# KEEPING PACE WITH VIDEO: ONLINE MULTIMODAL MEMORY CONSTRUCTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Long-form video assistants must make incoming observations searchable during ongoing use, before future questions are known. This requires online memory construction that keeps pace with the video stream under a fixed compute budget. When construction falls behind, unfinished work accumulates and access to recent evidence is delayed. We formulate online memory construction as a joint quality-stability problem: supporting useful retrieval while keeping construction workload bounded under sustained arrivals. Building on the insight that linguistic indexing does not require generating descriptions, we introduce **LoreMem** (**L**exicon-**O**rdinal **Re**trieval **Mem**ory), a caption-free multimodal memory. Using frozen encoders, **LoreMem** selects concepts from a predefined lexicon to index visual observations and dialogue, storing sparse concept identities and ranks. The index provides semantic cues for locating evidence, while the original observations retain the details needed for answering. On completed indexes across four video QA benchmarks, **LoreMem** achieves 52.0% average accuracy, exceeding Qwen3-VL-4B caption memory by 7.9 percentage points. Separately, workload evaluation using construction times directly measured on a Galaxy S26 Ultra shows no backlog accumulation over six hours of 30-second arrivals.

## 1 INTRODUCTION

Advances in large vision-language models (LVLMs) and wearable cameras are making video assistants that observe everyday life increasingly feasible (Shen et al., 2025; 2024; Yang et al., 2025). Such assistants support diverse use cases, from retrospective search over completed recordings to questions asked while recording is still in progress. A user may, for example, ask where they left an item earlier in the day while the assistant continues to observe new events. Supporting such interactions without waiting for a later offline processing stage calls for a memory that makes observations searchable during ongoing use. A persistent semantic index allows relevant observations to be retrieved without repeatedly encoding and scanning the growing video archive. This motivates online memory construction that keeps pace with incoming video within a mobile device's memory and compute budgets, before future questions are known.

Existing video-memory research has advanced how observations are represented, organized, and retrieved for downstream answering (Wang et al., 2023; Fan et al., 2024; Ma et al., 2025; Yeo et al., 2026b). However, construction stability under sustained video arrivals has not been a central criterion in their design and evaluation. If construction persistently falls behind, unfinished work accumulates and the searchable memory becomes increasingly outdated, delaying access to recent evidence even when retrieval quality is high. We therefore formulate *online memory construction* as a joint quality-stability problem: a memory must support useful retrieval while keeping its outstanding construction workload bounded under sustained arrivals and a fixed compute budget. This workload includes unfinished work carried over from earlier segments, exposing construction debt that isolated construction-time measurements do not capture.



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

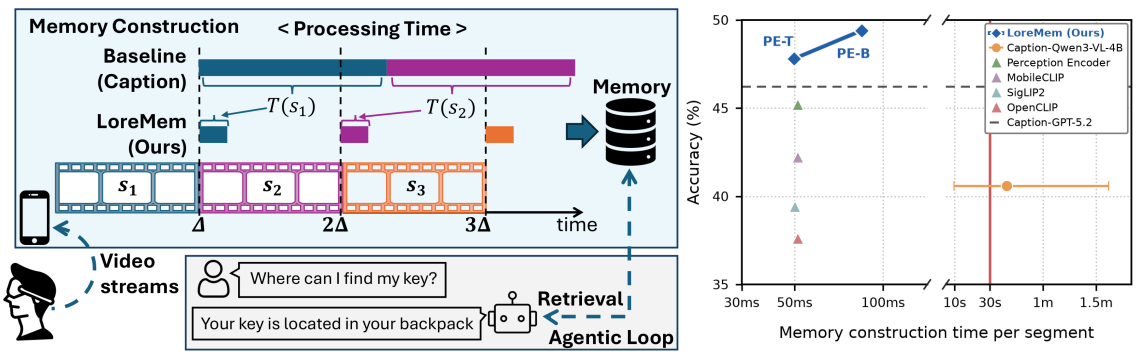



Figure 1: **Memory quality and measured construction cost. Left:** Backlog accumulation under online memory construction. The evaluated caption configuration falls behind the incoming video stream, accumulating backlog, while the embedding-based methods keep pace. **Right:** EgoLifeQA accuracy versus directly measured construction time with Galaxy S26 Ultra. Triangles denote visual embedding methods: Perception Encoder, MobileCLIP, SigLIP2, and OpenCLIP; `PE-T/B` are the Tiny/Base Perception Encoder variants. Qwen shows the mean and min-max range of measured construction times that depend on the number of output tokens. Red marks indicate 30-second arrivals, and gray indicates accuracy with GPT-5.2 caption.

Caption-based memories describe events in language for later retrieval (Wang et al., 2023; Zhang et al., 2024). These descriptions can capture rich details, but autoregressive generation is costly, and compact generators may sacrifice downstream quality. Visual-embedding memories offer inexpensive, non-autoregressive construction and natural-language retrieval (Cherti et al., 2023; Vasu et al., 2024; Tschannen et al., 2025; Bolya et al., 2025), with directly measured construction times below the arrival interval in Figure 1. Since everyday questions can depend on both visual events and dialogue, we seek an effective linguistic index that makes both sources searchable at similarly low construction cost.

Our approach builds on a simple observation: *linguistic indexing does not require generating descriptions.* A retrieval index need not describe an observation exhaustively; it needs semantic cues that help locate relevant evidence. We use a predefined lexicon as a fixed set of these cues, indexing visual observations and dialogue by their most relevant concepts. Although a fixed lexicon is less expressive than free-form captions, it can guide search over the original observations, whose details remain available for answering.

We introduce **`LoreMem`** (**L**exicon-**O**rdinal **Re**trieval **Mem**ory), a caption-free multimodal memory built from a predefined lexicon. We precompute lexicon embeddings in frozen visual and text encoder spaces. For each segment, we store the top concept identities and ranks separately for visual and dialogue evidence. Retrieval aggregates the stored ranks of query-selected concepts, avoiding direct comparison of raw similarity magnitudes across encoders. Construction is non-autoregressive and requires no task-specific paired training; the sparse records are built before future questions are known and reused across queries.

On completed indexes across four benchmarks, **`LoreMem`** achieves 52.0% average QA accuracy, exceeding both Qwen3-VL-4B caption memory and Perception Encoder visual memory by 7.9 percentage points. These results show that a fixed lexicon can support effective indexing without generated descriptions. Separately, workload evaluation using directly measured on-device construction times shows no backlog accumulation over six hours of 30-second arrivals.

Our contributions are threefold: **(1)** We formulate online memory construction through joint requirements on downstream utility and bounded construction workload **(Section 3)**. **(2)** We introduce **`LoreMem`**, a caption-free multimodal memory that indexes visual and dialogue evidence using a predefined lexicon, with sparse concept ranks for retrieval and no task-specific training **(Section 4)**. **(3)** We evaluate the indexing utility of this representation through completed-index retrieval and QA across four benchmarks, and directly measure mobile construction times to assess workload over six hours of segment arrivals **(Section 5)**.

## 2 Related Work

**Direct long-video QA and visual-embedding memory.** Long-video QA uses compression, selection, and recurrent memory. MovieChat (Song et al., 2024) and MovieChat+ (Song et al., 2026) construct sparse memories, MA-LMM (He et al., 2024) updates a visual memory bank, and LongVU (Shen et al., 2025) removes redundant tokens. ReWind (Diko et al., 2025) learns instruction-conditioned memory; VideoTree (Wang et al., 2025d) expands a hierarchy. Time-R1 (Wang et al., 2025a) and Video-RTS (Wang et al., 2025c) use reinforcement learning for temporal grounding and adaptive frame selection, respectively.

**Caption-style memory.** Encode–Store–Retrieve (Shen et al., 2024) verbalizes observations, Lifelong-Memory (Wang et al., 2023) organizes descriptions hierarchically, and Video ReCap (Islam et al., 2024) generates captions at multiple temporal scales. LLoVi (Zhang et al., 2024) answers from frame descriptions; LangRepo (Kahatapitiya et al., 2025) compresses language into a temporal hierarchy. Video-RAG (Luo et al., 2025) retrieves ASR, OCR, and object-detection text. LightRAG (Guo et al., 2025) builds an entity–relation graph; here it operates on generated segment descriptions.

