EGOREVIEW: ACTIVE EVIDENCE ACQUISITION AND MEMORY SELF-IMPROVEMENT FOR EGOCENTRIC VIDEO

Anonymous authors
Paper under double-blind review

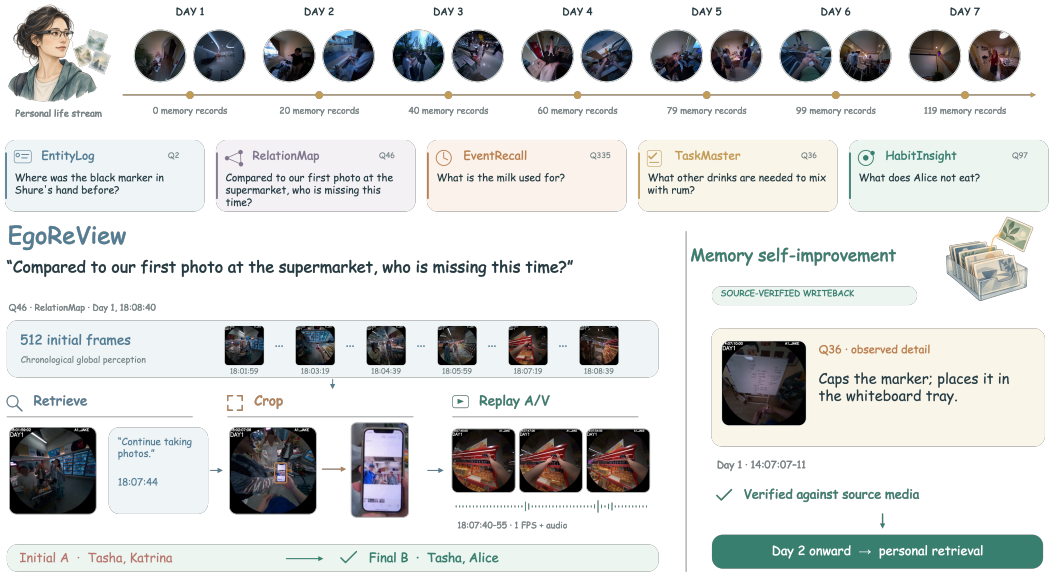

Figure 1: Overview of EgoReView + memory self-improvement. A chronological frame overview carries the initial reasoning, while retrieval, audiovisual revisiting, and cropping supply the evidence it lacks. Observations that pass source verification enter personal memory for future questions, with model weights and acquisition policy held fixed. The Q46 answer correction and the Q36 memory writeback are two separate real EgoLifeQA traces.

ABSTRACT

A personalized egocentric assistant must recover fine details from a long video history and retain the observations that later questions will depend on. Sparse frame overviews and caption memories, however, often omit exactly the evidence a person asks about. We present EgoReView, an agent harness that couples active evidence acquisition with memory self-improvement. Starting from an overview of up to 512 frames and an initial answer, the agent retrieves from memory, revisits video intervals, and inspects cropped regions as open questions require. It revises its answer only on decisive supplemental evidence, and observations that pass verification against their source media are added to the person's memory for subsequent days. Model weights and the acquisition policy remain fixed throughout. On EgoMemReason, EgoReView reaches 54.2% accuracy, the best among all compared methods, against 52.2% for its direct baseline. On a seven-day, single-person EgoLifeQA stream, retaining verified observations raises the fixed agent from 50.2% to 51.8%, while the direct baseline reaches 48.4%. These results suggest that memory refined through ordinary use is a practical route to personal adaptation, since every interaction can improve the evidence available to the next question.

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.

1 INTRODUCTION

An egocentric life assistant answers questions about a person's own experience: where an object was left, what happened during an activity last week, which routines recur (Yang et al., 2025; Wang et al., 2026a). Wearable cameras record the people, objects, and interactions behind that experience (Grauman et al., 2022), but a recording is not yet a memory. What the assistant needs is a memory that links observations across days, preserves brief details that later turn out to matter, and grows more useful as it answers more of its user's questions (Liu et al., 2026).

Reasoning across such extended visual context remains difficult even on curated long-form egocentric benchmarks (Mangalam et al., 2023). Memory-augmented video models lengthen the horizon a model can attend to (Song et al., 2024; He et al., 2024), and retrieval and tool use make hours of egocentric recordings searchable (Yang et al., 2025; Long et al., 2026). All of these approaches compress the recording. Sparse frame sampling trades detail for coverage and captions trade perception for searchability, so a short interaction can fall between two sampled frames and a caption can omit an object's state or an audible remark (Yeo et al., 2026; Li et al., 2026). Adaptive retrieval and targeted revisiting of the video recover part of what was lost (Tian et al., 2025; Yeo et al., 2026; Li et al., 2026), but they leave open which recovered observations should outlive the answer that prompted them. A memory built before any question arrives cannot anticipate every detail its user will later care about.

Language agents face a related problem. Persistent memory lets them reuse experience and personalize their responses (Park et al., 2023; Zhong et al., 2024), recent systems organize and update what they store dynamically (Xu et al., 2026; Chhikara et al., 2025), and long-term conversational benchmarks test whether that information stays accessible across sessions (Maharana et al., 2024; Wu et al., 2024a). In dialogue, however, the interaction history is itself text, so memory can be distilled from what was said. An egocentric assistant must often return to the raw recording to find an observation that its stored memory never contained.

We study *memory self-improvement* as a way to connect evidence recovery with memory retention. Questions expose gaps in memory, and resolving a gap yields exactly the kind of observation worth keeping. Our premise is that *evidence acquired for the current question can also refine the memory available to future questions* (Figure 1). Personalization then consists of accumulating facts about one person's experience, guided by the questions that person actually asks, while model weights and acquisition policy stay fixed.

EgoReView provides the active observation on which this process depends. It starts from an overview of up to 512 frames together with an initial answer, so direct video inference serves both as our baseline and as the agent's first stage. From there the model selects tools according to what it still does not know: it can search caption memory, revisit an interval through frame timestamps, sample a segment more densely, or crop a small region, and the audiovisual extension can replay video with its original audio. No order over these operations is prescribed. The original overview stays available throughout inference, and the initial answer is overturned only when supplemental evidence is decisive.

The memory self-improvement extension keeps the factual details these operations recover. Observations are described from the acquired media, checked against their sources, and stored with their temporal provenance. Once a day's questions have been answered, the verified records join the memory available the next day. Acceptance rests on source verification alone, so correctness labels for the questions are never required. Access is constrained by both event time and publication time, which prevents a later discovery from altering an earlier prediction.

