EGOCITE: CONTEXT-AUGMENTED INDEXING AND TIME-AWARE RETRIEVAL FOR LONG-HORIZON EGOCENTRIC MEMORY

Le Zhang **Ke Sun**
University of Michigan, Ann Arbor

ABSTRACT

Long-horizon egocentric memory transforms continuous first-person video and audio into a searchable record of past experiences. We demonstrate two bottlenecks in existing systems: indices built from context-poor captions are unreliable for agentic search, while retrieval ignores a question's temporal intent. To address both bottlenecks, we introduce `EgoCITE` (Egocentric Context-augmented Indexing and Time-aware Evidence retrieval), a long-horizon agentic memory framework for egocentric QA. `EgoCITE` comprises three components. `EgoScheme` uses local multimodal context to turn fragmentary video captions and speech transcripts into self-contained atomic memory indices. `EgoIndex` organizes complementary action, activity, utterance, and conversation representations into searchable multi-view memory indices at multiple granularities. `EgoRetrv` combines semantic search with question-conditioned temporal relevance scoring and curation of retrieved evidence. We evaluate `EgoCITE` on EgoLifeQA, EgoMem, and EgoR1-Bench in terms of answer accuracy and target-event retrieval alignment. `EgoCITE` improves accuracy over agentic memory baselines by at least 4.4–14.2% while achieving 36× lower cost than long-context LLM agents.

1 INTRODUCTION

Long-horizon egocentric memory aims to transform continuous first-person video and audio captured by wearable devices into a searchable record of daily life. Such memory enables an agentic assistant to answer questions about experiences that might otherwise be forgotten: whom a user met, where an object was left, what was discussed, or what typically occurs in a particular situation. This capability underpins emerging applications in memory augmentation, wearable augmented reality, and AI glasses that capture visual, audio, and social experiences (Zulfikar et al., 2024; Paruchuri et al., 2025; Kim et al., 2026; Wang et al., 2026a). Realizing this vision requires organizing dense multimodal observations into a searchable index whose entries remain interpretable outside their local context, retrieving the right evidence from a long history according to a question's temporal intent, and reasoning over that evidence to produce an accurate answer.

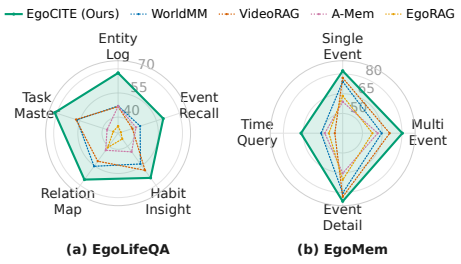

Figure 1: Per-category QA accuracy of agentic memory baselines on (a) EgoLifeQA (Yang et al., 2025b) and (b) EgoMem (Zheng et al., 2026) benchmarks.

Egocentric memory systems increasingly organize continuous experience into elaborate memory structures, such as knowledge graphs, semantic memories, and multi-scale hierarchies (Yeo et al., 2026; Sun et al., 2026; Yan et al., 2026). This mirrors agentic workflows in software engineering and computer use, where agents iteratively inspect and act on external workspaces (Yang et al., 2024; Xie et al., 2024). However, our investigation exposes a more fundamental bottleneck when this paradigm is applied to long-horizon egocentric memory. Memory entries are typically derived from short, independently generated video captions and speech transcripts, which often omit the local context

needed to interpret people, objects, places, and elliptical utterances. Once such context-poor content is indexed, neither a richer memory structure nor a stronger retrieval agent can reconstruct information that was never represented. We therefore introduce `EgoScheme`, which uses local multimodal context to transform fragmentary video captions and speech transcripts into self-contained atomic memory indices. `EgoIndex` then organizes disentangled and complementary action, activity, utterance, and conversation representations into searchable multi-view memory indices at multiple granularities.

Even with self-contained atomic memory indices, retrieval remains a second bottleneck. Existing systems typically let an agent search captions or entities by semantic similarity and rerank the returned candidates (Yang et al., 2025b; Yeo et al., 2026). Yet semantic similarity captures what an event is about, not when or how often it occurred. Questions over long-horizon egocentric memory frequently express temporal intent through qualifiers such as "last," "first," "usually," or "this morning." The most semantically similar event may therefore be neither the requested occurrence nor representative of a recurring habit. We address this limitation with `EgoRetrv`, a time-aware retrieval framework that combines iterative semantic search with question-conditioned temporal relevance scoring and curation of retrieved evidence. To separate targeted evidence discovery from cross-evidence reasoning, `EgoRetrv` adopts a dual-agent design: a drafting agent iteratively retrieves specific evidence using semantic and temporal queries, while a sampling agent reasons over the accumulated evidence in context to curate a coherent set aligned with the question's temporal intent.

We instantiate these ideas in `EgoCITE` (Fig. 3), a long-horizon agentic memory framework for egocentric QA that couples context-augmented indexing with time-aware retrieval. The design separates the responsibilities that existing systems entangle: local context is used at construction time to make evidence self-contained, while temporal and cross-event reasoning are applied at retrieval time. Our contributions are as follows:

- We identify two fundamental failures of current egocentric memory systems. (1) Fragmented and elliptical memory entries lack the context required for reliable indexing and agentic search. (2) Existing retrieval interfaces do not sufficiently integrate question context, particularly temporal intent, into memory search and selection.
- We introduce `EgoCITE`, which builds context-augmented, multi-view atomic memory indices and uses dual-agent, time-aware retrieval to select and curate retrieved atomic memory indices according to temporal intent.
- We evaluate `EgoCITE` on EgoLifeQA (Yang et al., 2025b), EgoMem (Zheng et al., 2026), and EgoR1-Bench (Tian et al., 2025) using answer accuracy and target-event retrieval alignment. `EgoCITE` improves accuracy over agentic memory baselines by at least 4.4–14.2% while achieving 36× lower cost than long-context LLM agents. Our project page and code are available at `https://egocite.github.io`.

2 Background and Related Work

Egocentric Lifelogging. Egocentric lifelogging captures continuous first-person experience to support memory recall and personal assistance. Ego4D (Grauman et al., 2022), EgoSchema (Mangalam et al., 2023), HD-EPIC (Perrett et al., 2025), and Nymeria (Ma et al., 2024) scaled egocentric video from short-clip action recognition to long-form temporal reasoning and multi-day multimodal recordings. Memoro (Zulfikar et al., 2024), EgoTrigger (Paruchuri et al., 2025), SpeechLess (Kim et al., 2026), and EgoSelf (Wang et al., 2026a) build wearable and AI glasses systems for memory logging across visual, audio, and interaction modalities. EgoLife (Yang et al., 2025b), LifeDialBench (Zheng et al., 2026), EgoMemReason (Wang et al., 2026b), and SuperMemory-VQA (Alam et al., 2026) offer multi-day video-speech data and QA benchmarks for life assistant evaluation. However, building memory from continuous multimodal sensor streams remains challenging: the streams are high-dimensional and cannot be indexed or queried directly. Handling them poorly results in information loss, retrieval errors, and weak long-horizon recall.

Agentic Memory. Agentic memory gives LLM agents persistent, queryable knowledge beyond a single context window. Prior work spans retrieval-augmented reasoning (Lewis et al., 2020; Yao et al., 2022; Asai et al., 2024) and long-term memory systems (Packer et al., 2023; Xu et al., 2026; Kang et al., 2025; Rasmussen et al., 2025). In multimodal memory, VideoRAG (Ren et al., 2026), AMEGO (Goletto et al., 2024), WorldMM (Yeo et al., 2026), EgoGraph (Sun et al., 2026), and AVA (Yan et al., 2026) enable long-context video QA through retrieval-augmented memory.

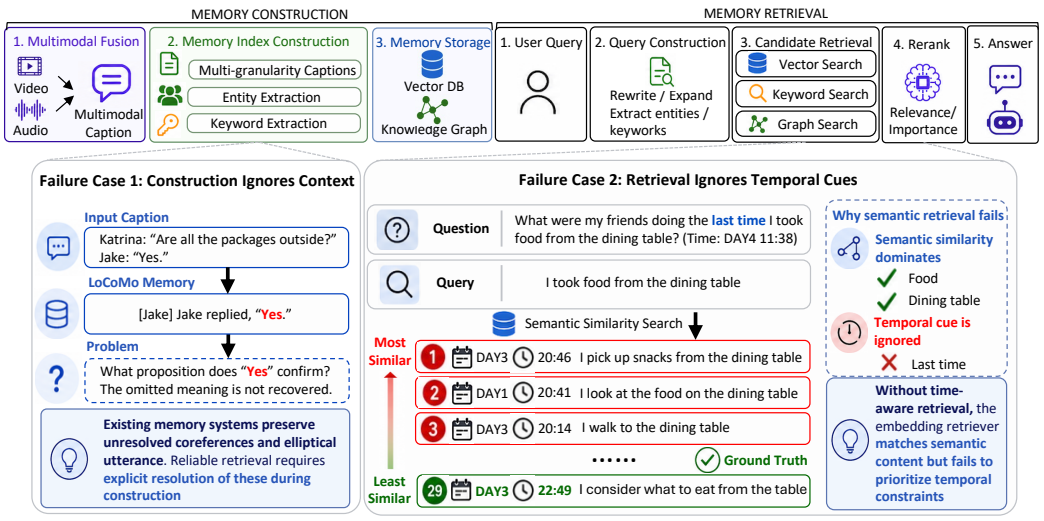

Figure 2: The workflow and failure cases of existing memory construction and retrieval approaches.

We summarize the workflow of existing multimodal memory systems in Fig. 2 through four key concepts. *(1) Memory entries.* During construction, multimodal signals (i.e., video and audio) are transcribed by multimodal LLMs into natural-language captions, which serve as memory entries. *(2) Memory indices.* Systems transform these entries into searchable indices through multi-granularity summarization (Yang et al., 2025b; Ren et al., 2026), entity extraction (Yeo et al., 2026), or keyword extraction (Xu et al., 2026). *(3) Memory structures.* The resulting indices are stored in structures such as vector databases (Yang et al., 2025b; Ren et al., 2026; Xu et al., 2026) or knowledge graphs (Yeo et al., 2026). *(4) Retrieved evidence.* At inference time, an LLM reformulates the user's question into search queries, retrieves candidate memory entries through keyword matching (Xu et al., 2026), vector search (Ren et al., 2026; Yang et al., 2025b), or graph traversal (Yeo et al., 2026), and reranks them by relevance. The highest-ranked entries and their associated multimodal signals then become retrieved evidence for response generation. In this work, we mainly investigate the role of high-quality memory indices in improving long-horizon egocentric memory question answering.

3 Motivations

We evaluate existing long-horizon egocentric memory systems (Yeo et al., 2026; Ren et al., 2026; Xu et al., 2026; Yang et al., 2025b) in Appx. A and identify two fundamental limitations.

Insight 1: Long-horizon egocentric agentic search is limited by what is indexed. Existing systems organize experience using increasingly elaborate memory structures, including vector databases, knowledge graphs, semantic memories, and visual memories (Yang et al., 2025b; Sun et al., 2026; Yeo et al., 2026). Yet their memory indices are typically built directly from short, independently generated memory entries, such as captions and speech transcripts. These entries often omit the local context needed to identify people, objects, actions, and elliptical utterances, making relevant experiences difficult for an agent to retrieve. Coreference resolution alone is insufficient because real-world conversations also contain ellipsis and omitted expressions (Aralikatte et al., 2019). For example, in failure case 1 of Fig. 2, Jake replies, "*yes*," but the isolated entry does not specify what he confirms. In our analysis of WorldMM (Yeo et al., 2026) and LoCoMo (Maharana et al., 2024), 10% of WorldMM's extracted entities contain unresolved pronouns, and 17% of LoCoMo memory entries contain unresolved verbatim quotes (Appx. A.1). *Context omitted from memory entries before indexing remains unavailable to downstream agentic search. Memory construction should therefore use local context to produce self-contained atomic memory indices.*

Insight 2: Retrieval must model temporal intent in addition to semantic relevance. A second limitation is that egocentric memory retrieval often lacks *time awareness*. Memory recall questions are inherently temporal (Gouveia et al., 2013), expressing recency ("last time"), habits ("usually"), or

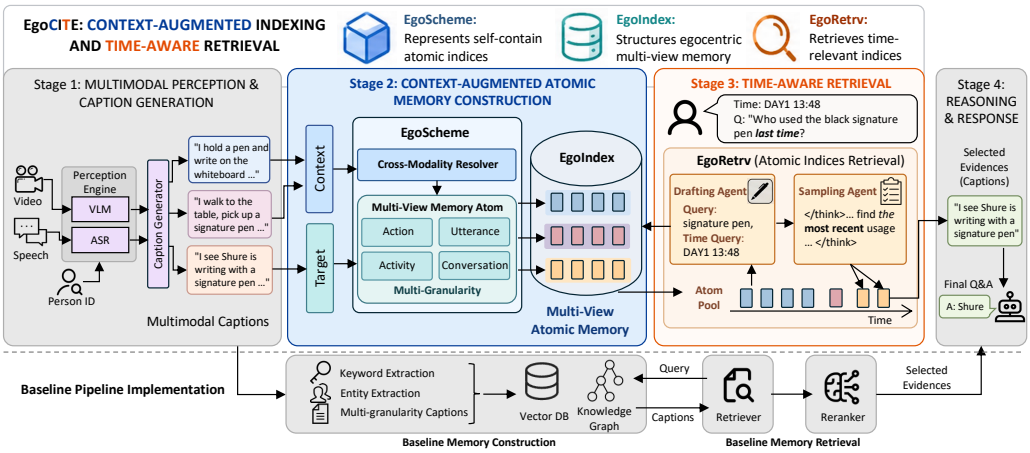

Figure 3: Overview of the pipeline: (1) Multimodal perception & caption generation, (2) Context-augmented memory indexing, (3) Time-aware atomic memory index retrieval, and (4) Reasoning & response.

explicit time ("yesterday afternoon"). More than 76% of EgoLifeQA (Yang et al., 2025b) and 48% of EgoMem (Zheng et al., 2026) questions contain temporal intent. Existing systems nonetheless rank by semantic similarity and treat time as passive metadata (Yang et al., 2025b; Yeo et al., 2026; Ren et al., 2026; Xu et al., 2026), so the top-k fills with semantically plausible but temporally wrong evidence. We show an example of this retrieval failure in Fig. 2, where semantically similar memories are retrieved while the temporally correct evidence is ranked much lower. This failure is reflected in EgoLifeQA, where accuracy on time-related questions is 9.8–10.3% lower for WorldMM (Yeo et al., 2026) and 13.3% lower for VideoRAG (Ren et al., 2026) than on time-unrelated questions (Appx. A.2). *We therefore argue that temporal reasoning should be an explicit component of retrieval, rather than a post hoc filter over semantically retrieved memories.*

Making retrieval time-aware is not as simple as adding timestamps to the memory query for two reasons. *(1) Sentence embeddings ignore temporal intent.* Embeddings are optimized for semantic similarity rather than temporal ordering, so appending temporal expressions does not reliably retrieve the correct evidence. *(2) Temporal expressions are often ambiguous.* Phrases such as "at noon" or "at breakfast" refer to approximate time ranges rather than exact timestamps, making hard temporal filtering prone to missing relevant memories (Sun et al., 2026).