**Agent-based and retrieval-augmented video QA.** EgoRAG (Yang et al., 2025) retrieves egocentric experiences, DrVideo (Ma et al., 2025) iteratively enriches retrieved documents, and VideoRAG (Jeong et al., 2025) combines retrieved visual and textual evidence. AdaVideoRAG (zhucun et al., 2025) adapts retrieval across text, visual, and graph stores. The VideoAgent systems use targeted frame acquisition (Wang et al., 2024) and searchable multimodal memory (Fan et al., 2024). GCAgent (Yeo et al., 2026a) organizes causal and temporal relations; WorldMM (Yeo et al., 2026b) retrieves from episodic, semantic, and visual memories. Video-EM (Wang et al., 2025b) selects temporally ordered episodic evidence. VideoLucy (Zuo et al., 2025) and Visual Agentic Memory (Li et al., 2026a) search hierarchical memories with coarse-to-fine backtracking and visual verification, respectively. MAGIC-Video (Li et al., 2026b) combines a multimodal memory graph with narrative chains for long-horizon retrieval. EgoGraph (Sun et al., 2026) encodes cross-entity temporal dependencies in a dynamic knowledge graph. ObjectStream (Dong et al., 2026) links latent objects across frames to preserve their histories and state changes. GROVE (Gong et al., 2026) consolidates observations into moments, episodes, and recurring patterns with retrieval at each temporal scale.

## 3 Online Memory Construction

**Segment-local construction.** We consider a continuous video stream divided into fixed-duration segments $\mathbf{x}_{1:T} = (x_1, \ldots, x_T)$, where a new segment becomes available every $\Delta$ seconds. Each $x_t$ contains the video observations and an aligned transcript supplied as input. The same transcripts are provided to all compared methods. We focus on a class $\mathcal{B}$ of segment-local memory builders. Each builder $b \in \mathcal{B}$ is fixed before any evaluation query is revealed and converts each segment independently into a reusable index record

$$m_t = b(x_t; \Theta),$$

where $\Theta$ denotes fixed model parameters and precomputed resources. The record $m_t$ may contain an index representation, temporal metadata, and pointers to observations used downstream. Under this segment-local restriction, $m_t$ depends only on $x_t$ and $\Theta$. The complete index for the first $T$ segments is

$$\mathcal{M}_T = \{m_1, \ldots, m_T\}.$$

During streaming, some of these records may still be under construction; segment arrival alone does not make a record available for retrieval. In our setting, *online* refers to the requirement that records be constructed while new segments continue to arrive.

**Construction stability.** Let $C_t(b)$ be the service time required by builder $b$ to construct $m_t$ on a fixed compute configuration, given the video observations and supplied transcript. Speech recognition is outside this cost; any transcript processing performed by $b$ is included. For a single work-conserving FIFO builder, let $V_t$ denote the outstanding construction workload, in seconds of service, immediately after segment $x_t$ arrives. The workload evolves as

$$V_{t+1} = \max\{0, V_t - \Delta\} + C_{t+1}(b), \tag{1}$$

with $V_1 = C_1(b)$. We use

$$\rho_t(b) = \frac{V_t}{\Delta} \tag{2}$$

to report the normalized construction workload. Unlike the per-segment ratio $C_t(b)/\Delta$, $\rho_t$ also accounts for unfinished work carried over from earlier segments. It captures the total outstanding work, including the current segment's service time and any accumulated backlog. When $\rho_t > 1$, the outstanding work cannot be completed before the next segment arrives. An isolated crossing can be cleared by shorter subsequent service times, whereas sustained growth in $\rho_t$ indicates that the builder is falling progressively further behind the video stream, delaying the availability of recent observations for retrieval.

**End-to-end pipeline.** Given a query $q$, let $\mathcal{C}_T \subseteq \mathcal{M}_T$ denote the candidate records eligible under the benchmark's temporal rules. We evaluate retrieval and QA using completed indexes to isolate representation quality from construction-induced evidence unavailability, giving methods access to the same eligible observations. Construction delay is assessed separately using directly measured on-device service times. Thus, $\mathcal{C}_T$ is not filtered by construction completion times. Streaming deployment would require completion before query arrival. A retriever $R_b$ associated with builder $b$ selects up to $K$ segment indices:

$$\mathcal{S}_q = R_b(q, \mathcal{C}_T; K).$$

The corresponding observations and temporal metadata are assembled into an evidence context $\mathcal{E}_q$ in temporal order, and an answer model $A$ produces

$$\mathcal{E}_q = \mathrm{Assemble}(\mathcal{S}_q), \qquad \hat{y}_q = A(q, \mathcal{E}_q).$$

Retrieval and QA quality measure the downstream utility of the memory, alongside the construction stability defined above. This pipeline serves to evaluate memory construction; answer generation is outside the construction workload $C_t(b)$.

# 4 LOREMEM: LEXICON-ORDINAL RETRIEVAL MEMORY

**LoreMem** is a non-parametric memory representation for making an incoming video segment searchable without generating a caption or knowing future queries. As illustrated in Figure 2, the central component is a fixed linguistic lexicon that acts as a common set of coordinates. Video frames, the accompanying dialogue, and a later query are independently projected onto these coordinates in their respective encoder spaces. Each segment stores only the highest-ranked lexicon entries for its visual and textual observations, and retrieval compares these sparse ordinal memories with the lexicon entries activated by the query.

## 4.1 FIXED LEXICON COORDINATES

Let $\mathcal{V} = \{v_1, \ldots, v_L\}$ be a fixed lexicon of $L$ entries. The lexicon is determined before processing any target video or evaluation query and remains unchanged thereafter. We use two frozen encoders, one for visual observations and one for text. For each modality $\mu \in \{\mathrm{vis}, \mathrm{txt}\}$, we precompute a normalized anchor for every lexicon entry,

$$e^{\mu}_v = \mathrm{norm}(g^{\mu}(v)), \qquad E^{\mu} = \left[e^{\mu}_{v_1}, \ldots, e^{\mu}_{v_L}\right]^{\top},$$

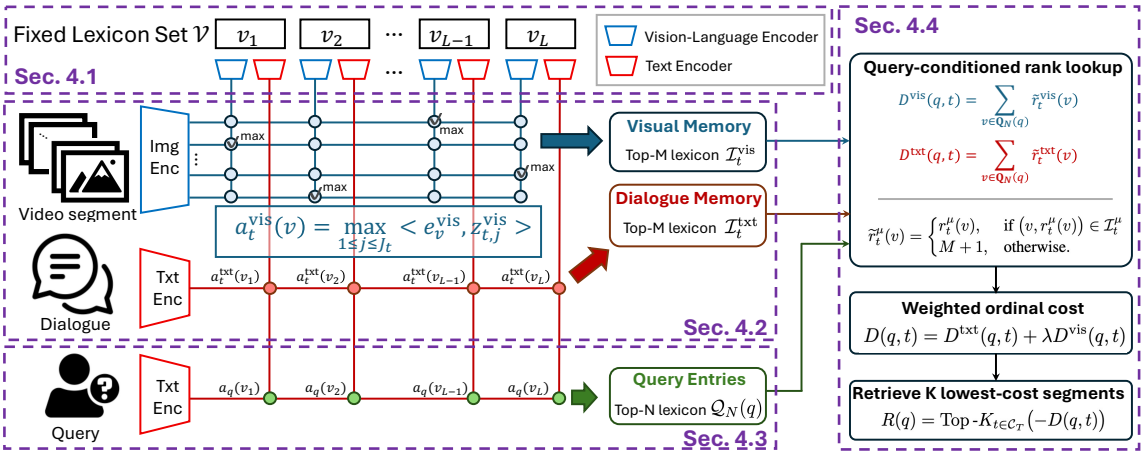



Figure 2: **Overview of `LoreMem`.** A fixed lexicon provides shared concept identities across frozen visual and text encoders. Each segment stores its top-$M$ visual and dialogue lexicon ranks, while the query selects its top-$N$ text concepts. Retrieval returns the $K$ segments with the lowest weighted rank cost across modalities.

where $g^{\mu}$ denotes the language-side encoder associated with modality $\mu$ and $\text{norm}(z) = z/\|z\|_2$. The two anchor tables may have different embedding dimensions and similarity distributions. Their rows refer to the same lexicon entries, so the lexicon identities align the modalities.