We evaluate the two stages on complementary tasks. On EgoMemReason (Wang et al., 2026b), EgoReView reaches 54.2% accuracy over 500 questions against 52.2% for the same model's direct baseline. On a seven-day, single-person EgoLifeQA (Yang et al., 2025) stream, we compare the baseline, EgoReView, and its memory extension under a shared annotation-assisted input protocol: retaining verified observations raises accuracy from 50.2% to 51.8% with the same model, audiovisual tools, and acquisition policy, while the direct baseline obtains 48.4%. Once the shared initialization day is excluded, memory refinement adds 2.01 percentage points over EgoReView. This is evidence of short-term improvement in a personal memory; whether the benefit persists over months or years can only be settled by longer evaluations.

Our contributions are:

- An agent harness that combines global video observation with adaptive evidence acquisition and verified answer revision for egocentric question answering.
- A memory self-improvement mechanism that turns observations acquired during use into source-verified personal records, with causal access across days and with model weights and acquisition policy held fixed.
- An evaluation that connects long-video reasoning on EgoMemReason with chronological memory adaptation on EgoLifeQA, including a matched agent comparison that isolates the benefit of retaining supplemental memory.

2 Related Work

Long-term egocentric video understanding. Video LLMs already handle short clips well (Yu et al., 2019; Li et al., 2024; Liu et al., 2024; Han et al., 2025), and the long-video benchmarks of Wang et al. (2025); Wu et al. (2024b) have since shifted attention to reasoning over far longer recordings, where progress has been substantial. The strongest current multimodal models, GPT-6 Astra and Gemini-3.8-Flash, reach minute- to hour-level understanding by combining extended context lengths with agentic video understanding strategies.

Most of this progress concerns third-person footage, whereas physical AI and human interaction are naturally recorded from the first person. GPT-6 Astra has recently advanced short-horizon egocentric understanding through its agentic capabilities, and short horizons suffice for settings such as robotic manipulation. Assisting a person is different, because the relevant history spans hours, days, and eventually months or years. Yang et al. (2025) study egocentric life assistance at this scale, where the context of an end-to-end model is no longer sufficient on its own and an explicit visual memory becomes necessary.

Memory-based video LLMs. Retrieval-augmented generation (RAG), developed for long-context language models, transfers to long videos by retrieving relevant frames or clips instead of sampling them uniformly. Such methods match textual and visual prompts against the query so that the model can concentrate on the clips that matter (Jeong et al., 2025; Luo et al., 2026; Xu et al., 2025), and more recent work organizes what is retrieved through graph structures built over visual features (Ren et al., 2026; Shen et al., 2026; Xue et al., 2025). Their reach remains limited, however: retrieval decisions rest on text representations or plain similarity scores between visual and query features, and the length of the underlying sequences still stands in the way of holistic understanding and multi-step reasoning.

Because visual matching alone scales poorly, captioning has become the common way to store visual content in textual form. EgoRAG (Yang et al., 2025) organizes hierarchical text memory over a first-person stream and reasons across an entire day of activity. Ego-R1 (Tian et al., 2025) extends this line with vision-centered tools and iterative reasoning for long-horizon questions. HippoMM (Lin et al., 2026) proposes a dual-process memory that pairs semantic summaries with multimodal cues, and M3-Agent (Long et al., 2026) builds entity-centered long-term memory from multimodal context and queries it iteratively.

Agent harness and self-improvement. An agent harness organizes tool access, observations, context, and execution constraints around a foundation model. ReAct (Yao et al., 2022) interleaves reasoning with environmental interaction so that later actions can depend on newly acquired information. Improvement is also possible across tasks without touching model weights: Reflexion (Shinn et al., 2023) stores verbal reflections on task feedback in episodic memory, and Agentic Context Engineering (Zhang et al., 2026b) incrementally curates reusable strategies and knowledge into an evolving context. A second family instead modifies the executable mechanism. AutoHarness (Lou et al., 2026) synthesizes harness code from game feedback, the Darwin Gödel Machine (Zhang et al., 2026a) evolves coding agents through code changes and empirical evaluation, and ModularRSI (Wu et al., 2026) localizes changes to functional modules using contrasts between successful and failed trajectories, validation gates, and evolution tasks kept separate from downstream benchmarks. Taken together, these studies draw a line between adapting stored experience and adapting the mechanisms that govern execution.

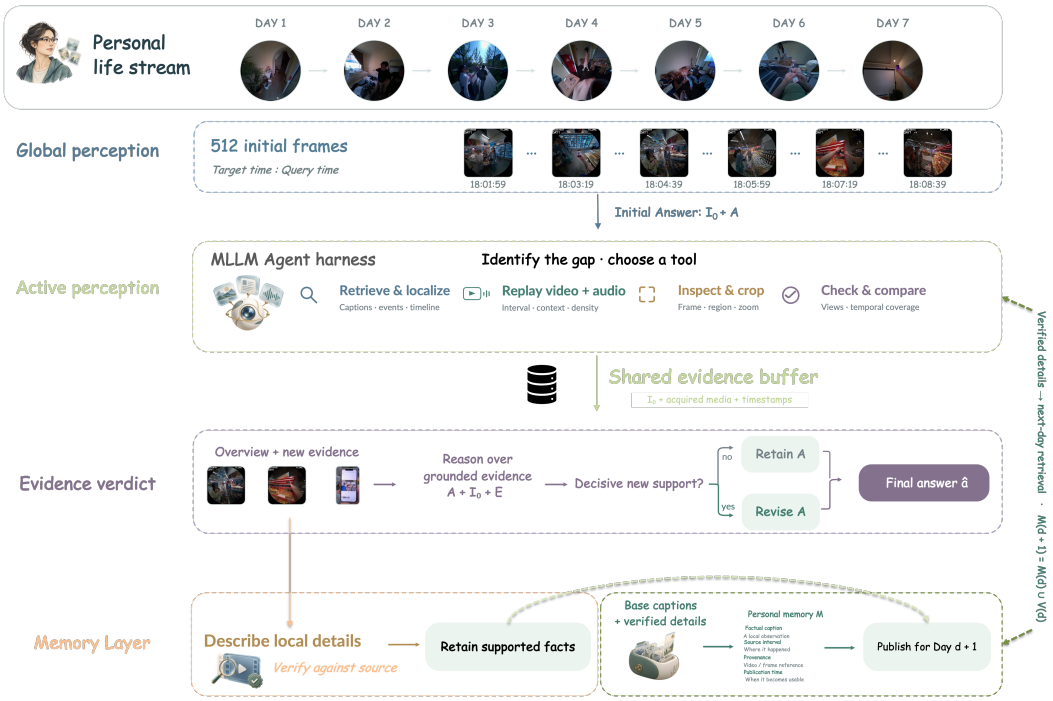

Figure 2: **Framework of EgoReView + memory self-improvement.** In the illustrated EgoLifeQA protocol, up to 512 timestamped frames from annotated target-to-query intervals form the overview $\mathcal{I}_0$, from which the model predicts an initial answer A. The harness then selects retrieval, audiovisual replay, cropping, and coverage checks as needed to acquire the missing evidence. The overview and the supplemental media together determine the final verdict, and A stands unless the new evidence justifies a revision. Local observations are separately verified against their source media and added to personal memory for the next day.