4 METHODOLOGY

Together, these findings motivate `EgoCITE`, a four-stage pipeline for long-horizon egocentric memory (Fig. 3). Stages 1 and 4 follow the standard perception and response architecture of existing egocentric memory systems, whereas our contributions lie in Stages 2 and 3.

Stage 1: Multimodal Perception and Caption Generation. We convert egocentric video and audio into dense visual captions and speech transcripts, resolve speaker and participant identities, and fuse these outputs into multimodal captions that serve as raw memory entries for Stage 2 (see implementation details in Appx. B.1). **Stage 2: Context-Augmented Memory Indexing.** Each raw memory entry is augmented with a context window of neighboring entries. Within this window, `EgoScheme` resolves coreferences and ellipsis to produce self-contained atomic memory indices. `EgoIndex` organizes these indices into four complementary views—actions, activities, utterances, and conversations—forming a searchable multi-view memory index (Sec. 4.1). **Stage 3: Time-Aware Retrieval.** Given a user question, `EgoRetrv` retrieves atomic memory indices using two agents. A *drafting* agent performs multi-round retrieval over `EgoIndex` using semantic similarity and *time queries*, accumulating candidates in an *index pool*. A *sampling* agent then curates this pool with explicit awareness of temporal qualifiers such as "last", "first", and "usually" in the question (Sec. 4.2). **Stage 4: Reasoning and Response.** The response agent receives the curated atomic memory indices

View	Component	Specification	Example
Action	HOI	Require a named **person**, a specific **object**, and **direct physical handling**.	"**I fasten** the **shelf bracket** with **a screwdriver**."
	Gesture	Require a named **person** and a **meaningful expression**.	"**Jake laughs loudly**."
	Movement	Require a named **person** and a named **destination** that the person reaches.	"**Shure** enters the **living room**."
Utterance		Require a named **speaker**, a **speech act**, and **self-contained content with resolved pronouns and ellipsis**.	"**Jake tells** Tasha that **dinner is ready**." "**Jake confirms** that **all the packages are outside**."
Activity		Require an **immediate goal**, **specific participants, an object, and a location**.	"**Alice, Tasha, and I prepare dinner** in **the kitchen**."
Conversation		Require **participants** and **specific, coherent topic** across captions.	"**Jake and I** debate **which model to use for the demo video**."

Table 1: Key specifications and examples for `EgoScheme`.

and maps their timestamps back to multimodal captions. It then examines the selected captions and produces the final multiple-choice answer (see implementation details in Appx. B.5).

4.1 Context-Augmented Memory Indexing

Context-augmented memory indexing addresses the first limitation identified in Sec. 3: raw memory entries often lack the context required for reliable agentic search. We first define the multi-view, multi-granularity atomic memory indices exposed to the retrieval agent through `EgoIndex`, then describe how `EgoScheme` constructs these self-contained indices from context-poor memory entries.

Design 1: `EgoIndex`—Multi-View Egocentric Memory Indexing. Existing systems typically construct memory indices directly from captions generated by multimodal LLMs, yielding a single, undifferentiated view of each observation (Yang et al., 2025b; Ren et al., 2026). A single caption may combine short actions, extended activities, individual utterances, and broader conversations, even though a question may target only one interaction type or level of abstraction. `EgoIndex` therefore separates experience along two axes: physical behavior versus spoken interaction, and fine- versus coarse-grained semantics. Inspired by recent egocentric vision research (Ma et al., 2024; Yeo et al., 2026), this organization yields four complementary memory *views*: *actions*, *activities*, *utterances*, and *conversations*. Actions and activities represent physical behavior at fine and coarse granularities, respectively, whereas utterances and conversations represent spoken interaction at fine and coarse granularities. Actions describe short-term behaviors, including hand-object interactions, meaningful gestures, and movements. Activities capture higher-level, goal-oriented interactions. Utterances record individual speech acts with self-contained content, whereas conversations summarize broader topic exchanges among participants. Each view-specific representation is an *atomic memory index* derived from one or more raw memory entries. Formally, each atomic memory index m_i is associated with a timestamp τ_i inherited from its source memory entry or entries. We define `EgoIndex` as a collection of timestamped atomic memory indices,

$$\mathcal{M} = \bigcup_{v \in \mathcal{V}} \mathcal{M}_v = \{(\tau_i, m_i)\}_{i=1}^{|\mathcal{M}|}.$$

where the four views are $\mathcal{V} = \{\textsc{Action}, \textsc{Utterance}, \textsc{Activity}, \textsc{Conversation}\}$. `EgoIndex` offers two advantages: (1) *Multi-granularity retrieval.* Fine- and coarse-grained atomic memory indices support queries at different abstraction levels. (2) *Separation of interaction types.* Physical behavior and spoken interaction capture complementary aspects of an experience.

Design 2: `EgoScheme`—Context-Augmented Atomic Memory Index Construction. To construct self-contained atomic memory indices for `EgoIndex`, `EgoScheme` defines a structured specification for each view, as shown in Tab. 1. All four views follow two principles: (1) *Human-centered*: each atomic memory index is anchored to a specific person; and (2) *Decoupled*: each index captures one coherent behavior, activity, utterance, or conversational topic, preventing compounded semantics from confusing similarity-based retrieval. When the available context does not support a unique resolution, `EgoScheme` preserves the original expression rather than introducing unsupported details.

For (a) action, we divide user behavior into hand-object interactions (HOIs), gestures, and movements. An HOI specifies the person, object, and direct physical interaction. A gesture captures a meaningful expression. A movement specifies a user's movement toward a named destination. We demonstrate examples in Tab. 1. For (b) utterance, each atomic memory index specifies the speaker, speech act,

and self-contained content with resolved pronouns and ellipsis. For example, "Jake confirms that all the packages are outside" identifies Jake as the speaker, "confirms" as the speech act, and "all the packages are outside" as the complete propositional content. To extract (a) actions and (b) utterances, we set the context segment length to T ($T = 5$ min by default), such that the context of each caption C_t consists of all captions from the preceding T minutes, denoted by $[C_{t-T}, \ldots, C_t]$. For each current caption C_t, we prompt an LLM to generate structured atomic memory indices following the specifications above, conditioned on the entire context window. The extracted index m_i inherits timestamp $\tau_i = t$ from the source caption.

For (c) activity, each atomic memory index summarizes one immediate goal grounded in its participants, object, and location. For (d) conversation, each atomic memory index names all participants and captures one coherent topic across captions. We demonstrate examples in Tab. 1. To extract (c) and (d), we partition the multimodal captions into disjoint T'-minute windows, denoted by $[C_{t-T'}, \ldots, C_t]$, with $T' = 30$ minutes by default following WorldMM (Yeo et al., 2026). We prompt an LLM to segment each window into activity or conversation atomic memory indices m_i following `EgoScheme`, with each index assigned a specific temporal interval $[t_i^{\text{start}}, t_i^{\text{end}}]$. The extracted index m_i inherits the start time of its source caption segment, i.e., $\tau_i = t_i^{\text{start}}$. To ensure that each memory index captures a single coherent behavior, we further apply a disentangling procedure that separates unrelated content joined by connectives such as "and" or "while." To reconcile indices across adjacent windows, we identify temporally continuous memory indices using sentence-embedding similarity and merge them with an additional LLM call (see Appx. B.2 for implementation details and Appx. F for examples).

4.2 Time-Aware Atomic Memory Index Retrieval

Motivated by Insight 2 in Sec. 3, `EgoRetrv` makes a question's temporal intent an explicit retrieval signal. It combines question-conditioned temporal relevance scoring with a dual-agent design: (1) A *drafting* agent issues multi-round semantic, view, and *time* queries to `EgoIndex` and progressively accumulates specific atomic memory indices in a persistent working memory named the *index pool*, prioritizing retrieval recall. (2) A *sampling* agent reasons over the accumulated indices in context and curates a coherent evidence set aligned with the question's temporal intent, improving retrieval precision by selecting temporally relevant indices and removing redundant ones.

Design 3: Drafting agent with time query. At each retrieval round, given a question Q, the drafting agent generates a structured query set $\mathcal{Q} = \{Q_s, Q_v, Q_t\}$, where Q_s is a semantic query, Q_v specifies the memory view to retrieve, and Q_t is an optional *temporal query*. Using a sentence-embedding model f, we compute the semantic relevance of each candidate memory index $m_i \in \mathcal{M}_{Q_v}$ as $S_i = \text{sim}\,(f(Q_s), f(m_i))$.

The temporal query $Q_t = [t_{\text{start}}, t_{\text{end}}]$ represents the drafting agent's estimated time range for the queried event, inferred from the user question using commonsense temporal knowledge and atomic memory indices accumulated in prior rounds. A point estimate is represented by $t_{\text{start}} = t_{\text{end}}$. For example, a "last time" qualifier with a reference time of `DAY1 12:00` may yield $Q_t =$ [`DAY1 12:00`, `DAY1 12:00`], whereas "this morning" may yield $Q_t =$ [`DAY1 06:00`, `DAY1 12:00`]. The agent may update Q_t across retrieval rounds as the retrieved atomic memory indices refine its temporal belief. We derive the temporal relevance score R_i directly from the temporal query $Q_t = [t_{\text{start}}, t_{\text{end}}]$. After mapping all timestamps onto a continuous timeline with hour as the unit, we compute

$$R_i = \begin{cases} 1, & t_{\text{start}} \le \tau_i \le t_{\text{end}}, \\ \lambda^{t_{\text{start}} - \tau_i}, & \tau_i < t_{\text{start}}, \\ \lambda^{\tau_i - t_{\text{end}}}, & \tau_i > t_{\text{end}}. \end{cases}, \tag{1}$$

where $0 < \lambda < 1$ controls the time-decay factor. We select the default $\lambda = 0.99$ based on ablation study in Sec. 5. Candidates within Q_t receive $R_i = 1$, whereas candidates outside the interval are exponentially downweighted according to their hourly distance from the nearest boundary of Q_t. If no temporal query Q_t is provided, we set $R_i = 1$ by default. The final retrieval score combines semantic and temporal relevance multiplicatively:

$$\text{score}(m_i \mid Q_s, Q_v, Q_t) = S_i R_i.$$

The top-k timestamped atomic memory indices are selected according to the final score:

$$\mathcal{K} = \text{TopK}_{(\tau_i, m_i) \in \mathcal{M}_{Q_v}} \text{score}(m_i \mid Q_s, Q_v, Q_t),$$

where $\mathcal{K} \in \mathcal{M}_{Q_v}$ contains the retrieved `(timestamp, index)` pairs from the specified view. The index pool $\mathcal{P}$ is updated as $\mathcal{P} \leftarrow \mathcal{P} \cup \mathcal{K}$. Implementation details are presented in Appx. B.3 and F.

Together, these mechanisms address the two challenges identified in Insight 2. (1) Modeling temporal relevance R_i outside the embedding space avoids the temporal insensitivity of sentence embeddings. (2) Soft decay via λ accommodates ambiguous temporal expressions by assigning nearby atomic memory indices non-zero scores. In contrast, hard temporal cutoffs (Sun et al., 2026) discard all memory indices outside a fixed window and are brittle to reasoning errors and linguistic ambiguity.

Design 4: Sampling agent with time-aware curation. The time query biases retrieval toward a predicted window but lacks a global view of the collected atomic memory indices. We therefore decouple retrieval and curation: the drafting agent collects high-recall candidates, while the sampling agent reviews them and selects the temporally correct atomic memory indices based on time cues in the question statement. In practice, the sampling agent reads the timestamped atomic memory indices from the *index pool* $\mathcal{P}$ and orders them chronologically. It then applies the curation $\mathcal{C}(Q, \mathcal{P})$ to select the most relevant indices based on the temporal cues in the question Q. For example, for recency-oriented questions containing qualifiers such as "last" or "most recent," the sampling agent prioritizes atomic memory indices with the latest timestamps while retaining related indices needed to distinguish among answer choices. For habitual cues such as "usually" or "habit," it selects a temporally diverse set of relevant atomic memory indices spanning multiple days to capture recurring behavior rather than a single event (see implementation details in Appx. B.4). This extends the recency-only selection algorithm in EgoRAG (Yang et al., 2025b) to questions that require multiple atomic memory indices to answer, such as habits, coreferences, and specific time ranges. This also outperforms semantic memory in WorldMM (Yeo et al., 2026), where habitual questions are poorly summarized during construction time due to imperfect memory consolidations (see Sec. 5).