### 4.2 Visual and Dialogue Memory Construction

Consider a video segment $x_t$ with sampled frames $\mathcal{F}_t = \{I_{t,1}, \ldots, I_{t,J_t}\}$ and dialogue $d_t$. For a frame $I_{t,j}$, the frozen vision encoder $f^{\text{vis}}$ produces

$$z_{t,j}^{\text{vis}} = \text{norm}\big(f^{\text{vis}}(I_{t,j})\big)\,.$$

We compare each frame with every visual lexicon anchor. Because a concept need only be visible in one frame to characterize the segment, its segment-level visual affinity is the strongest frame-level match:

$$a_t^{\text{vis}}(v) = \max_{1 \le j \le J_t} \big\langle e_v^{\text{vis}}, z_{t,j}^{\text{vis}} \big\rangle\,. \tag{3}$$

Thus, the maximum is taken independently for each $v$: different lexicon entries may obtain their highest affinities from different frames.

For the dialogue, the frozen text encoder $f^{\text{txt}}$ produces a single normalized representation,

$$z_t^{\text{txt}} = \text{norm}\big(f^{\text{txt}}(d_t)\big)\,, \qquad a_t^{\text{txt}}(v) = \big\langle e_v^{\text{txt}}, z_t^{\text{txt}} \big\rangle\,. \tag{4}$$

The visual and textual affinities are ranked separately. For $\mu \in \{\text{vis}, \text{txt}\}$, define

$$r_t^{\mu}(v) = 1 + \sum_{u \in \mathcal{V} \setminus \{v\}} \mathbf{1}[a_t^{\mu}(u) > a_t^{\mu}(v)]\,,$$

where exact ties are resolved using a fixed lexicon order. Rank one therefore denotes the strongest match. We retain only the top-$M$ lexicon–rank pairs in each modality:

$$\mathcal{I}_t^{\mu} = \big\{ \big(v, r_t^{\mu}(v)\big) \;\big|\; r_t^{\mu}(v) \le M \big\}\,.$$

The memory record for segment $t$ consists of $\mathcal{I}_t^{\text{vis}}$ and $\mathcal{I}_t^{\text{txt}}$, together with its temporal metadata. Each modality retains its top-$M$ lexicon entries and their ranks, and the fixed anchor tables are shared across all segments. The record depends only on the current segment and these fixed anchors, making construction causal and independent of any future query.

### 4.3 Query Projection

At retrieval time, a query $q$ is projected using the same text encoder and text lexicon anchors as the dialogue:

$$z^q = \mathrm{norm}\big(f^{\mathrm{txt}}(q)\big)\,, \qquad a^q(v) = \big\langle e_v^{\mathrm{txt}}, z^q \big\rangle\,.$$

Let $r^q(v)$ be the rank induced by $a^q(v)$, with rank one again denoting the largest affinity. We use only the $N$ lexicon entries most relevant to the query,

$$\mathcal{Q}_N(q) = \{v \in \mathcal{V} \mid r^q(v) \le N\}\,.$$

This selection maps the query to a sparse semantic intent expressed in the same lexicon identities used by both segment memories.

### 4.4 Weighted Rank-Sum Retrieval

Retrieval evaluates how highly each modality of segment $t$ ranks the query-relevant lexicon entries. Since only the top-$M$ pairs are stored, we define the lookup rank

$$\widetilde{r}_t^{\mu}(v) = \begin{cases} r_t^{\mu}(v), & \text{if } \big(v, r_t^{\mu}(v)\big) \in \mathcal{I}_t^{\mu}, \\ M+1, & \text{otherwise.} \end{cases}$$

In other words, a query lexicon entry absent from the modality's top-$M$ memory is assigned the worst possible lookup rank, $M+1$. The modality-specific rank cost is

$$D^{\mu}(q,t) = \sum_{v \in \mathcal{Q}_N(q)} \widetilde{r}_t^{\mu}(v), \qquad \mu \in \{\mathrm{vis}, \mathrm{txt}\}. \tag{5}$$

Each cost lies in $[N, N(M+1)]$, and a smaller value indicates stronger agreement with the query. Finally, we combine dialogue and visual evidence using a single visual weight $\lambda$:

$$D(q,t) = D^{\mathrm{txt}}(q,t) + \lambda D^{\mathrm{vis}}(q,t), \qquad \lambda \ge 0. \tag{6}$$

Given the candidate memory records $\mathcal{C}_T$ available at query time, retrieval returns the $K$ segments with the smallest fused cost,

$$R(q) = \mathrm{Top}\text{-}K_{t \in \mathcal{C}_T}\big(-D(q,t)\big).$$

The rank representation removes the need to calibrate the raw similarity scales of the visual and text encoders. The coefficient $\lambda$ controls only the relative contribution of the two ordinal costs: $\lambda = 0$ yields dialogue-only retrieval, whereas a larger value gives greater influence to visual evidence.

## 5 Experiments

### 5.1 Experimental Setup

**Datasets.** We evaluate on four long-form video QA benchmarks: EgoLifeQA (Yang et al., 2025) covers daily experiences in multi-day egocentric recordings; Ego-R1 (Tian et al., 2026) targets reasoning over ultra-long egocentric video; SuperMemory-VQA (Alam et al., 2026) tests long-horizon memory of visual events and conversations; and TeleEgo (Yan et al., 2025) evaluates memory, understanding, and cross-memory reasoning in everyday settings. Further dataset and evaluation details are provided in the Appendix A.

**Baselines.** We compare **`LoreMem`** with four baseline families. *No-memory* methods answer directly with GPT-5.4-mini, Qwen3-VL-8B (Bai et al., 2025), Time-R1 (Wang et al., 2025a), or Video-RTS (Wang et al., 2025c). *Visual-embedding* memories retrieve with OpenCLIP (Cherti et al., 2023), Perception Encoder (Bolya et al., 2025), MobileCLIP (Vasu et al., 2024), or SigLIP2 (Tschannen et al., 2025) features. *Caption-style* memory indexes captions generated from video and dialogue by Qwen3-VL-4B (Bai et al., 2025). *Agent-based and RAG-style* systems include EgoRAG (Yang et al., 2025), Video-RAG (Luo et al., 2025), and LightRAG (Guo et al., 2025).

| Approach | Memory Builder | Generator | EgoLifeQA | Ego-R1 | SuperMemory-VQA | TeleEgo | Average |
|---|---|---|---|---|---|---|---|
| No-Memory | - | GPT-5.4-mini | 35.8 | 37.3 | 41.9 | 50.1 | 41.3 |
| | - | Qwen3-VL-8B | 31.2 | 27.7 | 37.1 | 50.8 | 36.7 |
| | - | Time-R1 | 34.2 | 32.7 | 37.5 | 44.8 | 37.3 |
| | - | Video-RTS | 35.8 | 33.3 | 38.6 | 47.2 | 38.7 |
| Visual Embedding | OpenCLIP (86M) | Qwen3-VL-8B | 37.6 | 35.0 | 37.1 | <u>53.6</u> | 40.8 |
| | Perception Encoder (86M) | Qwen3-VL-8B | 45.2 | 40.0 | 38.3 | 52.7 | 44.1 |
| | MobileCLIP (86M) | Qwen3-VL-8B | 42.2 | 32.7 | 37.9 | 52.2 | 41.3 |
| | SigLIP2 (86M) | Qwen3-VL-8B | 39.4 | 33.7 | 37.9 | 52.8 | 41.0 |
| Caption-style | Qwen3-VL-4B | Qwen3-VL-8B | 40.6 | 37.0 | <u>44.2</u> | **54.7** | 44.1 |
| EgoRAG | EgoGPT-7B | Qwen3-VL-8B | 45.2 | 49.0 | 38.8 | 53.4 | 46.6 |
| Video-RAG | - | Qwen3-VL-8B | <u>48.0</u> | <u>54.3</u> | 41.8 | 47.5 | <u>47.9</u> |
| LightRAG | GPT-5.2 | Qwen3-VL-8B | 37.1 | 39.7 | 42.9 | 48.2 | 42.0 |
| **LoreMem** | Perception Encoder (86M) | Qwen3-VL-8B | **49.4** | **56.0** | **48.0** | **54.7** | **52.0** |

Table 1: **Main results on long-form video question answering.** All scores are QA accuracy (%). All methods use 0.5 fps and a 300-frame answer budget: the latest frames for no-memory methods, or 20 retrieved 30-second segments for retrieval-based methods. Bold and underline indicate the best and second-best distinct scores in each column (See Appendix B for retrieval results).