Personal egocentric assistance ties both forms of adaptation to one individual's continuing experience. EgoSelf (Wang et al., 2026a) couples interaction graph memory with user profiles for personalized responses, and StreamMemBench (Liu et al., 2026) evaluates whether agents retain egocentric evidence and act on interaction feedback in later assistance. Our concern is the refinement of perceptual memory during use: a question exposes an omitted detail, the agent revisits the relevant media, and source verification decides whether the recovered observation persists. The resulting records stay linked to their audiovisual sources and become available to the questions asked on subsequent days. The update relies on source verification rather than answer labels and leaves both model weights and acquisition policy unchanged. Feedback-driven policy editing is examined separately (Section 3.3), which lets factual memory accumulation and changes in evidence acquisition be evaluated as distinct sources of improvement.

3 METHOD

EgoReView pairs active observation within a single question with memory refinement across days (Figure 2). A fixed multimodal model first answers from a global overview, then acquires the evidence it is missing and verifies any revision it proposes. The memory self-improvement extension retains source-verified observations for later questions, leaving model weights and acquisition policy unchanged.

3.1 ACTIVE EVIDENCE ACQUISITION

Global perception. A question q specifies answer options $\mathcal{Y}_q$, a person p, and a query time τ_q. The accessible history $\mathcal{H}_q$ contains only that person's recordings up to τ_q, and the same cutoff applies

to caption memory $\mathcal{M}$; gaps in the recording stay unobserved. The initial images and direct answer are

$$\mathcal{I}_0 = \mathrm{Sample}_{512}(\mathcal{H}_q), \qquad a_0 = F_\theta^{\mathrm{direct}}(q, \mathcal{Y}_q, \mathcal{I}_0). \tag{1}$$

This inference is at once the baseline and the agent's first stage, so existing predictions are reused whenever the images, question, and model metadata match. On EgoMemReason, sampling selects up to 512 positions uniformly from a chronological 1 FPS candidate sequence spanning the recorded history. The agent also receives provenance $\mathcal{P}_0 = \{(h_i, t_i, c_i)\}$, which links each frame identifier to its actual time and source clip, and analyzes these images independently into z_0, a description of the available evidence and the gaps that remain, without ever seeing a_0.

Adaptive observation. At step k, the state $s_k = (\mathcal{I}_0, \mathcal{P}_0, z_0, \mathcal{E}_{<k}, \mathcal{L}_{<k}, b_k)$ collects the overview, the accumulated evidence, the action history, and the remaining budget. The fixed model chooses a tool, its arguments, and the gap that call should close:

$$u_k \sim \pi_{\theta,\phi}(\cdot \mid q, s_k), \qquad o_k = \mathrm{Execute}(u_k; \mathcal{H}_q, \mathcal{M}), \qquad \mathcal{E}_{\le k} = \mathcal{E}_{<k} \cup \{o_k\}. \tag{2}$$

The available tools cover caption and event retrieval, temporal revisiting, denser local sampling, spatial cropping, and coverage checks. Their order is not prescribed: a visible object can be cropped straight away, whereas an event absent from the overview is usually located through caption retrieval first. Temporal tools return ordered source frames, crops operate on original image content, and a caption match proposes a location rather than settling an answer.

The harness enforces person identity, causal access, valid references, and resource limits. When a requested time is unavailable, a nearby recorded frame may substitute for it within a fixed tolerance, with the actual timestamp exposed. Initial images stay attached at every decision alongside the selected supplemental evidence. Observation ends when the model stops, exhausts its budget, or repeatedly returns no new sources.

Evidence-based revision. From the retained supplemental images $\mathcal{I}_+$, the base agent forms an independent answer $a_1 = F_\theta^{\mathrm{answer}}(q, \mathcal{Y}_q, \mathcal{I}_0 \cup \mathcal{I}_+, \mathcal{P})$, where $\mathcal{P}$ extends $\mathcal{P}_0$ with the provenance of $\mathcal{I}_+$. This call sees neither a_0 nor the earlier analysis nor the retrieved caption text. When the two answers disagree, a verifier inspects both together with their visual support; writing $v = 1$ for a successful verification whose references to supplemental images are valid,

$$\widehat{a} = \begin{cases} a_1, & a_1 \neq a_0,\ \mathcal{I}_+ \neq \varnothing,\ v = 1, \\ a_0, & \text{otherwise.} \end{cases} \tag{3}$$

Missing evidence or a failed acquisition therefore leaves the initial answer standing. Verification is itself fallible, but the saved source references and trajectories keep every revision auditable. Section 4 specifies sampling, budgets, and the baseline's access to temporal metadata.

3.2 Memory Self-Improvement

Questions reveal details that an existing caption memory left out. EgoReView + memory self-improvement turns the evidence acquired to close these gaps into persistent, source-verified records. Personalization therefore lives entirely in the memory state, while the model parameters θ and the acquisition policy ϕ stay fixed.

Causal memory state. Let $\Delta\mathcal{M}_d$ hold the supplemental records available at the start of day d, and let τ_q be the query cutoff. The memory used for a question is

$$\mathcal{M}_d(q) = \mathcal{M}_{\mathrm{base}}^{\le \tau_q} \cup \{m \in \Delta\mathcal{M}_d : t_{\mathrm{end}}(m) \le \tau_q,\ t_{\mathrm{pub}}(m) \le \tau_q\}, \tag{4}$$

and the agent $\mathcal{A}_{\theta,\phi}$ returns the final answer and trajectory $(\widehat{a}_q, \ell_q) = \mathcal{A}_{\theta,\phi}(q, a_{0,q}, \mathcal{I}_{0,q}; \mathcal{M}_d(q))$. The supplemental store does not change within a day. Every record carries a caption, a source interval, a modality, an uncertainty, and a media reference. Because event time and publication time are both checked, a later discovery cannot retroactively affect an earlier prediction.

