substantially reducing construction overhead while providing the temporal and contextual links needed for REINSTATE.

Source availability. Let $P(f)$ denote the sources supporting annotation fragment f. Eligibility at query time τ requires

$$\mathrm{end}(f) \le \tau \quad \text{and} \quad b_j \le \tau \text{ for every } s_j \in P(f). \tag{1}$$

Partially available segments expose only their eligible fragments rather than the complete description.

SEARCH: dense and lexical retrieval. The runtime ranks eligible episode records in the time-filtered index. Qwen3-Embedding-4B supplies query and caption embeddings, which are compared by cosine similarity. BM25Okapi uses lowercased, whitespace-tokenized display text, which includes the caption, timestamp, and granularity. Its corpus statistics come from the active index, while returned candidates are restricted to episode records. Each ranking considers at most 2,048 candidates; only positive BM25 scores contribute. Equal-weight reciprocal rank fusion gives

$$R(s_j \mid u) = \frac{\mathbb{I}[j \in L_d(u)]}{60 + r_d(j,u)} + \frac{\mathbb{I}[j \in L_b(u)]}{60 + r_b(j,u)}, \tag{2}$$

where L_d, L_b are candidate lists and r_d, r_b their one-based ranks. The backend retains up to 2,048 fused candidates before the Controller filters for completed, valid 30-second sources and exposes at most 12 cards. Caption-preview selection does not rerank these sources. Source-linked trace descriptions are attached after ranking. The active locator does not use PPR, entity-edge expansion, or a learned reranker.

REINSTATE: temporal expansion. Each inspection follows at most 32 predecessor and successor links per direction, with horizons of 120 seconds before and 330 seconds after each anchor. A response presents at most 20 scene-trace segments and 16 source candidates, preserving anchors and temporal coverage. Selection follows chronology and coverage rather than semantic scores. Continuation through newly surfaced sources is limited to five levels.

OPEN and interaction budgets. The reference runtime allows five SEARCH calls, ten REINSTATE calls, sixteen source admissions through OPEN, and 32 total Controller turns/tool calls. The append-only transcript is bounded at 512 KiB. A temporal response is bounded at 32 KiB; each linked Reader context is bounded at 24 KiB and twenty segments, with at most sixteen distinct contexts per question. Windows linked to multiple admitted sources are deduplicated. When no inspected window is linked to an opened source, its eligible source-bound summary provides a fallback context.

Prefix reuse. The runtime verifies that each Controller request begins with the entire previous message sequence and records the new suffix and tool-schema fingerprint. It reserves space for an assistant/tool exchange before dispatch rather than rewriting prior messages to meet the budget. These measures preserve prefix stability for backend caching; they do not implement a new KV compression or eviction algorithm.

Table 7: **Three-controller, three-Reader evaluation.** Values are EgoLifeQA accuracy (%). The last row reports the mean ± sample standard deviation across the three Controller runs.

Controller	Reader 42	Reader 43	Reader 44	2/3 vote
r01	73.4	73.6	74.4	74.2
r02	72.8	72.8	73.2	73.2
r03	**74.8**	**74.4**	**75.8**	**75.8**
Mean ± SD	73.7 ± 1.03	73.6 ± 0.80	74.5 ± 1.30	74.4 ± 1.31

C Additional Experimental Analyses

C.1 Repeated Controller and Reader Runs

Table 7 reports the three-controller-run evaluation of the default Qwen3.8-27B configuration on the 500-question EgoLifeQA benchmark. Each Controller run produces a set of sources admitted through OPEN and their linked context, which is evaluated with three Reader seeds. The final column reports accuracy after majority voting over these three outputs for each question. The Qwen3.8-27B result in Table 1 is the mean of the three voting accuracies (74.2%, 73.2%, and 75.8%): 74.4%, with a sample standard deviation of 1.31 percentage points. The main table reports the mean; variability and individual runs are presented here. These repeated-run statistics apply only to this configuration.

C.2 Temporal Context Range

Table 8 varies the candidate range and retained scene-trace size on EgoLifeQA; $a \to b$ uniformly retains b entries from a range of a. With range 32, retaining 20 entries performs best among the tested counts. At a fixed retained count of 20, reducing the range from 32 to 20 lowers accuracy by 3.2 points, suggesting that temporal coverage matters beyond the number of entries shown. The candidate range in this ablation is distinct from the runtime's per-direction traversal limit (Appendix B).

Table 8: **Ablation on temporal context range.** EgoLifeQA accuracy (%); $a \to b$ uniformly retains b entries from a range of a.

Range → retained	Acc. (%)
$32 \to 12$	71.8
$32 \to 16$	73.4
$32 \to 20$	**74.4**
$32 \to 24$	73.8
$20 \to 20$	71.2

C.3 Retrieval Budget Evaluation Protocol

Figure 3 uses the same 200-question EgoLifeQA subset (ID $\bmod 5 \in \{0, 1\}$) for every SEARCH/REINSTATE call-limit setting. Each setting runs retrieval once per question and evaluates the resulting evidence with three Reader seeds (42, 43, and 44), averaging 600 correctness indicators without majority voting. OPEN remains capped at 16 sources. The marginal means average over the five settings of the other limit, with the same questions shared across settings. This subset and aggregation protocol differs from the main 500-question voting evaluation.