**Implementation details.** All methods use a 300-frame answer budget at 0.5 fps and the same provided transcripts. No-memory methods take the latest frames; retrieval methods select $K = 20$ 30-second segments, restore temporal order, and pass their original frames and transcripts to Qwen3-VL-8B (See Appendix F for prompt template). This shared answer model uses greedy decoding without token sampling. We report single-run QA accuracies and do not report variability across decoding seeds. Query-agnostic memories are built before evaluation questions are exposed and reused. QA uses completed indexes under benchmark temporal rules; the separate construction-workload evaluation does not restrict candidates. All text-retrieval methods, including **LoreMem**, caption-style memory, and Video-RAG, use the same frozen text encoder: `BGE-large-en-v1.5` (Xiao et al., 2024), with multilingual `BGE-M3` (Chen et al., 2024) used on TeleEgo. **LoreMem** uses frozen `PE-Core-B16` (Bolya et al., 2025) for visual encoding and precomputes anchors for 65,383 English WordNet entries[1] in each encoder's native space. The same English lexicon is used for TeleEgo's Chinese queries and dialogues. Unless otherwise stated, **LoreMem** uses max frame-to-lexicon pooling, per-modality memory support $M = 1{,}000$, query support $N = 100$, and visual weight $\lambda = 0.1$ across all four benchmarks. For measuring on-device memory construction times, we use Galaxy S26 Ultra with Qualcomm Snapdragon 8 Elite Gen 5 for Galaxy (See Appendix C).

### 5.2 Main Results

**Overall performance.** Table 1 shows that **LoreMem** achieves the highest average QA accuracy (52.0), outperforming Video-RAG (47.9) by **4.1 percentage points (pp)** and both Perception Encoder visual memory and Qwen3-VL-4B caption memory (44.1 each) by **7.9 pp**. It ranks first on EgoLifeQA, Ego-R1, and SuperMemory-VQA, and ties caption memory on TeleEgo. On TeleEgo, **LoreMem** achieves 54.7 accuracy on Chinese queries and dialogues using the same English lexicon with multilingual `BGE-M3`. This result supports the cross-lingual applicability of the lexicon representation when paired with a multilingual text encoder. These results show that ordinal lexicon memory improves average downstream QA accuracy without generating segment captions.

[1] `https://github.com/globalwordnet/english-wordnet`

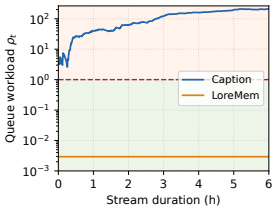



Figure 3: **On-device construction workload.** Workload $\rho_t$ over six hours of 30-second arrivals using directly measured construction times.

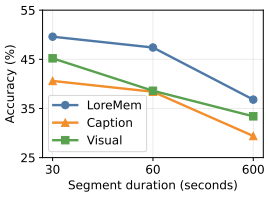



Figure 4: **Segment duration.** EgoLifeQA accuracy at a fixed 300-frame budget. Durations of 30, 60, and 600 seconds use top-20, top-10, and top-1 retrieval, respectively.

| Visual Enc. | Text Enc. | Acc. | N@20 |
|---|---|---|---|
| PE-Tiny-384 | BGE-Base | 45.8 | 12.6 |
| PE-Tiny-384 | BGE-Large | 47.8 | 14.8 |
| PE-Base | BGE-Base | 46.2 | 12.8 |
| PE-Base | BGE-Large | **49.4** | **15.1** |

Table 2: **Encoder size.** EgoLifeQA accuracy and nDCG@20 (N@20) for frozen visual/text encoder pairs. PE denotes Perception Encoder. Larger encoders improve both metrics; bold indicates the best result for each metric.

| Vis. Pool | Score | Acc. | N@20 |
|---|---|---|---|
| Mean | Cosine | 41.0 | 6.0 |
| Max | Cosine | 41.6 | 6.1 |
| Mean | Rank | 49.3 | 14.9 |
| Max | Rank | **49.4** | **15.1** |

Table 3: **Pooling and scoring.** EgoLifeQA QA accuracy and nDCG@20 (N@20) for mean/max visual pooling and cosine/rank retrieval score. Bold marks the best scores in each column.

**Queue stability.** Figure 3 reports construction workload over six hours of 30-second arrivals, using service times directly measured on a Galaxy S26 Ultra. Workload $\rho_t = V_t/\Delta$ includes unfinished work from earlier arrivals, as defined in Equation 2. With these measured service times, the evaluated caption configuration accumulates construction debt, while **LoreMem** remains below $\rho_t = 1$ without accumulating backlog. One might reduce the maximum caption length to lower the construction workload and improve queue stability; however, Appendix D.3 shows that this comes at a substantial cost to memory quality: halving the caption output cap from 1024 to 512 tokens lowers EgoLifeQA accuracy from $40.6\%$ to $35.0\%$.

### 5.3 Ablation Studies and Analysis

We further analyze **LoreMem**'s key design choices and retrieval behavior. Additional quantitative and qualitative analyses are provided in Appendix D.

**Segment duration.** Increasing segment duration could stabilize an overloaded caption builder by allowing more time between arrivals, provided that construction time grows slowly enough to bring the offered load below one. Figure 4 examines the QA cost of this alternative on EgoLifeQA at a fixed budget of 300 frames sampled at 0.5 fps: 30-, 60-, and 600-second segments use top-20, top-10, and top-1 retrieval, respectively. Accuracy decreases for all three memories as duration increases, most sharply at 600 seconds, while **LoreMem** remains the most accurate throughout. Under the fixed budget, longer segments combine more events into each retrieval unit and reduce the number of independently selectable evidence intervals, jointly limiting evidence granularity and retrieval flexibility. Although this analysis does not establish which durations would stabilize caption construction, it shows that coarser segmentation carries a QA cost. Together with Figure 3, these results show that the more accurate 30-second granularity is compatible with a stable queue for **LoreMem** under the directly measured construction times.

**Encoder scale.** Table 2 varies visual and text encoder sizes on EgoLifeQA. Scaling either encoder improves QA accuracy and N@20, with larger gains from text encoder scaling in the tested configurations. Replacing BGE-Base with BGE-Large improves accuracy by **2.0–3.2 percentage points (pp)** across the two visual encoders, compared with **0.4–1.6 pp** from replacing PE-Tiny-384 with PE-Base. N@20 follows the same pattern, and the larger pair reaches 49.4 accuracy and 15.1 N@20.

**Visual pooling and retrieval score.** Table 3 compares mean/max pooling of frame-lexicon similarities and raw cosine/ordinal rank scoring. For each lexicon entry, max pooling takes the largest cosine similarity across the segment's sampled frames from Equation 3; mean pooling uses the arithmetic mean over the same frames. Pooling precedes lexicon ranking. In the cosine variant, we directly sum the pre-ranking lexicon

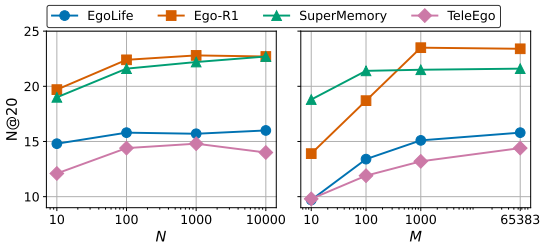



Figure 5: **Hyperparameter sensitivity.** Temporal-evidence nDCG@20 as query support $N$ (left) and per-modality segment support $M$ (right) vary across four benchmarks.