Audiovisual evidence. For EgoLifeQA we reuse the exact baseline images drawn from annotated target-to-query intervals (Section 4.3). Revisiting uploads video with its original audio, physically trimmed before τ_q. The audiovisual adapter conditions answering on a_0 and the collected evidence, and allows a revision only when the supplemental media references are decisive and valid. EgoReView and its memory extension share this adapter and keep all initial images attached during inference.

Verify and retain. The agent describes the acquired media in a call that sees neither the question nor the answer options. Once the day's predictions are saved, a separate verifier checks candidate details against their original media; it receives no question, no reference answer, and no correctness reward. Exact duplicates are removed by content and source hashes, and a deterministic, answer-independent rule selects at most B_M candidates per day for verification. Writing C_d for this selected set, the update is

$$V_d = \{m \in C_d : \mathrm{Verify}(m, \mathrm{source}(m)) = 1\}, \qquad \Delta\mathcal{M}_{d+1} = \Delta\mathcal{M}_d \cup V_d, \qquad \Delta\mathcal{M}_1 = \varnothing. \tag{5}$$

Verified records are published for the next day and supplement the provided memory. They preserve local observations rather than question-answer pairs or unsupported global conclusions. Source verification alone is the acceptance signal: daily correctness scores serve evaluation, never the selection of memory records.

3.3 Exploratory Policy Adaptation

We also consider a joint variant that proposes a policy edit ϕ'_d from correctness feedback and from diagnostic replays of tasks on completed days. Its next-day policy is

$$\phi_{d+1} = \begin{cases} \phi'_d, & g_d(\phi'_d; \phi_d) = 1, \\ \phi_d, & \text{otherwise.} \end{cases} \tag{6}$$

The gate g_d compares candidate against parent on separate historical validation tasks under identical memory, and demands improved accuracy within regression and cost bounds. Model weights stay fixed, and verified memory keeps accumulating through Equation 5 even when a policy edit is rejected. Appendix B specifies the editable modules and the acceptance thresholds.

4 Experiments

Core settings. We use Gemini-3.8-Flash with low thinking effort and fixed weights. Every agent reuses the baseline's initial answer and up to 512 images at 384×384, with at most eight tool calls per question. EgoMemReason allows 48 supplemental frames, 12 retained images, and sampling up to 8 FPS; EgoLifeQA allows 180 seconds of audiovisual replay at up to 4 FPS, with 1 FPS as the default. Memory self-improvement holds the policy fixed, verifies at most $B_M = 20$ candidate records per day, and publishes the accepted ones the next day. Appendix D reports the remaining settings.

4.1 EgoMemReason: setup and main results

EgoMemReason contains 500 multiple-choice questions spanning cumulative state tracking (CST), temporal counting (TC), event ordering (EO), event linking (EL), spatial preference (SP), and activity pattern (AP). The first four categories hold 100 questions each, SP and AP hold 50 each, and the hosted evaluator scores predictions against hidden answers, weighting overall accuracy by question count.

Initial frames cover the same person's recorded history up to the query time. Beyond them, the agent receives the provided DenseCaption memory and per-frame timestamps, both subject to that cutoff.

EgoReView reaches 54.2% against 52.2% for the direct baseline, ten additional correct answers and the best overall accuracy in Table 1. The gains fall in CST ($+4.0$ points), TC ($+4.0$), and EL ($+2.0$), with the other categories unchanged, consistent with selective acquisition helping most on questions about object state, counts, and event associations. Because our direct baseline already exceeds the highest public score of 39.6%, the same-model comparison is the more informative measure of what the agent contributes, although that contribution includes the additional information sources available to the agent.

Table 1: EgoMemReason results. Accuracy is reported in percent; bold indicates the best value in each column.

	Entity memory		Event memory		Behavior memory		
Method	CST	TC	EO	EL	SP	AP	Overall
Public leaderboard							
Gemini-3-Flash	46.0	28.0	36.0	44.0	44.0	44.0	39.6
Gemini-3.1-Pro	40.0	26.0	44.0	33.0	40.0	48.0	37.4
Qwen-3-VL-32B	35.0	46.0	27.0	27.0	**50.0**	46.0	36.8
Qwen-3-VL-30B-A3B	36.0	48.0	25.0	26.0	40.0	30.0	34.0
AVP	34.0	42.0	31.0	27.0	38.0	34.0	34.0
Molmo2-8B	36.0	**50.0**	27.0	25.0	34.0	22.0	33.2
InternVL3.5-38B	33.0	40.0	27.0	24.0	46.0	32.0	32.6
WorldMM	32.0	44.0	21.0	21.0	34.0	36.0	30.6
VideoLLaMA3-8B	23.0	31.0	27.0	32.0	38.0	36.0	30.0
Qwen-3-VL-8B	35.0	28.0	23.0	21.0	40.0	42.0	29.6
InternVL3.5-8B	23.0	29.0	23.0	27.0	34.0	42.0	28.0
GPT-5	29.0	42.0	20.0	18.0	32.0	28.0	27.8
Ego-R1	30.0	18.0	23.0	18.0	48.0	32.0	25.8
InternVideo-2.5-8B	29.0	27.0	25.0	15.0	32.0	32.0	25.6
StreamingVLM	25.0	29.0	21.0	20.0	20.0	32.0	24.2
SiLVR	31.0	14.0	27.0	17.0	18.0	28.0	22.4
LongVA-7B	22.0	18.0	20.0	20.0	20.0	22.0	20.6
Random	19.6	16.7	11.1	17.3	19.3	19.2	16.8
Our runs with Gemini-3.8-Flash							
Gemini-3.8-Flash (direct)	55.0	34.0	**68.0**	54.0	48.0	**52.0**	52.2
EgoReView	**59.0**	38.0	**68.0**	**56.0**	48.0	**52.0**	**54.2**

Public rows include named leaderboard entries. Our submissions were scored by the hosted evaluator. The highlighted row reports EgoReView.

Table 2: Component removals on EgoMemReason. Scores are accuracy (%); Δ is the difference from the full configuration. All six categories are retained.

Configuration	CST	TC	EO	EL	SP	AP	Overall	Δ
EgoReView (full)[†]	59.0	38.0	68.0	56.0	48.0	52.0	54.2	0.0
w/o memory retrieval	56.0	32.0	68.0	57.0	48.0	50.0	52.4	-1.8
w/o video revisiting	57.0	32.0	68.0	54.0	48.0	50.0	52.0	-2.2

[†]Full uses the earlier configuration reported in Table 1; the removal runs use a later code revision with the same cached baseline, base memory signature, and nominal budgets. Differences are descriptive and are not isolated effects from a matched code revision.