5 Experiments

Benchmarks. We evaluate `EgoCITE` on three egocentric memory benchmarks on EgoLife (Yang et al., 2025b), including EgoLifeQA (Yang et al., 2025b), EgoMem (Zheng et al., 2026), and EgoR1-Bench (Tian et al., 2025). For benchmarks, we use question statements, answer choices, and question time as the input context, and use ground-truth answers and ground-truth timestamps that requires to answer the question for evaluation (See details in Appx. C).

Baselines. We compare `EgoCITE` with long-context LLM agents Gemini-3.1-Pro agent (Team et al., 2023) and GPT-5.4 agent (OpenAI, 2026a), agentic memory baselines, including EgoRAG (Yang et al., 2025b), A-MEM (Xu et al., 2026), VideoRAG (Ren et al., 2026), and WorldMM (Yeo et al., 2026). Most baselines are evaluated using GPT-5.4 for retrieval agents and WorldMM is evaluated with Qwen3.6-27B-FP8 (Qwen Team, 2026) additionally. We evaluate `EgoCITE` with both GPT-5.4 and Qwen3.6-27B-FP8 at default 5-round retrieval and 15 memory indices to align with WorldMM. Additional details on dataset preprocessing, metrics, baseline implementations, model configurations, and hardware are provided in Appx. B and C.

Metrics. We use multiple-choice accuracy as the main metric. We use input token numbers, retrieval latency, and normalized cost as efficiency metrics. *Normalized Cost* measures the normalized API token cost of GPT-5.4 (OpenAI, 2026b). Short- and long-context models use normalized input/output costs of $1\times/6\times$ and $2\times/9\times$, respectively. `EgoCITE-GPT` and other agentic memory baselines use short-context pricing and the long-context GPT agent uses long-context pricing. The normalized cost is computed as the weighted sum of input and output token costs (details in Appx. C.2).

Main Results. We report the results in Tab. 2. `EgoCITE-GPT` is the strongest baseline, outperforming the long-context LLM-agent baselines by 3.6–8.9% on EgoLifeQA, 0.9–4.1% on EgoMem, and 2.6–4.7% on EgoR1-Bench. Compared with agentic memory baselines, it achieves average gains of at least 14.2%, 4.4%, and 9.0% on the three benchmarks, respectively. `EgoCITE-Qwen` trails `EgoCITE-GPT` by 1–3% average accuracy while still outperforming all baselines. Long-context LLM agents consistently outperform existing agentic memory systems. Nevertheless, `EgoCITE` surpasses these long-context LLM agents while requiring $23\times$ fewer input tokens, demonstrating that accurate long-term memory retrieval does not require million-token contexts.

Retrieval Hit Rate. We report the retrieval hit rate over the agent's retrieved memory indices, with a maximum retrieval budget of 15 memory indices for all baselines before question answering (Table 3). Both `EgoCITE` implementations substantially outperform existing retrieval-based memory methods across all three benchmarks. `EgoCITE-GPT` achieves the highest hit rates of 49.6%, 89.6%, and 62.7%

Table 2: Baseline accuracy (%) and number of input tokens on three benchmarks. Bold, underlined, and italicized values denote the best, second-best, and third-best results, respectively.

Method	Model	Input Token	EgoLifeQA (Yang et al., 2025b)						EgoMem (Zheng et al., 2026)					EgoR1-Bench (Tian et al., 2025)
			Ent.	Evt.	Hab.	Rel.	Task	Avg.	Sgl.	Mul.	Det.	Time	Avg.	
GPT-5.4 agent (OpenAI, 2026a)	–	783k	56.2	48.1	63.9	57.3	*69.7*	54.9	80.6	**83.8**	84.6	49.5	75.1	71.0
Gemini-3.1-Pro agent (Team et al., 2023)	–	804k	*61.1*	*56.7*	*59.8*	*62.8*	70.1	*60.2*	*81.0*	81.2	81.7	**68.5**	*78.3*	*72.7*
EgoRAG (Yang et al., 2025b)	GPT	16k	34.4	31.8	34.4	41.0	33.2	34.3	63.3	57.6	70.4	45.5	59.5	49.5
A-MEM (Xu et al., 2026)	GPT	6k	46.7	41.7	44.3	43.1	37.2	42.9	58.9	61.6	65.4	48.6	58.8	52.0
VideoRAG (Ren et al., 2026)	GPT	21k	46.5	39.3	58.5	51.5	57.5	46.5	77.0	70.7	83.3	41.0	68.6	64.7
WorldMM-Qwen (Yeo et al., 2026)	Qwen	87k	46.1	41.3	49.5	49.2	54.6	46.4	77.8	77.3	*84.6*	58.1	74.8	66.3
WorldMM-GPT (Yeo et al., 2026)	GPT	134k	46.9	44.4	53.6	55.4	56.9	49.6	74.2	64.6	80.4	51.4	68.1	64.0
EgoCITE-Qwen (Ours)	Qwen	34k	62.7	**59.5**	58.8	63.6	67.0	61.5	**83.5**	78.2	**88.3**	67.6	**79.9**	73.0
EgoCITE-GPT (Ours)	GPT	32k	**67.4**	**59.5**	**64.3**	**65.6**	**71.4**	**63.8**	82.3	*80.3*	86.7	*66.7*	79.2	**75.3**

Table 3: Retrieval memory index hit rate (%). Green color for improvements compared to the best baseline.

Method	EgoLifeQA	EgoMem	EgoR1
EgoRAG (Yang et al., 2025b)	4.3	41.1	7.6
A-Mem (Xu et al., 2026)	15.4	41.4	24.7
VideoRAG (Ren et al., 2026)	2.0	17.4	6.0
WorldMM-Qwen (Yeo et al., 2026)	31.0	*72.6*	38.3
WorldMM-GPT (Yeo et al., 2026)	*34.5*	66.9	*40.0*
EgoCITE-Qwen (Ours)	43.4(+8.9)	86.8(+14.2)	59.0(+19.0)
EgoCITE-GPT (Ours)	**49.6**(+15.1)	**89.6**(+17.0)	**62.7**(+22.7)

Table 4: Retrieval-round ablation for EgoCITE-GPT on EgoLifeQA.

Metric	1-Round	3-Round	5-Round
Hit rate (%)	36.5	48.6	49.6
Average accuracy (%)	58.7	62.6	63.8
EntityLog	61.9	66.1	67.4
EventRecall	54.7	58.6	59.5
HabitInsight	60.1	60.1	64.3
RelationMap	62.6	65.4	65.6
TaskMaster	61.0	71.0	71.4

on EgoLifeQA, EgoMem, and EgoR1, respectively, improving over the best baseline by up to 15.1%, 17.0%, and 22.7%. These results indicate that EgoCITE's memory representation and time-aware retrieval enable substantially more accurate retrieval, regardless of the underlying model's capability.

Temporal-Aware Questions. We demonstates the accuracy of EgoLifeQA with temporal-aware questions (see Appx. C.1) in Tab. 9. EgoCITE-GPT achieves the highest accuracy at 62.6%, followed by EgoCITE-Qwen at 60.9%. Both variants outperform the strongest long-context LLM baseline Gemini-3.1-Pro agent by 3.9% and 2.2%. Compared with the strongest agentic memory WorldMM-GPT, they improve accuracy by 15.5% and 13.8%. These results demonstrate that EgoCITE more effectively captures temporal intent than both long-context LLM agents and existing agentic memory baselines.

Method	Accuracy (%)
EgoRAG (Yang et al., 2025b)	34.8
A-Mem (Xu et al., 2026)	42.2
VideoRAG (Ren et al., 2026)	43.4
WorldMM-Qwen (Yeo et al., 2026)	43.8
WorldMM-GPT (Yeo et al., 2026)	47.1
GPT-5.4 agent (OpenAI, 2026a)	51.6
Gemini-3.1-Pro agent (Team et al., 2023)	*58.7*
EgoCITE-Qwen (Ours)	60.9
EgoCITE-GPT (Ours)	**62.6**

Table 5: Time-aware questions.

Habits and Multiple Evidences. Habitual and multi-evidence questions require retrieving information across multiple timestamps. We report the results on habitual (Hab.) questions in EgoLifeQA and multi-evidence (Mul.) questions in EgoMem in Tab. 2. EgoCITE-GPT ranks first and third behind the long-context GPT-5.4 agent. It outperforms agentic memory baselines by 5.8–29.9% and 3.0–22.7% on two categories, while requiring more than 24× fewer input tokens than long-context LLM agents. These results demonstrate that EgoRetrv effectively retrieves information across long temporal horizons without million-token contexts.

Memory Scaling. Fig. 4 reports cumulative accuracy on EgoLifeQA as the memory horizon increases from DAY1 to DAY7. EgoCITE achieves the highest accuracy from DAY3 onward and shows only a 4.8% decline from DAY1 to DAY7, compared with an 11.9% drop for GPT-5.4 agent. Gemini-3.1-Pro agent, WorldMM-GPT, and VideoRAG maintain lower accuracy throughout. These results show that EgoCITE's structured multi-view memory and time-aware retrieval scale more effectively than both long-context reasoning and existing retrieval-based approaches.

Retrieval Rounds. We ablate retrieval depth in EgoCITE-GPT using 1-, 3-, and 5-round variants. Increasing retrieval depth improves hit rate and QA accuracy, with gains of 13.1% and 5.1% from 1

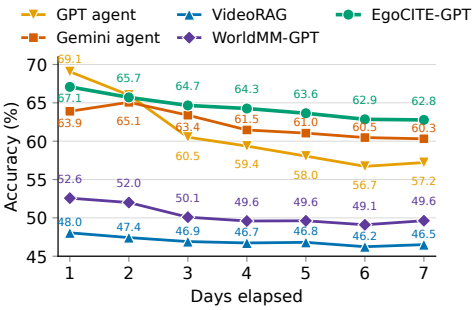

Figure 4: Memory scaling: EgoLifeQA accuracy over increasing memory horizon. Differences are compared to DAY1.

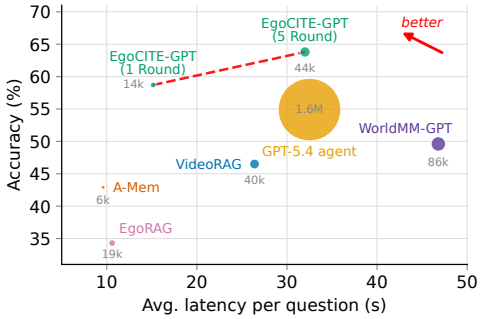

Figure 5: EgoLifeQA accuracy, latency, and normalized cost trade-offs. Gray annotations and marker size indicate normalized cost.

to 5 rounds (Tab. 4). Most gains are achieved within the first three rounds and later rounds provide only modest improvements, demonstrating the benefit of progressive multi-round retrieval.

Retrieval Efficiency. Since existing baselines differ in retrieval workflows, memory representations, and modalities, we compare their trade-offs among accuracy, retrieval latency, and normalized cost. 5-round variant achieves the highest accuracy with 2× lower normalized cost and 15s lower latency than WorldMM, thanks to atomic memory index retrieval that reduces retrieved context by 5.7×. 5-round variant also outperforms GPT-5.4 agent by 8.9% accuracy and 36× cost savings with comparable latency. Five-round retrieval improves accuracy by 5.1% over one round, with 2× higher latency and 3.2× higher normalized cost. Overall, `EgoCITE` is Pareto-optimal in accuracy, latency, and cost.

Time-Decay Factor. We ablate the temporal decay factor λ of `EgoCITE-GPT` on a 100-question validation subset of EgoR1-Bench in Tab. 11. Performance improves as λ increases from 0.97 to 0.99, peaking at 81.0% accuracy and 67.0% hit rate. Increasing λ further to 0.995 reduces both metrics, indicating that $\lambda = 0.99$ best balances semantic similarity and temporal relevance.

Component Ablation. We perform component ablations on `EgoCITE-GPT` and report the results on EgoR1-Bench in Tab. 10. We incrementally add `EgoCITE` components on EgoRAG-like caption RAG. Starting from caption RAG, adding `EgoIndex`, `EgoScheme`, and `EgoRetrv` progressively improves accuracy from 61.3% to 69.7%, 71.3%, and 75.3%, and hit rate from 40.7% to 54.3%, 59.3%, and 62.7%, demonstrating the contribution of each component.

Video Caption. Following WorldMM (Yeo et al., 2026), we use the human-annotated dense captions and speech transcripts from EgoLife as the memory source for fair comparison with baselines. For realistic evaluation, we replace dense captions with narrations generated by visual-language model (VLM) Gemini-3-Flash (Team et al., 2023) and Gemma-4-31B (Team et al., 2026) (Appx. B.1). We use EgoLifeQA as benchmark and Qwen3.6-27B-FP8 as retrieval agent. Results show Gemini-3-Flash and Gemma-4-31B achieve 60.3% and 57.1% accuracy, only 1.2% and 4.4% lower than human-annotated dense captions (see details in Sec. D). These results demonstrate that `EgoCITE` generalizes well to realistic VLM-generated narrations while maintaining competitive long-term memory performance.