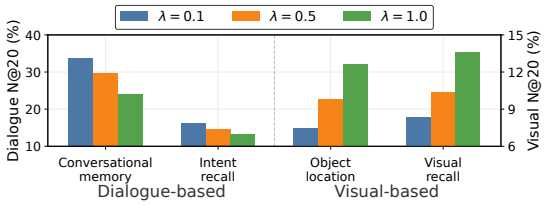



Figure 6: **Effect of visual weight $\lambda$.** N@20 by SuperMemory-VQA task type for $\lambda \in \{0.1, 0.5, 1.0\}$. Dialogue-based tasks use the left axis (Dialogue N@20); visual-based tasks use the right axis (Visual N@20).

affinities $a_t^{\mu}(v)$ from Equation 3 and 4 over the query-selected entries $\mathcal{Q}_N(q)$ in place of the rank values in Equation 5. We combine these modality-specific sums using the same visual weight $\lambda$ as in Equation 6 and retrieve the segments with the largest scores. With max pooling, rank scoring improves accuracy (**41.6** $\rightarrow$ **49.4**) and N@20 (**6.1** $\rightarrow$ **15.1**). Mean pooling with rank scoring already reaches 49.3 accuracy and 14.9 N@20, so the additional gain from max pooling is small. This ablation supports ordinal scoring in the tested setting, but does not isolate whether its gain comes from cross-modal scale differences or other properties of the score distributions.

**Memory and query support.** Figure 5 examines the coverage-sparsity trade-off controlled by per-modality segment support $M$ and the query selectivity controlled by $N$. Increasing $M$ from 10 to 1,000 improves N@20 on all four benchmarks. Retaining the full lexicon yields further gains on EgoLifeQA and TeleEgo, but little change on SuperMemory-VQA and a slight decline on Ego-R1. Thus, $M = 1{,}000$ retains useful concept coverage in a sparse memory. For $N$, moving from 10 to 100 improves all four benchmarks, whereas further expansion gives smaller, benchmark-dependent changes; TeleEgo declines at $N = 10{,}000$. These trends support the shared defaults $M = 1{,}000$ and $N = 100$ as a balance between memory coverage and query selectivity. The same setting is used across benchmarks, and the sweeps show that it need not maximize each benchmark individually.

**Visual-dialogue trade-off.** Figure 6 breaks down SuperMemory-VQA by question type for $\lambda \in \{0.1, 0.5, 1.0\}$. Increasing the visual weight raises N@20 for object-location and visual-recall questions, but lowers it for conversational-memory and intent-recall questions. These opposing trends show that the preferred modality balance depends on the evidence a task requires. The default $\lambda = 0.1$ favors dialogue-based retrieval, while larger values favor visually grounded tasks; no tested value maximizes every task type. Because dialogue and visual tasks use different vertical scales in the figure, comparisons should follow the changes within each task rather than the bar heights across task groups.

## 6 Conclusion

Existing video-memory evaluations emphasize downstream quality, giving limited attention to whether construction keeps pace with incoming video. We formulated *online memory construction* as a joint quality-stability problem under a fixed compute budget, reflecting the needs of assistants queried during ongoing recording. We also introduced **`LoreMem`** considering this setting, indexing visual observations and dialogue through sparse lexicon identities and ranks without caption generation. On completed indexes across four benchmarks, **`LoreMem`** improves average QA accuracy over Qwen3-VL-4B caption memory by **7.9 percentage points**. Separately, workload evaluation using directly measured on-device construction times shows **no backlog accumulation** over six hours of 30-second arrivals. See Appendix E for limitation.

## 7 AI USE STATEMENT

Generative AI tools were used solely for language editing purposes, including grammar correction and refinement of English wording. These tools did not contribute to the study conception, methodology development, experimental design, experimentation, data analysis, interpretation of findings, or preparation of technical content. All AI-assisted revisions were thoroughly reviewed and approved by the authors.

## 8 ETHICS STATEMENT

Persistent memory for egocentric and wearable video raises important privacy and misuse concerns because continuous recordings may capture bystanders, private spaces, conversations, locations, and other sensitive information. Although this work focuses on the technical problem of efficient retrieval and does not advocate unrestricted recording, a practical deployment should obtain appropriate consent, provide clear recording indicators and controls, minimize data collection and retention, and protect both memory records and referenced media through access control and secure storage. Retrieved evidence may also reflect biases or errors in the frozen encoders and can enable sensitive inferences beyond a user's original intent; consequently, high-stakes uses require careful auditing, user-visible provenance, and human oversight. Caption-free indexing does not eliminate these risks: ordinal records and the media to which they refer can still reveal sensitive information. Any use of such a system must therefore comply with applicable privacy, data-protection, and research-ethics requirements.

## 9 REPRODUCIBILITY STATEMENT

The paper specifies the queue-stability metric under online memory construction setup and details of **`LoreMem`**. The code is not publicly available at the current submission stage. Upon acceptance, we will release the implementation together with preprocessing and indexing scripts, configuration files, prompt templates, and evaluation code required to reproduce the main results. We will also document software and hardware dependencies, random seeds, and instructions for obtaining the benchmark data under their respective licenses and terms of use.

## REFERENCES

Samiul Alam, Shakhrul Iman Siam, Michael J Proulx, James Fort, Richard Newcombe, Hyo Jin Kim, and Mi Zhang. Supermemory-vqa: An egocentric visual question-answering benchmark for long-horizon memory. *ArXiv Preprint*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *ArXiv Preprint*, 2025.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, et al. Perception encoder: The best visual embeddings are not at the output of the network. In *Proceedings of Advances in Neural Information Processing Systems*, 2025.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (Findings)*, 2024.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive

language-image learning. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2023.

Anxhelo Diko, Tinghuai Wang, Wassim Swaileh, Shiyan Sun, and Ioannis Patras. Rewind: Understanding long videos with instructed learnable memory. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2025.

Mingkang Dong, Muxin Pu, Jie Li, Bohan Guo, Songruo Chen, Bin Ren, Xu Zheng, Chen Zhao, Tianwen Qian, Mohamed Elhoseiny, et al. Objectstream: Latent objects as memory anchors for streaming video understanding. *ArXiv Preprint*, 2026.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *Proceedings of European Conference on Computer Vision*, 2024.

Sitong Gong, Caixin Kang, Tianyu Yan, Guo Chen, Bo Zheng, Kaipeng Zhang, Yunzhi Zhuge, Xiang Ruan, Huchuan Lu, and Yifei Huang. Grove: Growing and reasoning over temporally stratified memory from streaming video experience. *ArXiv Preprint*, 2026.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. In *Proceedings of Empirical Methods in Natural Language Processing (Findings)*, 2025.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2024.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2024.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. Videorag: Retrieval-augmented generation over video corpus. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (Findings)*, 2025.

Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. Language repository for long video understanding. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (Findings)*, 2025.

Aiden Yiliu Li, Nels Numan, and Anthony Steed. Visual agentic memory: Enabling online long video understanding via online indexing, hierarchical memory, and agentic retrieval. *ArXiv Preprint*, 2026a.

Jiazheng Li, Chi-Hao Wu, Yunze Liu, Kaize Ding, Jundong Li, and Chuxu Zhang. Bridging modalities, spanning time: Structured memory for ultra-long agentic video reasoning. *ArXiv Preprint*, 2026b.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, et al. Video-rag: Visually-aligned retrieval-augmented long video comprehension. In *Proceedings of Advances in Neural Information Processing Systems*, 2025.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2025.

Junxiao Shen, John J Dudley, and Per Ola Kristensson. Encode-store-retrieve: Augmenting human memory through language-encoded egocentric perception. In *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, 2024.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of International Conference on Machine Learning*, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2024.

Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Shitong Sun, Ke Han, Yukai Huang, Weitong Cai, and Jifei Song. Egograph: Temporal knowledge graph for egocentric video understanding. *ArXiv Preprint*, 2026.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Agentic chain-of-tool-thought for ultra-long egocentric video reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *ArXiv Preprint*, 2025.

Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. Mobileclip: Fast image-text models through multi-modal reinforced training. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2024.

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

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. In *Proceedings of Advances in Neural Information Processing Systems*, 2025a.

Ying Wang, Yanlai Yang, and Mengye Ren. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos. *ArXiv Preprint*, 2023.

Yun Wang, Long Zhang, Jingren Liu, Jiaqi Yan, Zhanjie Zhang, Jiahao Zheng, Ao Ma, Run Ling, Xun Yang, Dapeng Wu, et al. Video-em: Event-centric episodic memory for long-form video understanding. *ArXiv Preprint*, 2025b.