4.2 EgoReView component analysis

Removing caption, event, entity, and timeline lookup lowers accuracy to 52.4%, and disabling temporal acquisition while keeping initial frames and crops yields 52.0% (Table 2). Either removal drops TC from 38.0% to 32.0%, consistent with both localization and revisiting being needed to recover scattered events. The effect varies by category: EL rises to 57.0% without retrieval but falls to 54.0% without revisiting. Because the full result comes from an earlier code revision, these comparisons point to complementary roles for the two components without isolating their causal effects.

4.3 EgoLifeQA: memory self-improvement

We evaluate all 500 questions from participant A1_JAKE over seven days, 202 of which require audio. Initial frames are allocated to the annotated target-to-query intervals in proportion to their duration, with separate allocations kept for overlaps and recording gaps omitted; all methods share this annotation-assisted input protocol. The comparison covers the direct baseline, EgoReView,

Table 3: Daily forward accuracy (%) on EgoLifeQA, comparing the baseline, EgoReView, memory self-improvement, and exploratory joint memory and policy updates.

Day	N	Baseline	EgoReView	EgoReView + memory self-improvement	EgoReView + memory and policy updates
Day 1	102	49.02	**51.96**	**51.96**	**51.96**
Day 2	59	55.93	57.63	**59.32**	55.93
Day 3	85	51.76	51.76	51.76	**55.29**
Day 4	84	47.62	50.00	53.57	**54.76**
Day 5	75	41.33	44.00	**45.33**	40.00
Day 6	76	48.68	48.68	52.63	**56.58**
Day 7	19	36.84	**42.11**	**42.11**	31.58
Days 1–7	500	48.40	50.20	**51.80**	51.60
Days 2–7	398	48.24	49.75	**51.76**	51.51

Accuracy (%). Bold marks the best value in each row. Day 1 is the shared agent initialization. Aggregate scores pool correct answers and question counts; no per-day method selection is used.

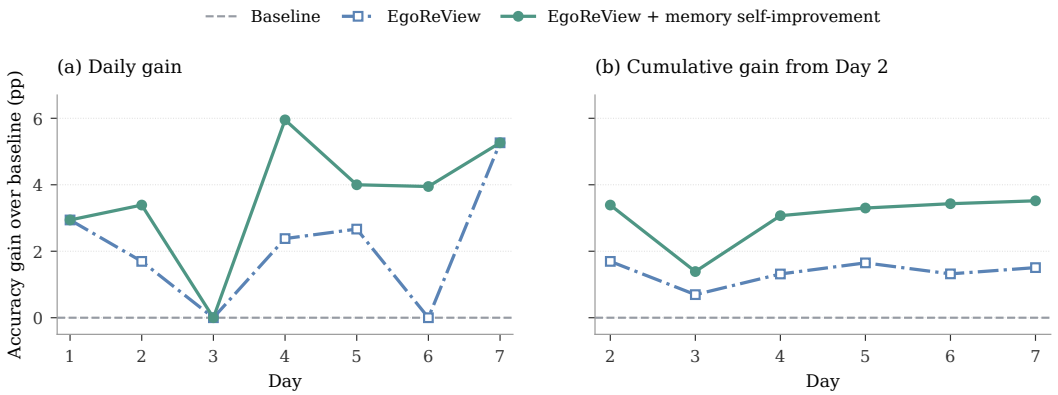

Figure 3: Gains over the direct baseline on EgoLifeQA. (a) Daily gain on the same questions. (b) Cumulative gain, pooling correct answers and question counts from Day 2 through each day. The dashed zero line denotes the baseline. Day 1 is the shared agent initialization.

and EgoReView + memory self-improvement. Both agents share audiovisual tools, a fixed policy, and the provided captions, and only the memory extension retains verified supplemental records. Since the baseline works from silent images, gains over it also reflect access to audio, whereas the agent-to-agent comparison isolates memory retention.

Predictions are saved before the daily update and never rerun under later memory. Day 1 serves as shared initialization, and Days 2–7 measure forward adaptation. The memory extension retains 119 verified records, and its maintenance cost is accounted for separately from inference. Aggregate accuracy pools correct answers across questions.

EgoReView + memory self-improvement achieves 51.8%, against 50.2% for EgoReView and 48.4% for the baseline (Table 3), or eight and seventeen more correct answers, respectively. Over Days 2–7, memory refinement gains 2.01 points on EgoReView (206/398 versus 198/398). It improves on Days 2, 4, 5, and 6 and ties on the others, with its largest gain on Day 6 ($+3.95$ points). Figure 3 plots daily and cumulative gains over the baseline; because question composition varies from day to day, the daily curve should not be read as monotonic improvement.

The category and modality breakdowns in Appendix A show gains in both modalities but considerable variation across reasoning categories. These results come from a single person, on data seen during development, and they support a modest benefit within seven days rather than sustained adaptation over months or years.

4.4 POLICY SELF-IMPROVEMENT AND THE DATA GAP

We also run the exploratory joint variant of Section 3.3, which adds feedback-driven policy edits to source-verified memory refinement and evaluates each edit on subsequent days. Joint memory and policy updates reach 51.6% (Table 3), 3.2 points above the baseline and 1.4 points above EgoReView, but one answer short of memory refinement alone (51.8%). They improve on memory refinement on Days 3, 4, and 6 and regress on Days 2, 5, and 7, so the policy edits add no further overall gain. Appendix B gives the update protocol.

The central distinction is between *verifying an observation* and *validating a reusable strategy*. A factual addition to memory can be checked against its source video, whereas a policy change needs recurring tasks and future feedback before it can be called general. Six update opportunities, each with at most eight historical validation questions, offer thin evidence for the second judgment, and selection may consequently favor recent questions over stable personal needs. Limited feedback is a plausible constraint on policy stability, though our experiment cannot establish that it causes the observed fluctuations.

The longitudinal resource behind these experiments, EgoLife, contains roughly 300 hours across six participants over one week, well under 1,000 hours in total, and our personalized evaluation covers a single participant. A substantial gap therefore separates the observation horizons now available from the months or years over which a personal assistant should adapt. Larger general egocentric corpora do not close it by themselves, since they lack continuous personal histories with recurring requests and linked feedback. What progress requires is both longer recordings per person and repeated opportunities to tell durable acquisition strategies apart from adaptation to isolated tasks.