6 Conclusion

We presented `EgoCITE`, a memory system for long-horizon egocentric life assistants. The system is built around a simple premise: a memory index is only as useful as the entries it organizes and the temporal reasoning used to retrieve evidence from them. `EgoScheme` converts local multimodal context into self-contained atomic memory indices, `EgoIndex` organizes complementary views without requiring global entity consolidation, and `EgoRetrv` combines soft temporal scoring with time-aware curation of retrieved evidence. Across the three-benchmark evaluation protocol, this formulation separates memory construction, evidence retrieval, and answer reasoning into inspectable stages. Future work should test the system under noisier perception, open-world identity resolution, and longer personal recordings.

REFERENCES

Samiul Alam et al. Supermemory-vqa: An egocentric visual question-answering benchmark for long-horizon memory. *arXiv preprint arXiv:2606.00825*, 2026.

Gabor Angeli et al. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 344–354, 2015.

Rahul Aralikatte et al. Ellipsis resolution as question answering: An evaluation. *arXiv preprint arXiv:1908.11141*, 2019.

Akari Asai et al. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

Zhoujun Cheng et al. Batch prompting: Efficient inference with large language model apis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 792–810, 2023.

Matthijs Douze et al. The faiss library. *IEEE Transactions on Big Data*, 2025.

Gabriele Goletto et al. Amego: Active memory from long egocentric videos. In *European Conference on Computer Vision*, pp. 92–110. Springer, 2024.

Rúben Gouveia et al. Footprint tracker: supporting diary studies with lifelogging. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 2921–2930, 2013.

Kristen Grauman 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.

Jiazheng Kang et al. Memory os of ai agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25972–25981, 2025.

Yoonsang Kim et al. Speechless: Micro-utterance with personalized spatial memory-aware assistant in everyday augmented reality. In *2026 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 217–227. IEEE, 2026.

Woosuk Kwon et al. Efficient memory management for large language model serving with page-dattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Lingni Ma et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pp. 445–465. Springer, 2024.

Adyasha Maharana et al. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024.

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

OpenAI. GPT-5.4 Thinking System Card. Technical report, OpenAI, March 2026a. URL https://deploymentsafety.openai.com/gpt-5-4-thinking/gpt-5-4-thinking.pdf. Accessed: 2026-07-07.

OpenAI. Openai api pricing. https://developers.openai.com/api/docs/pricing, 2026b. Accessed: 2026-07-27.

Charles Packer et al. Memgpt: towards llms as operating systems. 2023.

Akshay Paruchuri et al. Egotrigger: Toward audio-driven image capture for human memory enhancement in all-day energy-efficient smart glasses. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

Toby Perrett et al. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23901–23913, 2025.

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

Preston Rasmussen et al. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025.

Xubin Ren et al. 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.

Shitong Sun et al. Egograph: Temporal knowledge graph for egocentric video understanding. *arXiv preprint arXiv:2602.23709*, 2026.

Gemini Team et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Shulin Tian et al. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2506.13654*, 2025.

Yanshuo Wang et al. Egoself: From memory to personalized egocentric assistant. *arXiv preprint arXiv:2604.19564*, 2026a.

Ziyang Wang et al. Egomemreason: A memory-driven reasoning benchmark for long-horizon egocentric video understanding. *arXiv preprint arXiv:2605.09874*, 2026b.

Tianbao Xie et al. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, 2024.

Wujiang Xu et al. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026.

Yuxuan Yan et al. {AVA}: Towards agentic video analytics with vision language models. In *23rd USENIX Symposium on Networked Systems Design and Implementation (NSDI 26)*, pp. 1939–1957, 2026.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Jingkang Yang et al. Egolife: Towards egocentric life assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28885–28900, 2025b.

John Yang et al. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems*, 2024.

Shunyu Yao et al. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Woongyeong Yeo et al. 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.

Jianjie Zheng et al. Evaluating memory capability in continuous lifelog scenario. *arXiv preprint arXiv:2604.11182*, 2026.

Xu Zheng et al. Mllms are deeply affected by modality bias. *arXiv preprint arXiv:2505.18657*, 2025.

Wazeer Deen Zulfikar et al. Memoro: Using large language models to realize a concise interface for real-time memory augmentation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2024.

Table 6: Examples of unresolved memory produced by existing agentic memory baselines.

Baseline	Input Captions (excerpt)	Memory Entry (Triple / Observation)
WorldMM	"Come on, everyone mark **it**."	[I | says to Katrina, Alice, Tasha | Come on, everyone mark **it**]
	"Keep **this** box."	[I | say | Keep **this** box]
	"**There** are a total of six computers."	[I | states to Shure | **there** are a total of six computers]
	"This thing, you can put **these** down."	[I | say | This thing, you can put **these**]
LoCoMo	Lucia: "Is it a little mouse?" Jake: "What is this?"	[Jake] Jake asked "**What is this?**"
	Shure: "Great Northeast."	[Shure] Shure said "**Great Northeast.**"
	Jake: "Or not, haha."	[Jake] Jake said "**Or not, haha.**"
	Katrina: "Are all the packages out?" Jake: "Yes."	[Jake] Jake replied "**Yes.**"
	Jake: "Or sit here."	[Jake] Jake said "**Or sit here.**"

A Baseline Failure Case Study

A.1 Failure in Memory Construction

We ask whether existing memory representations produce self-contained and semantically complete memory indices. We analyze the memory indices constructed by two representative systems on EgoLife (Yang et al., 2025b): WorldMM (Yeo et al., 2026), which converts captions into OpenIE (Angeli et al., 2015) `(subject, predicate, object)` triples, and LoCoMo (Maharana et al., 2024), which extracts speaker-attributed observation facts. We hypothesize that neither representation produces context-independent memories: WorldMM leaves referring expressions unresolved, while LoCoMo preserves utterances verbatim. To quantify this, we flag a WorldMM triple if its subject or object contains unresolved pronouns and a LoCoMo observation if it contains a quoted span, indicating that the original utterance is copied rather than resolved. Across 92,131 WorldMM triples and 254,997 LoCoMo observations, 10.1% of WorldMM triples contain unresolved references and 17.4% of LoCoMo observations contain verbatim quotes. We show examples in Tab. 6. These results show that a substantial portion of existing memory indices are not self-contained, motivating a memory representation that explicitly resolves both coreference and ellipsis during memory construction.

A.2 Failure in Time-Aware Retrieval

We ask whether existing retrieval-based memory systems handle time-aware questions as effectively as time-unaware questions. To answer this, we construct a time-aware subset of EgoLifeQA (Appx. C.1) and evaluate two representative retrieval-based baselines, WorldMM (Yeo et al., 2026) and VideoRAG (Ren et al., 2026), using GPT-5.4 as the retrieval agent. Neither method explicitly models temporal intent during retrieval, instead relying primarily on semantic similarity. On the corresponding time-unaware questions, WorldMM and VideoRAG achieve 56.9% and 56.7% accuracy, respectively. However, their performance drops to 47.1% (-9.8%) and 43.4% (-13.3%) on time-aware questions. This substantial performance gap indicates that semantic retrieval alone is insufficient for long-horizon egocentric memory retrieval. We present representative failure cases in Fig. 7.

B Implementation Details

B.1 Perception and Caption

We primarily use the the human-annotated dense captions and speech transcripts as the source of memory in EgoLife (Yang et al., 2025b) for fair baseline comparisons on agentic memory performance, following WorldMM (Yeo et al., 2026). For realistic evaluations, we also deploy VLMs as narrator to replace the dense captions as the source of video narrations as follows.

Question ID	Question Statement	Time Cue
A1_JAKE Q104	When was the *last time* cherries were mentioned?	"last"
A1_JAKE Q1	Who used the screwdriver *first*?	"first"
A1_JAKE Q90	Where is the place I will *never* go to?	"never"
A3_TASHA Q903	Who mic'd the chorus *last night*?	"last night"
A3_TASHA Q1091	What did Lucia hand me *last time*?	"last"
A3_TASHA Q479	Who do I *usually* sit with when eating?	"usually"
A4_LUCIA Q432	Who mentioned AI *last time*?	"last"
A4_LUCIA Q256	When did we *first* mention juice?	"first"
A4_LUCIA Q930	Who bit the bullet *yesterday morning*?	"yesterday"
A5_KATRINA Q321	What did I *last* put on the shelf?	"last"
A5_KATRINA Q111	Who *usually* participates when we select products?	"usually"
A5_KATRINA Q1156	How *often* do I clean my room every day?	"often"
A6_SHURE Q170	When was the *last time* I danced?	"last"

Table 7: Time-aware questions answered incorrectly by both WorldMM-GPT and VideoRAG.

Person ID Annotation Example: A1_JAKE DAY1 11AM–5PM

```
# Camera wearer (I)
I am Jake, the camera wearer.

# Other people who may appear
Katrina - pink long hair, white short-sleeve shirt, dark blue pants, yellow shoes
Alice - brown short hair, white long-sleeve t-shirt, grey skirt, grey shoes
Tasha - black long hair in a ponytail, black long-sleeve clothing, dark blue jeans
Lucia - long black hair, blue dress, blue shoes
Shure - male, short curly black hair, pink or orange short-sleeve shirt, brown pants, colorful shoes
```

Figure 6: Human expert annotated person ID example.

Person ID. We assume person identification is prior knowledge, since personal relationships are relatively stable and closed-set in real-world settings. This prior knowledge includes the mapping among a person's visual appearance, voice identity, and name. In the EgoLife dataset (Yang et al., 2025b), visual identity is difficult to obtain because *faces are redacted in this dataset*, in contrast to EgoGPT and EgoButler (Yang et al., 2025b), which was developed by the dataset collectors with access to the raw videos. We therefore manually annotate each person's visual appearance, including clothing colors and styles, hairstyles, and other visual cues, across different recording scenes over the seven-day recordings (See example in Fig. 6). We generate identity files for each recording scene to map visual descriptions to person names. These identity files are used during VLM captioning to match people to their names. Empirically, we evaluate both proprietary VLM Gemini-3-Flash (Team et al., 2023) and open-weight VLM Gemma-4-31B (Team et al., 2026).

Video Caption. We first process 15-second, 1 FPS video clips independently from speech transcripts. Empirically, generating captions from joint video–speech inputs introduces hallucinations in objects, actions, and person identities, which we attribute to multimodal LLMs' tendency to prioritize textual over visual information (Zheng et al., 2025). The resulting visual captions are then aligned with speech transcripts to construct multimodal captions, following WorldMM (Yeo et al., 2026).

B.2 Index Memory Construction

Action & Utterance. Given the generated 5-minute context window, we resolve coreference and ellipsis following EgoScheme to construct action and utterance atomic memory indices from multimodal captions. Specifically, the context-resolution LLM receives the current caption together with its preceding 5-minute context, enabling it to rewrite each memory index into a self-contained and semantically complete representation. The LLM then extracts structured action and utterance indices according to the prompt in Fig. 16. Action indices capture hand-object interactions, large movements, and meaningful gestures. Utterance indices are normalized into explicit speech acts or statements with resolved speakers, referents, and topics. Compared with the original captions,

Component	Specification	Example
Category	Extract **only** four event types: (A) Hand–Object Interaction, (B) Big Displacement, (C) Utterance/Speech, (D) Meaningful Gesture. Discard all other events.	"I fasten the shelf bracket with a screwdriver." "I walk into the kitchen." "Jake asks Tasha about the dinner menu." "Jake laughs out loudly."
Atomicity	Extract exactly one entry per distinct action. Never combine actions connected by *and*, *while*, *then*, or *as*; split them into separate entries.	**Good:** "I'm working on computer to edit video." "Jake says weather will be bad tomorrow." **Bad:** "I'm working on computer while listening to Jake talking about weather."
Action Format	Each entry should contain 8–20 words in the form *subject + verb + object*. Use "I" for the camera wearer and a real name for everyone else. Preserve named objects and places verbatim.	"Shure hands me the screwdriver at desk." "I message Lucia on my phone." "I fasten the shelf bracket with a screwdriver."
Entity Resolution	Resolve every pronoun and generic coreference to an explicit person, object, or place. Discard an entry only if the referent is truly unresolvable.	**Good:** "Jake picks up the bottle on table." **Bad:** "He picks it up."
Speech Normalization	Convert dialogue into a speech act with a complete, self-contained topic expressed using a WHAT or THAT clause.	**Good:** "Jake tells Tasha that dinner is ready." "I ask Alice about the shelf height." **Bad:** "Jake says OK." "I ask Alice about it."
Motion Filtering	Discard micro-gestures and undirected motion. Keep movement only when it reaches a named destination.	**Good:** "I walk into the kitchen." **Bad:** "I walk forward."
Output	Return JSON only.	`{"actions": ["entry1", "entry2", ...]}`

Table 8: Components of the action and utterance extraction prompt.

action indices become more specific by grounding actions to explicit objects and locations. Utterance indices become context-independent by resolving pronouns and ellipsis. Finally, the resulting indices are embedded using Qwen3-Embedding-4B (Yang et al., 2025a) and indexed with FAISS (Douze et al., 2025).

Activity & Conversation. Following WorldMM, we prompt an LLM to extract coarse-grained, goal-oriented activities from each 30-minute multimodal context, as shown in Fig. 17. The LLM may generate one or more activity indices, each associated with a specific time range. We observe that the generated activities are often coupled, using conjunctions such as "and" or "while" to combine multiple unrelated activities into a single index. Such coupled indices are difficult to retrieve semantically, while naively shortening them often removes important details. We therefore introduce a disentangling harness that separates coupled activities while preserving the original semantics. To reconcile activities across adjacent 30-minute segments, we use Qwen3-Embedding-4B (Yang et al., 2025a) to identify temporally continuous activities with a cosine similarity above 0.9, followed by an additional LLM call to merge them. The resulting activity indices are embedded and indexed using the same backend. Conversation indices are constructed in the same manner. We prompt the LLM to extract coarse-grained, goal-oriented conversations from each 30-minute multimodal context, as shown in Fig. 18. The extracted conversations are disentangled, merged across segment boundaries, and embedded into the conversation memory using the same pipeline.