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *Proceedings of Empirical Methods in Natural Language Processing*, 2025c.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2025d.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

Jiaqi Yan, Ruilong Ren, Jingren Liu, Shuning Xu, Ling Wang, Yiheng Wang, Xinlin Zhong, Yun Wang, Long Zhang, Xiangyu Chen, et al. Teleego: Benchmarking egocentric ai assistants in the wild. *ArXiv Preprint*, 2025.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2025.

Jeong Hun Yeo, Sangyun Chung, Sungjune Park, Dae Hoe Kim, Jinyoung Moon, and Yong Man Ro. Gcagent: Long-video understanding via schematic and narrative episodic memory. *IEEE Transactions on Multimedia*, 2026a.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2026b.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. In *Proceedings of Empirical Methods in Natural Language Processing*, 2024.

Xue zhucun, Jiangning Zhang, Xie Xurong, Yuxuan Cai, Yong Liu, Xiangtai Li, and Dacheng Tao. Adavideorag: Omni-contextual adaptive retrieval-augmented efficient long video understanding. In *Proceedings of Advances in Neural Information Processing Systems*, 2025.

Jialong Zuo, Yongtai Deng, Lingdong Kong, Jingkang Yang, Rui Jin, Yiwei Zhang, Nong Sang, Liang Pan, Ziwei Liu, and Changxin Gao. Videolucy: Deep memory backtracking for long video understanding. In *Proceedings of Advances in Neural Information Processing Systems*, 2025.

# A Dataset and Evaluation Details

## A.1 Datasets

Table 4 summarizes the question counts and total recording durations of the evaluated subsets.

**EgoLifeQA.** EgoLifeQA (Yang et al., 2025) is built on the EgoLife recordings, in which six participants captured a week of daily activities using wearable cameras, producing approximately 300 hours of video. Its life-assistant questions include recalling past events, locating objects, and reasoning about habits and routines. Relevant evidence may be brief and interspersed with long periods of routine activity, and spoken interactions can provide information that is not recoverable from frames alone.

| Dataset | Num. queries | video duration (h) |
|---|---|---|
| EgoLifeQA | 500 | 44.3 |
| Ego-R1 | 300 | 44.3 |
| SuperMemory-VQA | 4,853 | 52.9 |
| TeleEgo | 1,743 | 71.9 |

Table 4: Evaluation scale for each benchmark. TeleEgo includes only single-choice questions used in our experiments.

**Ego-R1.** Ego-R1 (Tian et al., 2026) emphasizes temporal dependencies, causal reasoning, and the combination of evidence from distant moments. Whereas EgoLifeQA emphasizes episodic recall and daily-life assistance, Ego-R1 places greater weight on reasoning across retrieved events.

**SuperMemory-VQA.** SuperMemory-VQA (Alam et al., 2026) contains 52.9 hours of recordings from ten participants wearing Meta Aria glasses. The dataset provides synchronized video and audio transcriptions alongside additional sensor streams. Its 4,853 multiple-choice questions span object and location memory, intent recall, visual scene recall, timeline reconstruction, conversational memory, and in-context retrieval. Each question includes an explicit unanswerable option, allowing evaluation of whether an answer is supported by the available evidence. The diversity of question types makes it possible to examine retrieval behavior separately for visually grounded and conversation-based tasks.

**TeleEgo.** TeleEgo (Yan et al., 2025) includes memory, understanding, and cross-memory reasoning questions in everyday egocentric settings. Its Chinese queries introduce a language difference from the English lexicon used by **`LoreMem`**. This benchmark therefore complements the other evaluations by examining retrieval when the query language and the lexicon language differ. We evaluate only the 1,743 single-choice multiple-choice (`mc_single`) questions across all five participants, excluding multiple-choice questions with multiple answers, binary questions, and open-ended questions.

## A.2 Metrics

**Temporal-evidence retrieval.** We evaluate retrieved segments using nDCG@20. A segment is considered relevant if its temporal interval overlaps any annotated evidence interval by a positive duration. Relevance is binary, and the gain at rank $i$ is discounted by $1/\log_2(i+1)$. DCG is normalized by the ideal ranking formed from all relevant segments in the candidate set. Only questions with temporal evidence annotations are included in retrieval evaluation, whereas all benchmark questions remain eligible for QA evaluation.

**Question answering.** For single-choice questions, QA accuracy is the percentage of questions for which the predicted option matches the ground-truth answer. Unparseable or failed responses are counted as incorrect. For questions with an unanswerable option, correctness is determined by the annotated option. Cross-benchmark averages are unweighted means computed from the unrounded per-benchmark scores.

| Approach | Memory Builder | Generator | EgoLifeQA Acc | EgoLifeQA N@20 | Ego-R1 Acc | Ego-R1 N@20 | SuperMemory-VQA Acc | SuperMemory-VQA N@20 | TeleEgo Acc | TeleEgo N@20 | Average Acc | Average N@20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Visual Embedding | OpenCLIP | Qwen3-VL-8B | 37.6 | 4.0 | 35.0 | 2.0 | 37.1 | 9.2 | <u>53.6</u> | 3.5 | 40.8 | 4.7 |
| | Perception Encoder | Qwen3-VL-8B | <u>45.2</u> | <u>6.3</u> | <u>40.0</u> | <u>3.4</u> | 38.3 | 14.2 | 52.7 | 3.4 | <u>44.1</u> | 6.8 |
| | MobileCLIP | Qwen3-VL-8B | 42.2 | 4.6 | 32.7 | 1.9 | 37.9 | 12.2 | 52.2 | 4.4 | 41.3 | 5.8 |
| | SigLIP2 | Qwen3-VL-8B | 39.4 | 5.7 | 33.7 | 2.5 | 37.9 | 10.8 | 52.8 | 9.7 | 41.0 | 7.2 |
| Caption-style | Qwen3-VL-4B | Qwen3-VL-8B | 40.6 | 3.7 | 37.0 | 2.3 | <u>44.2</u> | <u>17.1</u> | **54.7** | **15.5** | <u>44.1</u> | <u>9.7</u> |
| **LoreMem** | Perception Encoder | Qwen3-VL-8B | **49.4** | **15.1** | **56.0** | **23.5** | **48.0** | **21.5** | **54.7** | <u>15.0</u> | **52.0** | **18.8** |

Table 5: **Temporal-evidence retrieval and question-answering results for methods with reported retrieval scores.** Acc. denotes question-answering accuracy and N@20 denotes nDCG@20. All methods retrieve 20 30-second segments sampled at 0.5 fps, for a 300-frame answer budget. Text retrieval in **LoreMem** and caption memory uses the same encoder per dataset: BGE-Large, or BGE-M3 for TeleEgo. Bold and underline indicate the best and second-best distinct scores among the methods shown in each column.

## B Temporal-Evidence Retrieval Results

Table 5 reports temporal-evidence nDCG@20 alongside QA accuracy for the six methods in Table 1 with reported retrieval scores: four visual-embedding memories, Qwen3-VL-4B caption memory, and **LoreMem**. All methods follow the evaluation protocol in Section 5.1. As defined in Appendix A.2, nDCG@20 evaluates the ranking of the top-20 retrieved segments using binary relevance determined by temporal overlap with annotated evidence. Retrieval evaluation uses only questions with temporal evidence annotations; QA accuracy uses all evaluated questions.

**Retrieval performance.** **LoreMem** achieves the highest average nDCG@20 among the methods in Table 5, reaching 18.8 compared with 7.2 for SigLIP2, the strongest visual-embedding baseline on average, and 9.7 for caption memory. These correspond to nDCG@20 gains of 11.6 and 9.1, respectively. **LoreMem** ranks first on EgoLifeQA, Ego-R1, and SuperMemory-VQA, with nDCG@20 scores of 15.1, 23.5, and 21.5. The largest gain over the strongest baseline per benchmark occurs on Ego-R1: 23.5 versus 3.4 for Perception Encoder. These results support ordinal lexicon memory's ability to retrieve temporally annotated evidence without generating segment captions.

**Relation to QA accuracy.** The average retrieval improvement accompanies higher average QA accuracy for **LoreMem** among these methods, but the two metrics need not rank methods identically. On TeleEgo, caption memory achieves slightly higher nDCG@20 than **LoreMem** (15.5 vs. 15.0), while both obtain 54.7 QA accuracy. Thus, the retrieval results complement the end-to-end QA comparison in Table 1; they assess the ranking of annotated evidence, while accuracy also depends on how the answer model uses the retrieved context.