A deployed egocentric assistant would share a person's visual and auditory context, meet the same information needs repeatedly, and receive corrections grounded in that person's own experience. Our benchmark feedback is only an offline proxy for this setting. Longer studies with real user feedback are needed to test whether memory and policy updates keep their benefits as routines and needs change.

5 CONCLUSION

We presented EgoReView with memory self-improvement, which connects active evidence acquisition to a persistent personal memory. Global observation supplies the initial context, targeted revisiting recovers the details it misses, and source verification makes those observations available to future questions. The full system reaches the best accuracy among all compared methods on EgoMemReason (54.2%), 2.0 points above its direct baseline, and on the seven-day EgoLifeQA stream, retaining verified observations lifts the fixed agent from 50.2% to 51.8%. Memory refinement thus emerges as a practical component of personalized egocentric assistance, preserving facts about a person's experience in response to the information needs that arise during use.

The broader opportunity is to let everyday assistance progressively enrich the memory that supports it. Our evidence covers one person over seven days, and exploratory policy updates add no overall gain. Larger datasets spanning more people and months or years of egocentric life, together with recurring tasks and reliable feedback, would permit a stronger evaluation of policy iteration and help separate reusable acquisition strategies from adaptation to a handful of recent questions. Future work should examine whether memory and policy updates sustain their benefits on subsequent tasks while keeping errors, forgetting, and cost under control, which would bring personal memory agents closer to reliable long-term use.

AI USE STATEMENT

In this work, we used generative AI tools only to edit the manuscript for language and readability. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, propose or refine hypotheses, design or give feedback on the research methodology or experiments, implement methods, translate, clean or reformat datasets, support qualitative analysis, or interpret results. The multimodal model Gemini-3.8-Flash serves as the backbone of the evaluated agents and is therefore part of the object of study (Section 4), not an aid to conducting the research. The authors reviewed all AI-edited text

and checked that the edits did not alter the technical content, claims, numerical results, or citations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

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

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, et al. Video-bench: Human-aligned video generation benchmark. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18858–18868. IEEE, 2025.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13504–13514. IEEE, 2024.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. Videorag: Retrieval-augmented generation over video corpus. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21278–21298, 2025.

Chenglin Li, Qianglong Chen, Feng Han, Yikun Wang, Xingxi Yin, Yan Gong, Ruilin Li, Yin Zhang, and Jiaqi Wang. Videothinker: Building agentic videollms with llm-guided tool reasoning. *arXiv preprint arXiv:2601.15724*, 2026.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22195–22206, 2024.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5968–5977, 2026.

Guanming Liu, Yuqi Ren, Hansu Gu, Peng Zhang, Weihang Wang, Jiahao Liu, Ning Gu, and Tun Lu. Streammembench: Streaming evaluation of agent memory for future-oriented assistance. *arXiv preprint arXiv:2606.14571*, 2026.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024.

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. In *International Conference on Learning Representations*, volume 2026, pp. 146197–146246, 2026.

Xinghua Lou, Miguel Lázaro-Gredilla, Antoine Dedieu, Carter Wendelken, Wolfgang Lehrach, and Kevin P Murphy. Autoharness: improving llm agents by automatically synthesizing a code harness. *arXiv preprint arXiv:2603.03329*, 2026.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, et al. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *Advances in Neural Information Processing Systems*, 38:168008–168033, 2026.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents, 2024. URL https://arxiv.org/abs/2402.17753.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. Videorag: Retrieval-augmented generation with extreme long-context videos. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2390–2401, 2026.

Xiaoqian Shen, Wenxuan Zhang, Jun Chen, and Mohamed Elhoseiny. Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding. *Advances in Neural Information Processing Systems*, 38:100805–100830, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232. IEEE, 2024.

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. *arXiv preprint arXiv:2506.13654*, 2025.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967. IEEE, 2025.

Yanshuo Wang, Yuan Xu, Xuesong Li, Jie Hong, Yizhou Wang, Chang Wen Chen, and Wentao Zhu. Egoself: From memory to personalized egocentric assistant. *arXiv preprint arXiv:2604.19564*, 2026a.

Ziyang Wang, Yue Zhang, Shoubin Yu, Ce Zhang, Zengqi Zhao, Jaehong Yoon, Hyunji Lee, Gedas Bertasius, and Mohit Bansal. Egomemreason: A memory-driven reasoning benchmark for long-horizon egocentric video understanding, 2026b. URL https://arxiv.org/abs/2605.09874.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024a.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024b.

Siwei Wu, Jincheng Ren, Yizhi Li, Haau-Sing Li, Chengran Yang, Yuxuan Zhang, Weicheng Gu, Jian Yang, Riza Batista-Navarro, Chuanyi Zhang, et al. Modularrsi: Modular and generalizable recursive harness self-improvement. *arXiv preprint arXiv:2609.14857*, 2026.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026.

Zeyu Xu, Junkang Zhang, Qiang Wang, and Yi Liu. E-vrag: Enhancing long video understanding with resource-efficient retrieval augmented generation. *arXiv preprint arXiv:2508.01546*, 2025.

Zhucun Xue, Jiangning Zhang, Xurong Xie, Yuxuan Cai, Yong Liu, Xiangtai Li, and Dacheng Tao. Adavideorag: Omni-contextual adaptive retrieval-augmented efficient long video understanding. *arXiv preprint arXiv:2506.13589*, 2025.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

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*, pp. 25599–25609, 2026.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 9127–9134, 2019.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin gödel machine: open-ended evolution of self-improving agents. In *International Conference on Learning Representations*, volume 2026, pp. 104223–104294, 2026a.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, et al. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, volume 2026, pp. 86069–86100, 2026b.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

A EgoLifeQA Category and Modality Analysis

Table 4 breaks the three main methods down by reasoning category and required modality, using the chronological predictions reported in Section 4.3. The upper block covers all 500 questions; the lower block drops the shared Day 1 initialization and evaluates forward adaptation on the remaining 398.

Table 4: EgoLifeQA category and modality accuracy (%) for the three main methods.