B.3 Drafting Agent

Index Pool. The *index pool* stores all atomic memory indices retrieved from the database across retrieval rounds. Whenever new indices are retrieved, they are added to the pool, deduplicated by

content, and sorted chronologically by timestamp. At the beginning of each retrieval round, the drafting agent receives the entire index pool together with the newly retrieved indices, allowing it to progressively refine subsequent retrieval queries.

Tool Calling. We use the agentic tool calling capabilities of the latest LLMs for the drafting agent. We employ persistent thinking mode in Qwen models and function calling in GPT models for multi-round Chain-of-Thought (CoT) reasoning and atomic memory index drafting. In each round, the drafting agent decides one of five tool calls: `search_action_utterance`, `search_activity`, `search_speech`, `curate_evidence`, or `answer`. We present the detailed tool call schema in Fig. 19.

`search_action_utterance` takes one required `query` for short-term fine-grained action or utterance vector database search, which should include specific objects, persons, and/or utterances. `search_activity` takes one required `query` for long-term coarse-grained activity vector database search. `search_conversation` takes one required `query` for long-term coarse-grained conversation vector database search, which should include specific conversational topics. All three tools take an optional `time_query`, which indicates the most likely occurrence time of the searched events. It is captured by the database harness and translated into the standard timestamp format `DAYX HHMMSSFF`. This enables the agent to perform time-aware retrieval and query refinement for adaptive searching. Each time the database returns atomic memory index, the harness stores the index in a working memory *index pool*. Meanwhile, new retrieved atomic memory indices and the existing index pool are added to the agent context for the next round of CoT reasoning.

`curate_evidence` is a drafting-agent curation tool used to keep the *index pool* clean and organized, rather than the later curation agent. The agent can call this tool with a list of the retrieved memory indices to decide which memory indices to keep. The harness captures the list of indices to keep and removes the rest from the *index pool*. Empirically, we observe that strong agents such as GPT-5.4 and Sonnet-4.6 are good at curating indices, while weaker agents such as Qwen3.6-27B frequently make mistakes and become biased toward certain indices. We therefore only enable this tool call for strong agents. `answer` is the tool used to terminate the drafting agent CoT. It collects all indices in the *index pool* and sends it to the curation agent for final index curation before question answering.

Harness. `Query` and `curate_evidence` are automatically captured by the harness through string and array-list parsing. `time_query` is parsed into a canonical timestamp format, `DAYX HHMMSSFF`, which can represent either a timestamp or a time range. Since the dataset recordings can be intermittent, we concatenate the recordings into a single uniform timeline. Following Eq. 1, we map the timestamps of each candidate memory index in the vector database to a time-relevance score. We further compute the semantic similarity score by embedding the `query` using Qwen3-Embedding-4B model and search it in the pre-embedded FAISS vector database (Douze et al., 2025). The resulting semantic similarity score and time-relevance score are multiplied as the final ranking score.

B.4 Sampling Agent

The sampling agent is implemented as a single LLM call that curates the atomic memory indices in the *index pool*. Its goal is to keep the reasoning context of the final response agent concise and relevant. We prompt the agent to explicitly reason over temporal qualifiers in the question, such as "usually", "habit", "last", and "this morning" (see prompt in Fig. 20). Based on the inferred temporal intent, the agent selects memory indices that best match the question. For example, habitual questions favor diverse memories across time, whereas recency questions prioritize the most recent indices. Conversely, indices outside the relevant temporal scope are discarded. In practice, the sampling agent typically returns fewer than 15 atomic memory indices for the response agent.

B.5 Reasoning & Response

The response agent receives the curated atomic memory indices and maps their timestamps back to the corresponding multimodal captions. Given the question, answer options, query timestamp, and retrieved captions, it performs multimodal reasoning and generates the final answer.

B.6 Models & Configurations

Memory Perception & Construction. We evaluate three experimental setups for memory perception and construction. (1) Dense captions: GPT-5.4 (without thinking) is used for all memory construction calls over the human-annotated dense captions. (2) VLM narration: Gemini-3-Flash (without thinking) generates video narrations, followed by GPT-5.4 (without thinking) for memory construction. (3) Open-source pipeline: Gemma-4-31B is used for both video captioning and memory construction.

Retrieval & Response. We primarily evaluate `EgoCITE` on two agentic LLMs, GPT-5.4 and Qwen3.6-27B-FP8 with both tool calling mode. GPT-5.4 enables both function calling (tool calling) and thinking mode at medium effort with maximum 4096 output tokens for drafting agent and sampling agent. The response agent enables thinking mode at medium with maximum 4096 output tokens without tool calling. Qwen3.6-27B-FP8 enables both persistent thinking (tool calling) and thinking mode with maximum 4096 output tokens for drafting agent and sampling agent. The response agent enables thinking mode with maximum 4096 output tokens without tool calling. Both models we use the top-k values for retrieval with action and utterance memory at $k = 20$, activity memory at $k = 10$, and conversation memory $k = 10$. We set the drafting agent tool-calling round number to 5 with at most 15 retrieved atomic memory indices to align with WorldMM (Yeo et al., 2026).

C Experiment Details

C.1 Datasets & Benchmarks

EgoLife[1] is a socially intensive daily lifelogging dataset collected from six participants, comprising 7 days $\times$ 8 hours of continuous multimodal recordings per person, including egocentric RGB video and audio (Yang et al., 2025b). EgoLife provides human-annotated dense captions describing both first-person and third-person actions, as well as speech transcripts. Since the dataset is primarily in Chinese, we translate all dense captions and transcripts into English using Qwen3.6-27B-FP8 (Qwen Team, 2026). Following the practice of WorldMM (Yeo et al., 2026), we further align the translated captions and transcripts into 30-second windows to construct the multimodal captions used as the primary inputs to the memory systems. This dataset's content is used in three benchmarks: EgoLifeQA (Yang et al., 2025b), EgoR1-Bench (Tian et al., 2025), and EgoMem (Zheng et al., 2026).

EgoLifeQA[2] contains 2,905 manually annotated multiple-choice questions from EgoLife (Yang et al., 2025b) with five categories: Entity Log, Event Recall, Habit Insight, Relation Map, and Task Master. Each question includes one question statement, four answer choices, the question time, the ground-truth answer, and a target timestamp indicating the timestamps needed to answer the question. We use the question statement, answer choices, and question time as the question context for retrieval, curation, and response agent. We use the ground-truth answer and target timestamp as evaluation metrics.

EgoLifeQA Time-Aware Questions. Additionally, we split the EgoLifeQA benchmark into two subsets: *time-aware questions* and *time-unaware questions*. Time-aware questions contain keywords that clearly indicate temporal cues, including "usually", "often", "never", "morning", "afternoon", "DAY X", and "last", etc. (see Tab. 9). This subset contains 2,223 questions, covering 76% of the EgoLifeQA benchmark.

EgoMem[3] contains 939 multiple-choice questions from LifeDialBench (Zheng et al., 2026) based on contents of EgoLife (Yang et al., 2025b), including four categories: Single Event, Multiple Event, Event Detail, and Time Query. Each question includes a question statement, four answer choices, the question time, the ground-truth answer, and a target timestamp indicating the timestamps needed to answer the question. We use the question statement, answer choices, and question time as the question context for retrieval, curation, and response generation. The ground-truth answer and index timestamp are used for evaluation. Since the questions are written in third person, where camera wearers are referred to by name rather than as "I", we use GPT-5.4 (OpenAI, 2026a) to normalize the questions and choices from third person to first person. To further align with EgoLifeQA (Yang et al.,

[1] https://huggingface.co/datasets/lmms-lab/EgoLife
[2] https://huggingface.co/datasets/Ego-R1/Ego-R1-Data
[3] https://github.com/RayNeo-AI-2025/LifeDialBench

Table 9: Keywords used to classify temporal-aware questions in EgoLifeQA.

Category	Keywords
Time of day	morning, noon, afternoon, evening, night, midnight, dawn, dusk, tonight
Calendar / days	today, yesterday, tomorrow, day, days, date, week, weekend
Frequency / habit	usually, often, always, never, sometimes, frequently, rarely, routinely, regularly, habit
Ordering / recency	first, last, latest, earlier, earliest, before, previous, previously, recent, recently, ago, then, initially, originally, prior, final, finally
Generic temporal	when, again, already, just, past, moment, hour, minute, o'clock, order, sequence

2025b), we normalize the date and time into the same format, e.g., "DAY1" for dates and "12000000" for timestamps.

C.2 METRICS

Accuracy evaluates whether the response agent selects the correct choice among the candidate answers compared with the ground-truth answer. We prompt the response agent to generate a JSON output containing its prediction. Since LLMs may produce invalid JSON outputs, we implement an additional answer extraction step using an LLM call to extract the predicted choice.

Hits and NearHits evaluate *multi-round* retrieval performance by measuring whether the retrieved memory indices are temporally aligned with, or close to, the target multimodal captions needed to answer the question. *We use these metrics to evaluate multi-round retrieval agent performance.* We define *Hits* as whether the retrieved memory indices, temporally overlap with the target time. We define *NearHits* as whether the retrieved memory indices temporally overlap with a ± 5 minute window around the target time. This metric is motivated by the observation that answers can often still be inferred from nearby context even when the exact target time is missed. However, *Hits* and *NearHits* do not necessarily guarantee the correctness of the final answer.

Normalized Cost measures retrieval efficiency using the normalized API token cost of GPT-5.4 (OpenAI, 2026b). According to the GPT-5.4 pricing, short-context models incur a normalized cost of $1\times$ and $6\times$ per input and output token, respectively, while long-context models incur $2\times$ and $9\times$. `EgoCITE-GPT` and other GPT-based agentic memory baselines use short-context pricing, whereas the long-context GPT agent uses long-context pricing. The normalized cost is computed as the weighted sum of input and output token costs.

C.3 BASELINES

Long-context LLM Agents. State-of-the-art LLMs, such as GPT-5.4 (OpenAI, 2026a) and Gemini-3.1-Pro (Team et al., 2023), provide built-in agentic capabilities through tool use (e.g., keyword search and Python execution). We therefore treat them as long-context LLM agent baselines. Since the 56-hour video recording exceeds the maximum context window of commonly used VLMs, we use JSON files containing timestamps and multimodal captions as the inputs to LLMs with 1M-token context windows. We set the thinking effort of GPT-5.4 to medium and Gemini-3.1-Pro to medium. The LLMs are prompted to gather index from the input JSON documents and answer the questions. Each JSON document consumes 0.7–0.9M input tokens. To efficiently evaluate the questions, we employ a batch prompting strategy (Cheng et al., 2023), grouping 50 questions at a time. Questions without valid answers are re-processed. For latency and cost evaluation, we randomly sample 100 questions across the three benchmarks and prompt each question independently to obtain unbiased per-question efficiency measurements.

Table 10: Component ablation study.

Metric	Caption RAG	+EgoIndex	+EgoScheme	+EgoRetrv
Accuracy (%)	61.3	69.7	71.3	75.3
Hit Rate (%)	40.7	54.3	59.3	62.7

EgoRAG[4] uses multi-granularity summaries as the indices of the vector database (Yang et al., 2025b). During memory construction, we use GPT-5.4 to generate summaries from multimodal captions at multiple granularities and store them in the vector database. During retrieval, we use GPT-5.4 to generate vector database queries, and use GPT-5.4 with medium thinking effort to answer the questions. We primarily follow the original codebase, except for replacing the backend LLMs.

A-MEM[5] uses keyword tags as the memory index (Xu et al., 2026). During memory construction, we use GPT-5.4 to extract keywords and tag the multimodal captions accordingly. During retrieval, we use GPT-5.4 to generate queries, and use GPT-5.4 with medium thinking effort to answer the questions. We retrieve the top-$k = 15$ captions. We primarily follow the original codebase, except for replacing the backend LLMs and providing keyword extraction, retrieval, and response prompts from EgoRAG.

VideoRAG[6] employs both knowledge graph entity and visual vector for memory indices. At retrieval stage, it utilizes both semantic similarity comparison and graph walk to retrieve relevant captions and video clips. We use the multimodal captions for entity knowledge graph construction and videos for visual memory. We primarily follow the code base except replacing GPT-4o model in entity extraction and final answering stage into GPT-5.4 model and GPT-4o model in retrieval flows into GPT-5.4 model. We reuse the visual embeddings from WorldMM built by VLM2Vec for visual memory.

WorldMM[7] employs three types of memory, including episodic, semantic, and visual memory, together with multi-round adaptive retrieval (Yeo et al., 2026). Episodic and semantic memory are stored in knowledge graphs and visual memory is stored in vector database. We primarily follow the pipeline of the original codebase. We use GPT-5.4 to construct the three memory types from multimodal captions instead of GPT-5-mini. Similarly, we use GPT-5.4 instead of GPT-5 as the retriever, and GPT-5.4 with medium reasoning effort instead of GPT-5 for final question answering. The visual memory is built by visual embedding model VLM2Vec. We set the maximum number of retrieval rounds to 5 and maximum number of memory indices retrieved as 15, following default implementation of WorldMM.