## C On-Device Memory Construction Time

**Device and numerical precision.** We directly measure memory construction times on a Samsung Galaxy S26 Ultra powered by the Qualcomm Snapdragon 8 Elite Gen 5 for Galaxy. The Qwen3-VL-4B caption-memory builder uses `Q4_0` quantization. All other models used in the construction-cost comparison run in floating-point precision, including OpenCLIP, Perception Encoder, MobileCLIP, SigLIP2, and the BGE text encoder.

**Measurement scope.** We measure construction time per 30-second segment with 15 frames sampled at 0.5 fps and a supplied transcript. These timings provide the construction-cost coordinates in Figure 1; the QA scores come from the completed-index benchmark evaluation. The Qwen point and range summarize

| | EgoLifeQA | | Ego-R1 | | SuperMemory-VQA | | TeleEgo | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|
| Memory | Acc. | N@20 | Acc. | N@20 | Acc. | N@20 | Acc. | N@20 | Acc. | N@20 |
| Caption–Qwen3-VL-4B | 40.6 | 3.7 | 37.0 | 2.3 | 44.2 | 17.1 | 54.7 | 15.5 | 44.1 | 9.7 |
| Caption–GPT-5.2 | 46.2 | 12.3 | 55.0 | 23.3 | 46.1 | 22.7 | 56.2 | 20.8 | 50.9 | 19.8 |

Table 6: **Effect of the captioning model.** Acc. denotes QA accuracy and N@20 denotes nDCG@20. All memories retrieve 20 30-second segments sampled at 0.5 fps, for a 300-frame answer budget.

| Visual | Dialogue | **LoreMem** | Acc. |
|---|---|---|---|
| ✓ | ✗ | ✗ | 45.2 |
| ✗ | ✓ | ✗ | 48.0 |
| ✓ | ✓ | ✗ | 47.2 |
| ✓ | ✓ | ✓ | 49.4 |

Table 7: **Raw-score fusion and LoreMem on Ego-LifeQA.** Both multimodal rows use $\lambda = 0.1$. Acc. denotes QA accuracy (%).

the mean and min–max of measured construction times across segments. As defined in Section 3, speech recognition, retrieval, and answer generation are outside the memory-construction workload; transcripts are provided as input, and lexicon anchors are precomputed.

**Construction workload.** Figure 3 evaluates six hours of 30-second segment arrivals. We use the directly measured service times $C_t(b)$ in Equation 1 and report $\rho_t = V_t/\Delta$. This workload accounts for unfinished work from earlier segments as well as the current segment's construction time. Construction workload is evaluated separately from QA, which uses completed indexes under the benchmark's temporal rules.

# D Additional Analyses

## D.1 Effect of the Captioning Model

Table 6 examines how the captioning model affects caption-memory performance, with **LoreMem** included for reference. GPT-5.2 serves as a cloud-based caption-quality reference outside our on-device construction-cost comparison. All three memories follow the evaluation protocol in Section 5.1, using Qwen3-VL-8B for answering, the same text encoder per dataset, and the same retrieval budget. Replacing Qwen3-VL-4B with GPT-5.2 improves QA accuracy and nDCG@20 on all four benchmarks, raising their averages from 44.1 to 50.9 and from 9.7 to 19.8, respectively. These results show that caption memory benefits from a stronger captioning model, while **LoreMem** achieves competitive QA accuracy without generating segment captions.

## D.2 Direct Multimodal Fusion and Lexicon Projection

Table 7 compares single-modality retrieval, uncalibrated multimodal score fusion, and **LoreMem** on Ego-LifeQA. We reuse the PE-B visual-memory and Video-RAG dialogue-retrieval results from Table 1 as the visual-only and dialogue-only baselines, respectively. The direct-fusion baseline combines their segment scores as

$$S_{\text{fusion}}(q,t) = s^{\text{txt}}_{\text{BGE}}(q,t) + \lambda s^{\text{vis}}_{\text{PE}}(q,t), \qquad \lambda = 0.1, \tag{7}$$

where $s^{\text{txt}}_{\text{BGE}}$ and $s^{\text{vis}}_{\text{PE}}$ are the original segment similarity scores from the corresponding direct-retrieval baselines, without score normalization; the 20 highest-scoring segments are retrieved. **LoreMem** uses the same PE-B and BGE-Large encoders, projects observations and queries onto the fixed lexicon, and applies $\lambda = 0.1$ to the ordinal costs in Equation 6. The identical numerical coefficient does not equate the effective modality balance of the two scoring rules. All configurations use 30-second segments sampled at 0.5 fps, a 300-frame answer budget, and the same Qwen3-VL-8B answer model and segment evidence specified in Appendix F.1.

The evaluated raw-score fusion reaches 47.2% QA accuracy, which exceeds visual-only retrieval (45.2%) but falls below dialogue-only retrieval (48.0%) by 0.8 percentage points. **LoreMem** reaches 49.4%, outperforming direct fusion by 2.2 percentage points and dialogue-only retrieval by 1.4 percentage points. With the

same encoders and evidence budget, the combined lexicon-ordinal representation therefore yields higher QA accuracy than this raw-score fusion configuration. This comparison evaluates the complete representation; it does not separate the contributions of lexicon projection and ordinalization.

### D.3 Caption Output Budget and Construction Stability

To examine the accuracy cost of reducing caption-construction load, we halve the output cap from $1024$ to $512$ tokens while keeping the remaining protocol fixed (See Table 8).

EgoLifeQA accuracy drops from $40.6\%$ to $35.0\%$, a loss of $5.6$ percentage points, while **LoreMem** achieves $49.4\%$. Shortening captions therefore sacrifices accuracy. The construction-workload evaluation using directly measured times (See Figure 3) shows that the default caption builder accumulates backlog, while **LoreMem** maintains a stable queue. **LoreMem** achieves higher accuracy while avoiding autoregressive caption generation during construction.

| Memory | Max tokens | Acc. (%) |
|---|---|---|
| Caption–Qwen3-VL-4B | 1024 | 40.6 |
| | 512 | 35.0 |
| **LoreMem** | – | 49.4 |

Table 8: **Caption output budget on EgoLifeQA.** Max tokens caps caption generation per segment.

### D.4 Qualitative Analysis of Visual and Dialogue Memories

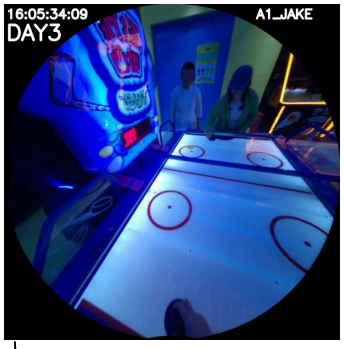
Each example pairs a query with a ground-truth evidence segment and the memory constructed from that segment before the query was known. We show a video frame, the accompanying dialogue, and the nine highest-ranked entries from each modality's memory, $\mathcal{I}_t^{\mathrm{vis}}$ and $\mathcal{I}_t^{\mathrm{txt}}$, to examine how the stored concepts relate to the information sought by the query. Figures 7, 8, and 9 illustrate relevant cues in visual memory, dialogue memory, and both modalities, respectively. The failure example in Figure 10 shows the ground-truth segment's memory, whose leading concepts emphasize a broader context than the query's specific target.

**Visual evidence with sparse dialogue.** Figure 7 shows the ground-truth arcade segment for the question, "Now Tasha and I are clapping hands in the arcade, what happened the last time we played games together?" The visual memory ranks *arcade* first, followed by related activity concepts such as *playing* (rank 3), *play* (rank 5), and *game* (rank 6). These entries capture the setting and activity visible around the air-hockey table and provide semantic links to the question. The accompanying dialogue is only "Jake: Yay!" Its highest-ranked text entries, including *yay*, *hooray*, and *hurrah*, express excitement but offer little information about where the event occurred or what game was being played. The visual coordinates capture the activity sought by the query, while the dialogue coordinates mainly capture the speaker's excitement. The original video preserves the event details needed to answer what happened during the earlier interaction.