Method	Ent.	Rel.	Event	Task	Habit	Visual	Audio	Overall
All questions: Days 1–7 ($N = 500$)								
Baseline (512 frames)	55.20	32.80	58.73	50.79	42.62	50.67	45.05	48.40
EgoReView	54.40	36.80	59.52	55.56	44.26	52.01	47.52	50.20
EgoReView + memory self-improvement	59.20	36.00	59.52	58.73	45.90	53.02	50.00	51.80
Forward adaptation: Days 2–7 ($N = 398$)								
Baseline (512 frames)	56.07	33.67	57.27	47.73	38.46	51.41	42.95	48.24
EgoReView	55.14	37.76	57.27	52.27	41.03	52.21	45.64	49.75
EgoReView + memory self-improvement	60.75	36.73	57.27	56.82	43.59	53.41	48.99	51.76

Ent.: EntityLog; Rel.: RelationMap; Event: EventRecall; Task: TaskMaster; Habit: HabitInsight. Visual and Audio denote questions without and with required audio. The five categories contain 125, 125, 126, 63, and 61 questions; the modality subsets contain 298 and 202. Day 1 is shared initialization.

Across all seven days, memory self-improvement raises EntityLog accuracy from 54.40% to 59.20%, TaskMaster from 55.56% to 58.73%, and HabitInsight from 44.26% to 45.90% relative to EgoReView. EventRecall holds at 59.52%, while RelationMap slips from 36.80% to 36.00%. The overall gain thus comes from several categories rather than from a uniform benefit. Audio-required accuracy rises from 47.52% to 50.00% and visual accuracy from 52.01% to 53.02%; since both agents use the same audiovisual tools, this comparison measures what retaining verified observations contributes under a fixed policy.

The forward comparison repeats the same pattern: gains in EntityLog, TaskMaster, and HabitInsight, unchanged EventRecall, and a small decline in RelationMap. Audio-required accuracy improves by 3.35 percentage points and visual accuracy by 1.20 points. Retained observations therefore help subsequent questions in both modalities, although a seven-day stream cannot establish which categories would benefit consistently over longer periods.

B EXPLORATORY JOINT MEMORY AND POLICY UPDATES

Protocol. The joint condition starts from the same audiovisual agent and permits edits to five restricted modules: retrieval, acquisition, evidence organization, memory retention, and decisions. Correctness feedback is released only after all of a day's predictions have been saved. Six historical diagnosis tasks are replayed twice, and at most two candidates are proposed per day. Each candidate then undergoes code review and paired replay on up to eight historical validation tasks, which shared event anchors keep separate from the diagnosis tasks. The model, the tool interfaces, the access rules, and the per-question budgets all stay fixed.

With N_+ corrected answers, N_- regressions, and model-call cost C, the gate in Equation 6 is

$$g_d(\phi'_d; \phi_d) = \mathbf{1}[\mathrm{Valid}(\phi'_d) \wedge N_+ - N_- \geq 1 \wedge N_- \leq 1 \wedge C(\phi'_d) \leq 1.5C(\phi_d)]\,. \tag{7}$$

Validity requires a successful paired execution and activation of the edited module. Parent and candidate use identical memory during validation, while each experimental condition otherwise follows its own memory trajectory. A rejected policy edit still leaves source-verified memory additions in place.

Accepted updates. Three of the six update opportunities produced an accepted edit. Day 1 relaxed the requirement for supplemental media when the original frames already support a correction. Day 3 added cross-language retrieval guidance and raised the default retrieval count from five to six. Day 5 added temporal search guidance and raised retrieval to eight candidates whenever progress stalled. Each accepted edit fixed one of eight validation answers without causing regressions. The Day 2 candidate fixed two answers but broke two others, and Days 4 and 6 produced no accepted change. The joint condition retained 115 supplemental records, against 119 for memory self-improvement alone.

Results. Joint updating reaches 51.6%, versus 51.8% for memory self-improvement and 50.2% for EgoReView (Table 3). It beats memory refinement on Days 3, 4, and 6 and trails it on Days 2, 5, and 7, the overall difference amounting to one answer out of 500. Table 5 shows its stronger RelationMap and TaskMaster scores next to weaker EntityLog and HabitInsight scores. None of this establishes an additional overall benefit from policy adaptation.

Interpretation. Eight-task validation and six update opportunities say little about reusable personal policies. Recent-task selection and small validation sets may contribute to unstable acceptance decisions, but the experiment cannot identify the cause. A month- or year-long task stream with repeated demands, reliable feedback, and evaluation on later tasks would allow a more informative test of sustained adaptation, although more data alone guarantees no improvement. The main paper accordingly studies memory refinement under a fixed policy, and joint updates remain an exploratory extension.

Table 5: Complete category and modality comparison, including joint updates. Scores are accuracy (%).

Method	Ent.	Rel.	Event	Task	Habit	Visual	Audio	Overall
All questions: Days 1–7 ($N = 500$)								
Baseline (512 frames)	55.20	32.80	58.73	50.79	42.62	50.67	45.05	48.40
EgoReView	54.40	36.80	59.52	55.56	44.26	52.01	47.52	50.20
EgoReView + memory self-improvement	59.20	36.00	59.52	58.73	45.90	53.02	50.00	51.80
EgoReView + joint updates	56.00	38.40	59.52	60.32	44.26	52.01	50.99	51.60
Forward adaptation: Days 2–7 ($N = 398$)								
Baseline (512 frames)	56.07	33.67	57.27	47.73	38.46	51.41	42.95	48.24
EgoReView	55.14	37.76	57.27	52.27	41.03	52.21	45.64	49.75
EgoReView + memory self-improvement	60.75	36.73	57.27	56.82	43.59	53.41	48.99	51.76
EgoReView + joint updates	57.01	39.80	57.27	59.09	41.03	52.21	50.34	51.51

Ent.: EntityLog; Rel.: RelationMap; Event: EventRecall; Task: TaskMaster; Habit: HabitInsight. Visual and Audio denote questions without and with required audio. The five categories contain 125, 125, 126, 63, and 61 questions; the modality subsets contain 298 and 202. Day 1 is shared initialization. Joint updates combine memory and policy changes.

C Daily Category Breakdown

Table 6: Daily category accuracy (%) for all four conditions.