C.4 Hardware

All small embedding models, including the Qwen3-Embedding-4B text embedding model and the VLM2Vec visual embedding model, are processed locally on a machine equipped with an RTX 4090 GPU, an Intel i5-13600KF CPU, and 128GB RAM. The Qwen3.6-27B-FP8 and Gemma-4-31B model are served on a cloud server equipped with 4× RTX 6000 Pro GPUs using vLLM (Kwon et al., 2023).

D Additional Results

NearHits. We further evaluate the retrieval agent using the NearHits metric (Tab. 12), where retrieval is considered successful if any of the at most 15 retrieved memory indices fall within a ±5-minute window of the ground-truth evidence. `EgoCITE-GPT` achieves the highest NearHits rates of 69.4%, 90.8%, and 70.3% on EgoLifeQA, EgoMem, and EgoR1, respectively, improving over the strongest baseline by up to 28.7%, 14.8%, and 24.3%. Compared with the Hit Rate results, the consistently

[4] https://github.com/EvolvingLMMs-Lab/EgoLife
[5] https://github.com/agiresearch/a-mem
[6] https://github.com/HKUDS/VideoRAG
[7] https://github.com/wgcyeo/WorldMM

Table 11: Time-decay factor λ ablation on EgoCITE-GPT.

Metric	$\lambda = 0.97$	$\lambda = 0.98$	$\lambda = 0.99$	$\lambda = 0.995$
Accuracy (%)	72.0	76.0	**81.0**	76.0
Hit Rate (%)	56.0	55.0	**67.0**	54.0

Table 12: Retrieval NearHits rate (%). NearHits indicate the retrieved memory indices fall into the $\pm$5-minute margin of ground truth evidence.

Method	**EgoLifeQA**	**EgoMem**	**EgoR1**
EgoRAG (Yang et al., 2025b)	14.5	44.9	13.5
A-Mem (Xu et al., 2026)	33.2	46.6	36.3
VideoRAG (Ren et al., 2026)	6.2	19.7	9.7
WorldMM-Qwen (Yeo et al., 2026)	39.5	76.0	46.0
WorldMM-GPT (Yeo et al., 2026)	40.7	73.5	45.3
EgoCITE-Qwen (Ours)	63.6	*75.6*	68.3
EgoCITE-GPT (Ours)	**69.4**(+28.7)	**90.8**(+14.8)	**70.3**(+24.3)

higher NearHits rates indicate that EgoCITE frequently localizes evidence to the correct temporal neighborhood even when the exact target timestamp is not retrieved.

Retrieval Efficiency and Cost. We report the per-question input tokens, output tokens, latency, and API cost in Tab. 13. EgoCITE-GPT offers a tunable accuracy-efficiency trade-off. The 1-round variant achieves 58.7% accuracy using only 9.3k input tokens, $0.034 per query, and 15.2s latency. Increasing retrieval depth to five rounds improves accuracy by 5.1% at the cost of 3.4$\times$ more input tokens, 2.1$\times$ higher latency, and 3.2$\times$ higher cost. Compared with GPT-5.4, the default 5-round configuration achieves higher accuracy while using 24.7$\times$ fewer input tokens and 36$\times$ lower cost with comparable latency. It also requires 2.3$\times$ fewer input tokens, 2$\times$ lower cost, and 14.8s lower latency than WorldMM-GPT while achieving substantially higher accuracy. Overall, EgoCITE achieves an effective balance between accuracy, latency, and computational cost.

Memory Granularity Ablation. We ablate the multi-view design of EgoIndex in EgoCITE-GPT by varying memory granularity on EgoR1-Bench. Fine-grained memory (actions and utterances) achieves 69.7% accuracy and 55.7% hit rate, while coarse-grained memory (activities and conversations) reaches 73.3% and 59.3%, respectively. Combining all four views further improves accuracy to 75.3% and hit rate to 62.7%. These results demonstrate that fine-grained and coarse-grained memories provide complementary information for long-horizon retrieval and QA.

Video Caption Results. We replace the default human-annotated dense-caption memory source with video narrations generated by Gemini-3-Flash and open-weight Gemma-4-31B. The Gemini pipeline uses GPT-5.4 for memory construction, consistent with the dense-caption setting, while the Gemma pipeline uses Gemma-4-31B throughout memory construction to evaluate a fully open-weight setting. All variants use Qwen3.6-27B-FP8 as the retrieval agent and are evaluated on EgoLifeQA. As shown in Tab. 14, replacing human-annotated dense captions with VLM-generated narrations results in only a modest performance drop. Gemini captions achieve 60.3% accuracy and 40.7% hit rate, only 1.2% and 2.9% lower than dense captions, respectively, while the fully open-weight Gemma pipeline achieves 57.1% accuracy and 36.3% hit rate. These results demonstrate that EgoCITE generalizes well to automatically generated video narrations, enabling practical deployment without relying on human-annotated dense captions.

E Future Work and Discussion

Although EgoCITE demonstrates that memory representation is critical for long-horizon egocentric retrieval, several challenges remain.

Table 13: Per-question cost breakdown for the latency–accuracy tradeoff. Cost is computed based on GPT-5.4 API pricing (OpenAI, 2026b).

Method	Input tok.	Output tok.	Latency (s)	Cost per Q ($)
EgoCITE-GPT (1 Round)	9,288	743	15.2	0.034
EgoCITE-GPT (5 Round)	31,670	2,073	32.0	0.110
GPT-5.4 agent (OpenAI, 2026a)	783,123	2,826	32.5	3.979
VideoRAG (Ren et al., 2026)	21,100	3,162	26.4	0.100
WorldMM-GPT (Yeo et al., 2026)	71,801	2,327	46.8	0.214
A-Mem (Xu et al., 2026)	1,598	805	9.6	0.016
EgoRAG (Yang et al., 2025b)	15,914	525	10.6	0.048

Table 14: Video captioning variants comparison to dense caption baseline on EgoLifeQA.

Metric	Human-annotated	Proprietary VLM	Open-source VLM
Video narrator	Dense Caption	Gemini-3-Flash	Gemma-4-31B
Memory constructor	GPT-5.4	GPT-5.4	Gemma-4-31B
Hit rate (%)	43.6	40.7	36.3
Average accuracy (%)	61.5	60.3	57.1
EntityLog	62.7	61.4	56.0
EventRecall	59.5	57.6	53.8
HabitInsight	58.8	59.8	59.1
RelationMap	63.6	60.5	59.9
TaskMaster	67.1	70.1	68.2

Multimodal Perception. Memory quality is still bounded by multimodal perception, particularly person identification under privacy-preserving face-redaction settings. An individual's appearance (e.g., clothing and hairstyle) may change substantially over long time horizons, making reliable visual identification difficult for both CV models and VLMs. Additionally, our system assumes accurate ASR transcripts with resolved speaker identities. In realistic settings, however, ASR errors, overlapping speech, and speaker diarization failures may further degrade memory construction and retrieval. Improving robust multimodal perception remains an important direction for future egocentric memory systems.

Cross-Modality Index Construction. We resolve coreference and ellipsis over text-based multimodal captions using LLMs, treating captions as an intermediate representation of the underlying multimodal signals. This abstraction inevitably loses cross-modal cues that are useful for reference resolution, such as gaze, pointing gestures, speaker localization, and object interactions. Future work could instead perform coreference and ellipsis resolution directly over synchronized multimodal signals, allowing speech, vision, and actions to jointly ground entities, events, and omitted references. Such cross-modal memory construction would produce richer memory indices and reduce the need to densely caption and index long-horizon videos, improving both efficiency and retrieval quality.

Memory Structure. We adopt a simple vector database to isolate the effect of memory representation in this work. We believe that the quality of the underlying memory indices is a prerequisite for any memory system, regardless of whether the indices are stored in a vector database, knowledge graph, or other structured memory. Richer memory structures may offer additional benefits, such as relational reasoning, hierarchical organization, or more efficient graph traversal. Future work could therefore integrate the proposed memory indexing scheme with sophisticated memory structures, enabling them to reason over cleaner, self-contained, and contextually complete memory indices.

Reasoning and Guessing. Hit and near-hit rates measure retrieval alignment but do not fully determine QA correctness. On EgoLifeQA, `EgoCITE-GPT` and `EgoCITE-Qwen` retrieve the target evidence but answer incorrectly on 11.3% and 9.1% of questions, while 12.6% and 15.3% are answered correctly despite missing the annotated evidence. These mismatches highlight two directions for future work: (1) improving agent reasoning over correctly retrieved memories to reduce reasoning

failures, and (2) improving benchmark questions and ground-truth evidence annotations, which may contain annotation noise or non-exclusive evidence that allows an answer to be inferred or guessed from memories outside the annotated target.

F EXAMPLES AND COMPARISONS

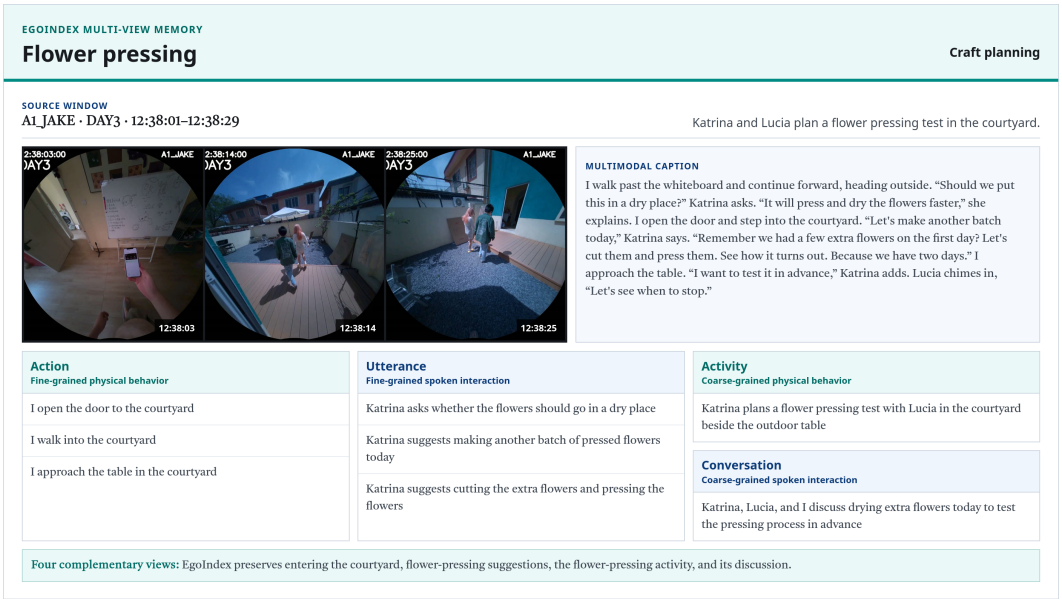

Figure 7: EgoIndex multi-view memory example for flower pressing. From one caption window, EgoIndex constructs fine-grained action and utterance indices alongside coarse-grained activity and conversation indices.

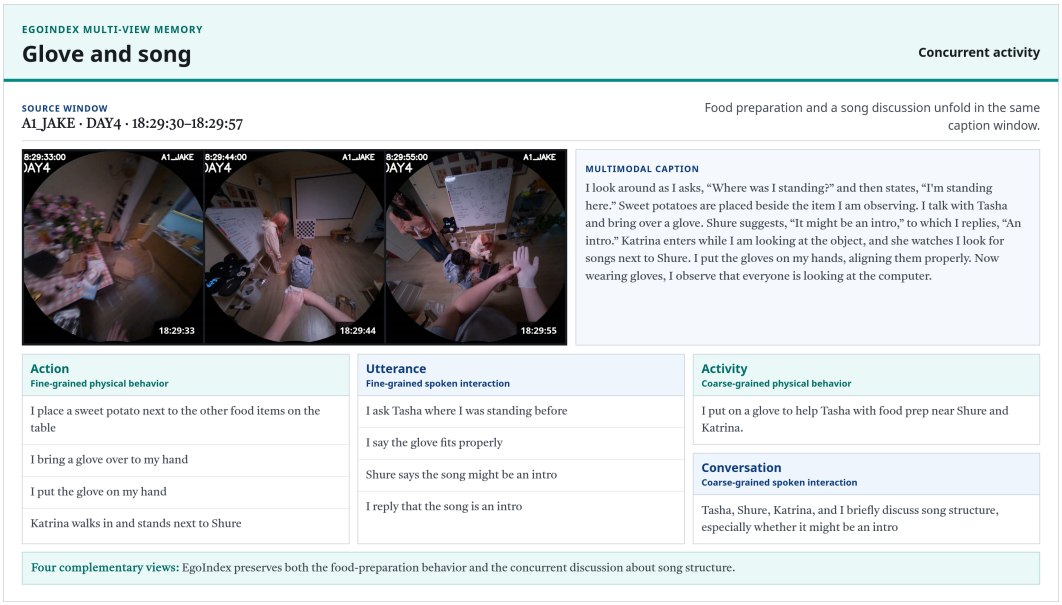

Figure 8: EgoIndex multi-view memory example for concurrent activity. EgoIndex separately preserves food-preparation actions, utterances about the song, the broader food-preparation activity, and the conversation about song structure.

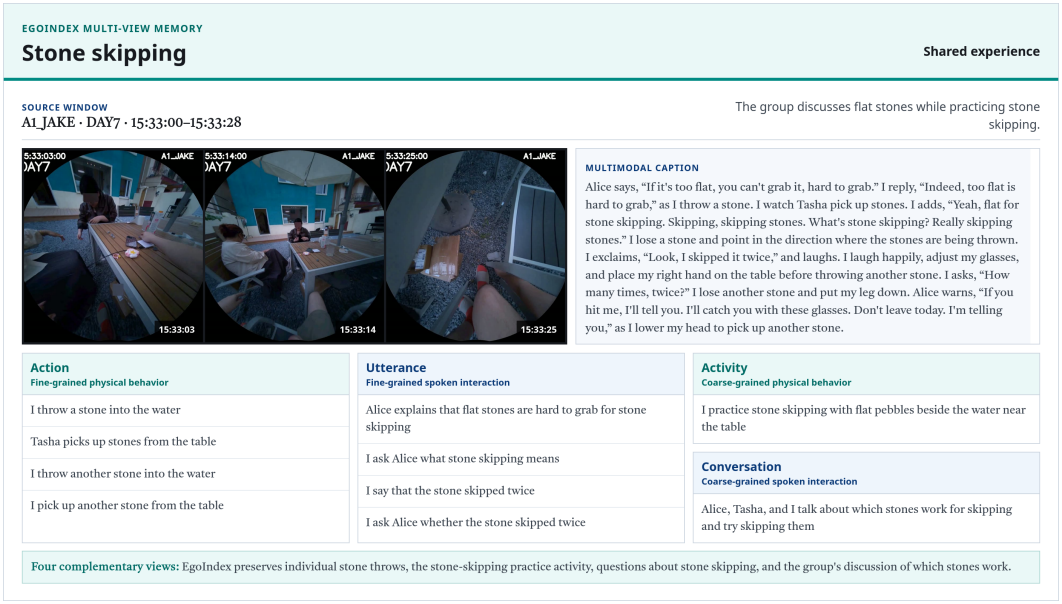

Figure 9: EgoIndex multi-view memory example for stone skipping. EgoIndex preserves individual stone-related actions and utterances together with the broader stone-skipping activity and conversation topic.

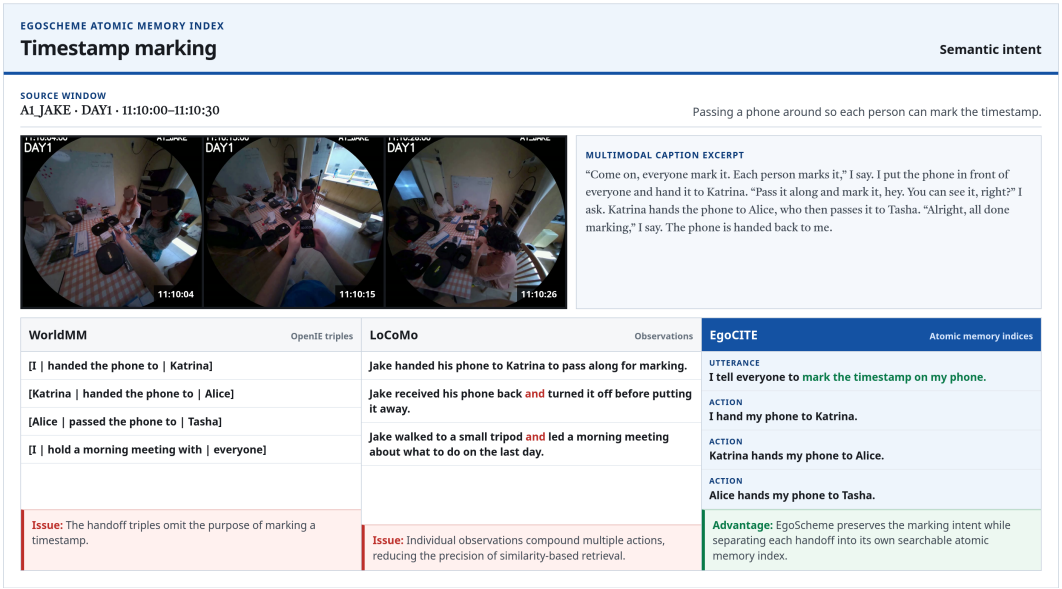

Figure 10: EgoScheme example for timestamp marking. Compared with WorldMM's context-poor triples and LoCoMo's compounded observations, EgoScheme preserves the marking intent and separates the phone handoffs into atomic memory indices.

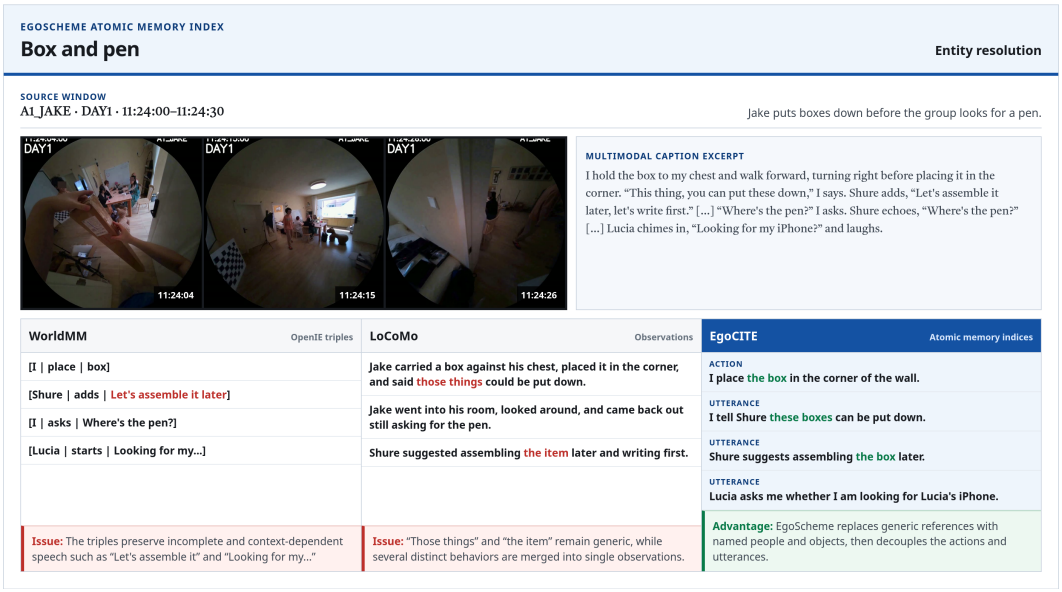

Figure 11: EgoScheme example for entity resolution. WorldMM and LoCoMo retain context-dependent references such as "it," "those things," and "the item," whereas EgoScheme resolves the referenced boxes and decouples the corresponding actions and utterances.

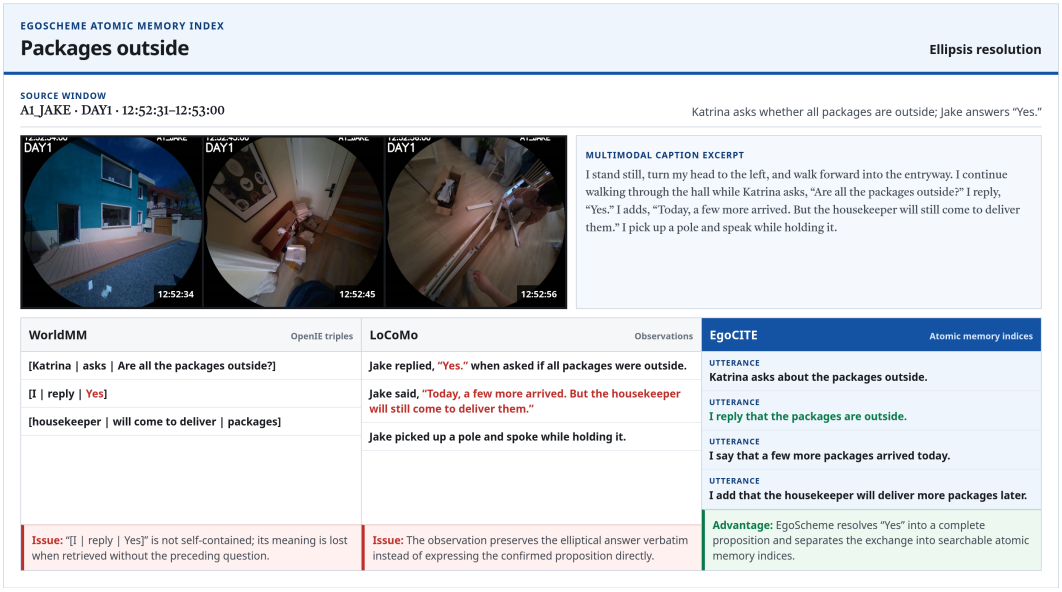

Figure 12: EgoScheme example for ellipsis resolution. WorldMM and LoCoMo preserve the elliptical reply "Yes," while EgoScheme constructs the self-contained atomic memory index "I reply that the packages are outside."

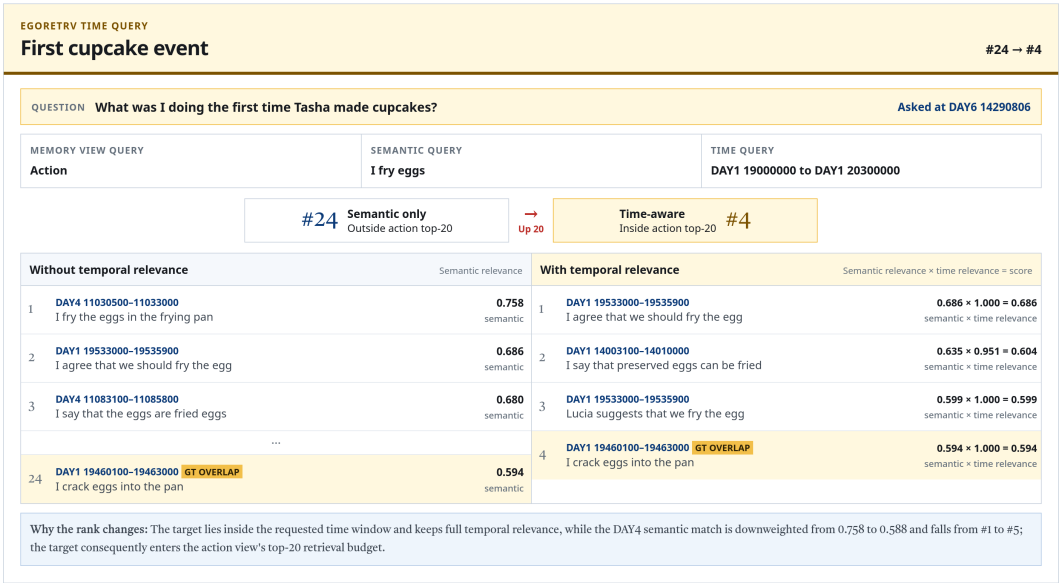

Figure 13: EgoRetrv time-query example for the first cupcake event. Temporal relevance scoring moves the target-overlapping action index from rank 24 under semantic-only retrieval to rank 4, placing it inside the action view's top-20 retrieval budget.

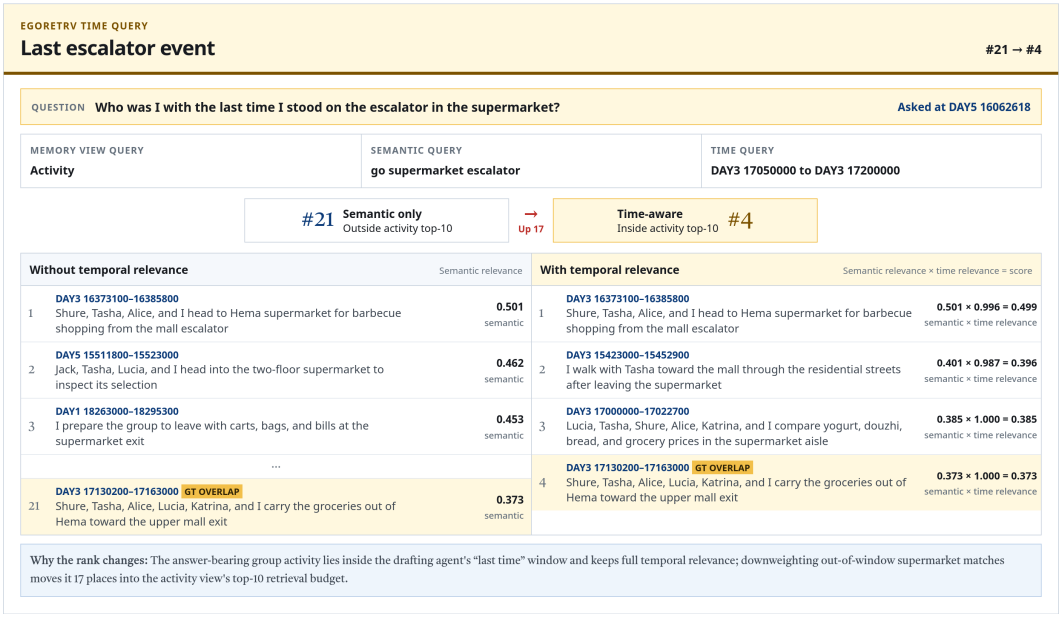

Figure 14: EgoRetrv time-query example for the last escalator event. Applying the question-conditioned time query moves the target-overlapping activity index from rank 21 to rank 4, placing it inside the activity view's top-10 retrieval budget.