Method	Ent.	Rel.	Event	Task	Habit	Overall
Day 1 ($N = 102$)						
Baseline	50.00	29.63	68.75	57.89	50.00	49.02
EgoReView	50.00	33.33	75.00	63.16	50.00	51.96
Memory SI	50.00	33.33	75.00	63.16	50.00	51.96
Joint updates	50.00	33.33	75.00	63.16	50.00	51.96
Day 2 ($N = 59$)						
Baseline	50.00	60.00	57.14	50.00	60.00	55.93
EgoReView	42.86	55.00	71.43	75.00	60.00	57.63
Memory SI	50.00	60.00	57.14	62.50	70.00	59.32
Joint updates	35.71	65.00	57.14	62.50	60.00	55.93
Day 3 ($N = 85$)						
Baseline	57.14	33.33	61.90	69.23	33.33	51.76
EgoReView	47.62	41.67	57.14	69.23	50.00	51.76
Memory SI	52.38	41.67	47.62	76.92	50.00	51.76
Joint updates	57.14	37.50	57.14	84.62	50.00	55.29
Day 4 ($N = 84$)						
Baseline	56.25	29.63	70.00	58.33	22.22	47.62
EgoReView	56.25	37.04	70.00	58.33	22.22	50.00
Memory SI	62.50	37.04	75.00	66.67	22.22	53.57
Joint updates	62.50	40.74	70.00	66.67	33.33	54.76
Day 5 ($N = 75$)						
Baseline	62.07	7.14	52.63	0.00	28.57	41.33
EgoReView	65.52	7.14	52.63	16.67	28.57	44.00
Memory SI	68.97	7.14	52.63	16.67	28.57	45.33
Joint updates	58.62	7.14	52.63	16.67	14.29	40.00
Day 6 ($N = 76$)						
Baseline	53.85	25.00	58.62	20.00	50.00	48.68
EgoReView	57.69	33.33	55.17	0.00	50.00	48.68
Memory SI	65.38	16.67	62.07	20.00	50.00	52.63
Joint updates	65.38	33.33	65.52	20.00	50.00	56.58
Day 7 ($N = 19$)						
Baseline	0.00	100.00	35.71	*n/a*	33.33	36.84
EgoReView	0.00	100.00	42.86	*n/a*	33.33	42.11
Memory SI	0.00	100.00	42.86	*n/a*	33.33	42.11
Joint updates	0.00	100.00	28.57	*n/a*	33.33	31.58

Memory SI denotes EgoReView + memory self-improvement; Joint updates additionally edit policies. Each block uses the same questions; updates apply only on subsequent days. *n/a* denotes a category with no questions. Abbreviations follow Table 4.

D Implementation and Reproducibility Details

Table 7 lists the inference settings for the reported EgoMemReason configuration and for the fixed-policy EgoLifeQA agents. Limits are per question unless stated otherwise. Cached initial images fall outside the supplemental image budgets, and native video parts carry a separate duration budget. A tool call differs from a model call, since planning, query translation, media description, and answering can each require generation; EgoLifeQA caps their combined generation attempts at 36 including retries, and memory maintenance is accounted for separately.

Table 7: Inference and memory settings. Values denote caps unless marked as defaults. The native audiovisual limits apply only to EgoLifeQA; its still-image tools retain separate budgets.

Setting	EgoMemReason	EgoLifeQA
Model and initial perception		
Model	Gemini-3.8-Flash, low thinking effort	
Initial images / image size	512 / 384×384	512 / 384×384
Input tokens per request	900,000	900,000
Ordinary output tokens	8,192	8,192
Tool calls	8	8
Total model-call cap	No separate cap	36
Still-image acquisition and retrieval		
Supplemental image budget	48	48
Images per acquisition call	12	12
Retained supplemental images	12	12
Supplemental pixel budget	12,582,912	12,582,912
Still-image longest side	1,024 pixels	1,024 pixels
Frame-inspection / crop calls	6 / 6	6 / 6
Crop depth	2	2
Nearest-frame tolerance	2 seconds	2 seconds
Retrieval k (default; range)	5; 1–10	5; 1–10
Stop after no new sources	2 consecutive actions	3 consecutive actions
Temporal acquisition		
Dense visual sampling	8 FPS, 10-second window	—
Visual temporal-zoom calls	4	—
Visual watch / skim span	180 / 600 seconds	—
Native video FPS (default; range)	—	1; 0.1–4
Native video longest side	—	Default 512; 128–768
Native audiovisual window	—	90 seconds
Total audiovisual duration / parts	—	180 seconds / 12
Persistent memory refinement		
Candidate records per question	—	2
Verification checks per update	—	20
Update / publication schedule	—	Daily / next day

Supplemental image limits exclude the initial overview and the internally sampled frames of native video inputs. The eight-call tool budget is shared across tools; individual caps do not add to it. FPS controls visual sampling, while replay retains continuous source audio when present.

Provided captions are stored in SQLite with full-text lookup over Chinese character bigrams and Latin words, followed by lexical ranking, and person and time filters apply before ranking. Non-Chinese queries are converted into Chinese search keywords without answering the question. Retrieval returns five results by default and accepts one to ten; event search groups neighboring matches no more than 60 seconds apart into candidate episodes, which are not treated as verified event counts. Frames and clips carry their actual source timestamps, crops retain original coordinates and provenance, and nearest-frame substitution exposes the time actually sampled.

All initial images stay attached during acquisition and final answering. On EgoMemReason, decision calls keep the latest six tool outcomes and ten retrieved captions and attach at most 12 additional

images, with final evidence including full-frame context for crops where available. On EgoLifeQA the corresponding limits are five outcomes, 15 captions, and 16 source notes per decision; its fixed policy orders supplemental still images chronologically and attaches up to 12 of them alongside the acquired audiovisual segments. Output contracts require valid option letters and source identifiers. The visual answerer forms its candidate independently of the baseline answer, whereas the audiovisual adapter sees the baseline and accepts a change only with decisive supplemental media references. Both adapters are shared across the corresponding within-benchmark comparisons.

For memory self-improvement, question-independent media descriptions supply at most two candidate records per question. Exact duplicates are removed using hashes of the source reference, the interval, and the caption. The remaining candidates are sorted deterministically by a hash of the day and the record hash, and at most 20 undergo source verification after each of Days 1–6. The verifier sees the proposed detail and the original media, but neither the question, the options, the answer, nor any reward. Accepted captions preserve uncertainty and audible negation and are published at the next day's boundary. The 119 retained records are an observed outcome rather than a preset memory capacity. Joint policy-update settings and acceptance criteria appear separately in Appendix B.

The Gemini adapter uses JSON-constrained outputs at medium media resolution. Temperature, top-p, and top-k decoding controls are not sent to the API, so the inherited local temperature setting does not specify Gemini decoding. Ordinary outputs allow 8,192 tokens, Chinese query translation 4,096, and EgoMemReason timeline descriptions 16,384. Malformed structured responses allow two retries and transient HTTP failures three, subject to the remaining call budget where one is set. Runs use seed setting zero, with request seeds derived from the context and the call index. We retain cached baseline inputs, configurations, policy snapshots, the prompts in the implementation, source identifiers, and raw API responses for auditing. Resumed runs reuse saved outputs, and API generation is not assumed to be bitwise reproducible.