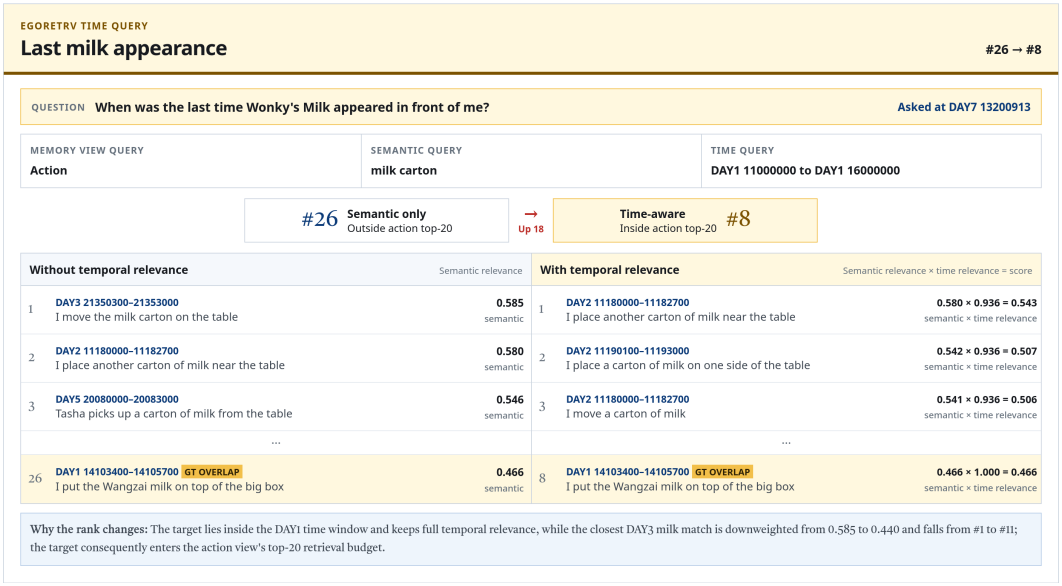

Figure 15: EgoRetrv time-query example for the last appearance of milk. Temporal relevance scoring moves the target-overlapping action index from rank 26 to rank 8, placing it inside the action view's top-20 retrieval budget.

G PROMPTS

Action & Utterance Extraction Prompt.

```
ONLY extract these FOUR types of entries — drop everything else:
A. HAND-OBJECT INTERACTION — someone (first-person "I" or a named housemate)
   directly handles or uses an object.
   e.g. "I fasten the shelf bracket with a screwdriver" / "Alice chops the onion".
B. BIG DISPLACEMENT — someone enters or moves into an explicitly named place.
   e.g. "I walk into the kitchen" / "Shure enters the living room" / "I sit down on the chair in living room
      ".
C. UTTERANCE OR SPEECH — someone talks, asks, tells, suggests, agrees, or explains.
   e.g. "Jake asks Tasha about the dinner menu", "Jake agrees with Tasha on hiking this afternoon".
D. MEANINGFUL GESTURES — someone laughs, cries, etc.
   e.g. "Jake laughs out loudly".
Anything that is not type A, B, C, or D — bare gestures, posture, glances, idle — must NOT be an entry.

RULES
1. ONE entry per distinct action/fact. Never join two with "and"/"while"/"then"/"as" — split instead.
2. Each entry is 8–20 words, "subject + verb + object". Use "I" for the wearer and a real NAME for anyone
      else; keep named objects/places verbatim.
3. NAME every person, object, and place. Resolve all pronouns and generic words to a real NAME, "I", or the
      actual object/place. Drop an entry only if a referent is truly unresolvable.
4. Turn dialogue into a speech act with a SPECIFIC topic, stated as a WHAT or THAT clause. The clause MUST
      be complete and self-contained.
   Good: "Jake tells Tasha that dinner is ready", "I ask Alice about the shelf height"
   Bad:  "Jake tells Tasha that" (fragment), "I ask Alice about it" (not specific)
5. Drop micro-gestures and undirected motion. Keep a motion only when it reaches a named destination.

GOOD EXAMPLES:
"I message Lucia on my phone", "I fasten the shelf bracket with a screwdriver", "Shure hands me the
      screwdriver at desk"

BAD EXAMPLES:
"Jake says OK." -- Lack of context. Should be "Jake agrees with Tasha on hiking this afternoon".
"I use tools." -- Lack of details. Should be "I use spoon to install the shelf".
"Jake says, 'I will handle it'." -- Unclear coreference and quotes. Should be "Jake says he will handle the
      cleanup job".
"I'm working on computer while listening to Jake talking about weather." -- Coupled actions. Should be "I'm
      working on computer to edit video" and "Jake says weather will be bad tomorrow".

RESOLVE: "He picks it up" -> "Jake picks up the bottle on table"; "I walk forward" -> "I walk into the
      kitchen"

Output JSON only: {"actions": ["entry1", "entry2", ...]}
```

Figure 16: EgoScheme action and utterance resolution and extraction prompt.

Activity Extraction Prompt

```
You are given timestamped 30-second captions, indexed 0..N-1.
Segment the captions into CONTIGUOUS activity blocks, where each block represents ONE
immediate goal pursued by the wearer (and the people present).

BOUNDARY RULES:
1. Each block must span between 1 and 10 minutes of wall time
   NEVER produce a block longer than 10 minutes.
2. Blocks must be contiguous and non-overlapping; every caption index belongs to
   exactly ONE block, and indices within a block must form a consecutive run.
3. Split when the immediate goal CHANGES — a different task, different place,
   different set of people, or a clear transition.

For EACH block, write a `summary` that MUST:
- Start with "I" or a real NAME (Jake/Alice/Tasha/Lucia/Katrina/Shure);
- name the SPECIFIC object, topic, people, and place;
- state the GOAL of the block, not raw motions or sub-steps,
  e.g. "Alice, Bob, and I discuss the dinner plan", "I organize the living room for the party";
- describe ONE thing — never use "and", "while", or "then";
- be specific and rich about the expression;
- never use a vague umbrella noun (techniques, topics, things, stuff, the project, ideas).

Good: "Alice, Bob, and I plan the group dinner in the kitchen"
Good: "I edit the Earth Day promo video at my desk"
Bad:  "We discuss the techniques"   (vague umbrella noun)
Bad:  "I pick up the bottle"        (a raw sub-step, not the block's goal)
Bad:  "I organize the room"         (too simple, needs details)

Output JSON only:
{"groups": [{"indices": [0,1,2,...], "summary": "..."}, ...]}
```

Figure 17: EgoScheme activity extraction prompt.

Conversation Extraction Prompt

```
You are given timestamped 30-second egocentric captions,
indexed 0..N-1. Find the CONVERSATION segments and label each with the broad
TOPIC being discussed.

Housemates — always use the real NAME: Jake, Alice, Tasha, Lucia, Katrina, Shure.

WHAT COUNTS AS CONVERSATION:
A caption is part of a conversation if the wearer ("I") is actively talking with
one or more named housemates — asking, replying, debating, joking, planning,
explaining, complaining, agreeing, etc. Reading aloud to oneself, watching a
video, or muttering does NOT count. Silent presence near someone does NOT count.

BOUNDARY RULES:
1. Each group must contain CONTIGUOUS caption indices (a consecutive run).
   Indices within a group should describe ONE coherent conversation topic.
2. A group ends when the topic shifts to a clearly different subject, OR when
   conversation pauses for >=1 minute of non-conversation activity (start a new
   group when conversation resumes).
3. Captions that are NOT part of any conversation must be LEFT OUT — do not
   force-assign them. The union of groups need NOT cover every index.
4. Empty groups list is fine if there is no conversation in this window.

For EACH group, write a `summary` that MUST:
- be 10–20 words;
- start with "I" or a real NAME;
- name every participant present in the conversation (including "I");
- state the BROAD TOPIC of the conversation, not raw motions or sub-quotes;
- describe ONE topic;
- be specific (what about the topic) — no vague umbrella nouns
  (techniques, topics, things, stuff, the project, ideas, plans).

Use plain English — write "and" normally; do NOT substitute "+", "&", "/", or "plus" for it.

Rules on connectives ("and" / "while" / "then"):
- OK to use them to list PARTICIPANTS, OBJECTS, or closely-related sub-items
  of the same topic.
- NOT OK to use them to join TWO DIFFERENT TOPICS into one summary.
  If two distinct topics were discussed, return TWO separate groups instead.

Good: "Alice, Lucia, and I plan the dinner menu and seating for tomorrow night"
Good: "Jake and I debate which model to use for the demo video"
Bad:  "We talk about stuff in the kitchen"              (vague umbrella noun)
Bad:  "I say hi to Alice"                               (gesture, not a topic)
Bad:  "Alice tells me she is going to the store"         (sub-quote, not the topic)
Bad:  "Jake + I talk about the model"                   (use "and", not "+")
Bad:  "I discuss the demo video and we plan dinner"      (two topics — split into two groups)

Output JSON only:
{"groups": [{"indices": [0,1,2,...], "summary": "..."}, ...]}
```

Figure 18: EgoScheme conversation extraction prompt.

Drafting Agent Tool Call Schema

```
{
  "tools": [
    {
      "name": "search_action_utterance",
       "description": "Search the ACTION index (atomic 30-sec observations of objects/people/physical
      movements/utterance). Adds matching captions to the working evidence pool with fresh integer IDs.",
      "parameters": {
        "query": {
          "type": "string",
          "required": true,
               "description": "Grounded action query (object / action+object / person+action / person+action+
      object / person+utterance),
              <=12 words."
           },
          "time_query": {
            "type": "string",
            "required": false,
          "description": "Optional time window in dataset format ('DAYn HHMMSSFF' or 'DAYn HHMMSSFF to DAYm
      HHMMSSFF'). Omit if no time signal."
          }
       }
    },
    {
      "name": "search_activity",
      "description": "Search the ACTIVITY index (5-min activity summaries). Adds matching captions to the
      working evidence pool.",
      "parameters": {
        "query": {
          "type": "string",
          "required": true
        },
        "time_query": {
          "type": "string",
          "required": false
        }
       }
    },
    {
      "name": "search_conversation",
      "description": "Search the CONVERSATION index (conversation-topic summaries). Use TELEGRAPHIC keyword
      utterance phrasing. Adds matching captions to the working evidence pool.",
      "parameters": {
        "query": {
          "type": "string",
          "required": true
        },
        "time_query": {
          "type": "string",
          "required": false
        }
       }
    },
    {
      "name": "curate_index",
      "description": "Prune the working evidence memory: keep only the listed . Use after each batch of
      searches to maintain a focused memory of 5-15 items. Do NOT over-curate: keeping fewer than 5 items
      risks discarding related evidence the QA agent needs.",
      "parameters": {
        "keep": {
          "type": "array[int]",
          "required": true,
          "description": "Evidence IDs to keep in working memory."
        }
       }
    },
    {
      "name": "answer",
      "description": "END the retrieval phase. Takes no arguments. A SEPARATE ANSWERING AGENT will be
      invoked with the captions in your working memory + the question + the choices, and will produce the
      final reasoning and answer letter itself. You do NOT pick the evidence, then call answer(). Call this
      ONLY after curate_index.",
      "parameters": {}
    }
  ]
}
```

Figure 19: EgoRetrv drafting agent tool call schema.

Sampling Agent Prompt

```
You are an evidence curator for a personal egocentric memory system.
You are given a pool of candidate evidence items retrieved for a multiple-choice question.
Your job is to select the most relevant evidence items for the final QA agent.

# Selection criteria
Keep between 5 and 15 items. Do NOT over-curate: leaving only 1 or 2 items
risks discarding related evidence the QA agent needs to answer correctly.
Prefer items that:
- Directly mention the key entity / action / object / person the question asks about.
- Help resolve time qualifiers ("last", "first", "yesterday", "recently") unambiguously.
- Distinguish between the answer choices — favour items that support one choice over another.

Drop items that are:
- Redundant (same event repeated → keep the most informative one).
- Clearly off-topic or unrelated to the question and choices.

# Curation policy (apply based on question wording)
- "usually" / "often" / "typically": curate a DIVERSE set spanning the full timeline.
- "last" / "most recent" / "latest": curate the LATEST timestamps.
- "first" / "earliest": curate the EARLIEST timestamps.
- "before X" / "after X": curate items in the relevant time window relative to event X.

Call `curate_index` once with your final selection, then stop.

User message:
Question: {question}
A) {choice_A}
B) {choice_B}
...

--- CANDIDATE EVIDENCE (N items) ---
  [E1] [DAY1 HHMMSSFF-HHMMSSFF] caption text...
  [E2] [DAY1 HHMMSSFF-HHMMSSFF] caption text...
  ...

Call `curate_index` with the IDs to keep (<=15).
```

Figure 20: EgoRetrv sampling agent prompt.

Response Agent Prompt

```
You are an egocentric memory assistant. The person wearing the camera is asking about their own past
      experiences. You are given a curated set of 30-sec VIDEO captions (with timestamps) that a retrieval
      agent already selected as the evidence for this question.

Use ONLY these captions to answer the multiple-choice question. Do not invent details that are not in the
      captions.

Follow these two grounding steps before committing to a letter:
  1. GROUND every element of the question in the captions. For each key noun, person, object, or action in
      the question, find the caption(s) that actually describe it. If a caption does not mention it, treat it
       as not happening.
  2. GROUND the time qualifier (if any). Words like 'last', 'yesterday', 'recently', 'first', 'before', '
      after', 'the day before yesterday' must be resolved against the caption timestamps and the memory query
       point — NOT against general daily schedules.

Output format (STRICT):
  - A few sentences of REASONING that cite the specific captions (by their timestamp) and explain why they
      support the chosen answer and rule out the others.
  - A single final line: `Answer: <letter>` where <letter> is A, B, C, or D.

User message:
[Memory query point: DAYn HHMMSSFF]

--- CURATED EVIDENCE (N raw 30-sec cap
  [DAY1 HHMMSSFF-HHMMSSFF] caption text...
  [DAY1 HHMMSSFF-HHMMSSFF] caption text...
  ...

Question: {question}
A) {choice_A}
B) {choice_B}
C) {choice_C}
D) {choice_D}

Reason from the captions above, citing
End with a single line: `Answer: <letter>`.
```

Figure 21: EgoCITE response agent prompt.