**Dialogue evidence for a specific fact.** Figure 8 pairs the question, "How much is an ego coin worth in RMB?" with its ground-truth evidence segment. The frame shows a group gathered indoors, and the leading visual entries, such as *jamming*, *singalong*, and *rehearsal*, do not convey the coin's value. The dialogue explicitly discusses ego coins and states, "each coin is now 3 RMB." Its text memory ranks *coin*, *exchange*, and *ego* in the first three positions, with related entries such as *barter* and *coinage* also appearing among the top nine. These concepts connect the question to the conversation containing the answer. The precise value, 3 RMB, is available in the retrieved transcript: the lexicon ranks provide a compact index to the evidence, while the original dialogue preserves the factual detail used for answering.

**Visual and dialogue evidence for the same event.** Figure 9 shows the ground-truth segment for the question, "What did we end up eating the last time we ordered takeout?" The frame shows food and takeout containers spread across a table. The visual memory ranks *takeout* third and *eating* fourth, with additional meal-related entries such as *feasting* and *mealtime*. The dialogue discusses food from KFC and McDonald's,

Now Tasha and I are clapping hands in the **arcade**, what happened the last time we **played games** together?




| | |
|---|---|
| **1** | **arcade** |
| 2 | tiddlywinks |
| **3** | **playing** |
| 4 | playday |
| **5** | **play** |
| **6** | **game** |
| 7 | duckpins |
| 8 | duckpin |
| 9 | playbox |

Lexicon-rank pair visual set $\mathcal{I}_t^{\text{vis}}$

**Jake:** Yay!

| | |
|---|---|
| 1 | yay |
| 2 | hooray |
| 3 | hurrah |
| 4 | jake |
| 5 | congratulation |
| 6 | happy |
| 7 | delighted |
| 8 | pleased |
| 9 | excited |

Lexicon-rank pair text set $\mathcal{I}_t^{\text{txt}}$

Figure 7: **Visual memory captures an event under sparse dialogue.** The ground-truth arcade segment yields visual concepts related to the question, including *arcade*, *playing*, *play*, and *game*. The dialogue contains only an exclamation, so its leading text concepts mainly describe excitement. Each list displays the top nine lexicon entries; smaller ranks indicate stronger matches.

How much is an **ego coin** worth in RMB?




| | |
|---|---|
| 1 | jamming |
| 2 | fisheye |
| 3 | singalong |
| 4 | rehearsal |
| 5 | lesson |
| 6 | streoscopic |
| 7 | housewarming |
| 8 | karaoke |
| 9 | rehearse |

Lexicon-rank pair visual set $\mathcal{I}_t^{\text{vis}}$

**Jake:** If you get it right, you'll earn an ego coin. This ego coin can be exchanged for desserts, and to prepare for future financial freedom.
**Jake:** Right now, each coin is priced at
**Jake:** What's the current price?
**Shure:** 3 RMB
**Jake:** 3 RMB
**Jake:** Okay, each coin is now 3 RMB
**Jake:** And tomorrow, we'll see what the price is
**Jake:** So, our first question!

| | |
|---|---|
| **1** | **coin** |
| 2 | exchange |
| **3** | **ego** |
| 4 | tenner |
| 5 | barter |
| 6 | coinage |
| 7 | egotrip |
| 8 | bargain |
| 9 | dollarfish |

Lexicon-rank pair text set $\mathcal{I}_t^{\text{txt}}$

Figure 8: **Dialogue memory locates a fact absent from the visual scene.** In the ground-truth segment, the leading text entries *coin*, *exchange*, and *ego* identify the topic of the conversation, whose transcript gives the answer, 3 RMB. The leading visual entries describe other aspects of the indoor scene. Each list displays the top nine lexicon entries.

and its text memory ranks *takeout* second and *takeaway* eighth. The two independently constructed memories therefore share a query-relevant *takeout* coordinate, providing complementary visual and dialogue cues for the same meal. The original dialogue supplies the specific food description, "This is KFC roast duck," needed for answer generation. This case illustrates how a shared concept identity makes agreement between the modalities explicit in the stored representation.

**Mismatch between query and concept specificity.** Figure 10 pairs the question, "This time I returned to the claw machine, what happened the last time I was at the claw machine?" with its ground-truth evidence segment. The visual memory ranks *arcade*, *amusement*, and *playtime* in the first three positions. These con-

What did we end up **eating** the last time we ordered **takeout**?




| | |
|---|---|
| 1 | cookout |
| 2 | takoyaki |
| **3** | **takeout** |
| **4** | **eating** |
| 5 | overseating |
| 6 | picnic |
| 7 | feasting |
| 8 | dejeuner |
| 9 | mealtime |

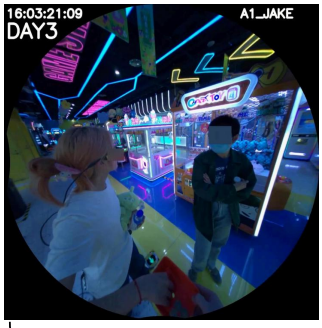
**Jake:** This is KFC roast duck.
**Jake:** Their KFC has nothing new, still the same old stuff.
**Alice:** This is Chinese-style boned duck.
**Lucia:** Yeah, the new items at KFC and McDonald's are quite different between here and abroad.

| | |
|---|---|
| 1 | chickenfeed |
| **2** | **takeout** |
| 3 | rotisseri |
| 4 | tunaburger |
| 5 | Entrée |
| 6 | poultry |
| 7 | chicken |
| **8** | **takeaway** |
| 9 | beefburger |

Lexicon-rank pair visual set $\mathcal{I}_t^{\text{vis}}$ — Lexicon-rank pair text set $\mathcal{I}_t^{\text{txt}}$

Figure 9: **Visual and dialogue memories jointly identify a takeout meal.** The ground-truth segment's visual memory ranks *takeout* and *eating* highly, while the text memory captures *takeout* and *takeaway* from the food-related conversation. Both modalities link the event to the question, and the original dialogue provides the specific description "KFC roast duck." Each list displays the top nine lexicon entries.

This time I returned to the **claw machine**, what happened the last time I was at the claw machine?




| | |
|---|---|
| 1 | arcade |
| 2 | amusement |
| 3 | playtime |
| 4 | aisle |
| 5 | jawbreaker |
| 6 | duckpins |
| 7 | playday |
| 8 | eggshake |
| 9 | fisheye |

**Jake:** Okay, there are still eight left.
**Alice:** The three-eyed guy.
**Katrina:** Let's play that one.
**Katrina:** Do you want to play the fighting game?

| | |
|---|---|
| 1 | eight |
| 2 | countdown |
| 3 | counting |
| 4 | eightsome |
| 5 | game |
| 6 | fighting |
| 7 | showdown |
| 8 | checkers |
| 9 | bunfight |

Lexicon-rank pair visual set $\mathcal{I}_t^{\text{vis}}$ — Lexicon-rank pair text set $\mathcal{I}_t^{\text{txt}}$

Figure 10: **Failure case: mismatch in concept specificity.** The ground-truth segment's leading visual entries describe the arcade setting, while the query asks about a specific claw-machine interaction. The text entries emphasize counting and a fighting-game suggestion. The top-nine lists show broad contextual cues with limited visible correspondence to the query's specific object.

cepts describe the scene's broad context, while the query singles out a particular object and interaction. The displayed entries do not make that object-specific cue explicit. In the dialogue memory, *eight*, *countdown*, and *counting* occupy ranks 1–3, with *game* and *fighting* at ranks 5 and 6. They reflect the conversation but offer little visible connection to the claw machine. The comparison with Figure 7 is informative: arcade-related concepts match the broader question about playing games, whereas the present query requires a more specific target. This failure case illustrates a mismatch between the specificity of the query and that of the ground-truth segment's leading memory concepts.

These examples show how the same representation exposes visual and dialogue cues with different relevance to a later query, consistent with the task-dependent modality trade-off in Figure 6. They also illustrate why the usefulness of a stored concept depends on its correspondence to the information the query seeks.

# E LIMITATIONS

**LoreMem**'s retrieval coverage is bounded by its fixed lexicon and the semantic resolution of its frozen encoders. Its global visual-dialogue weight may also be suboptimal for questions requiring different modality balances, and its ordinal representation does not explicitly model richer temporal relations between events. Future work could adapt modality weighting at query time, expand the lexicon without using evaluation queries, and introduce more explicit temporal structure while preserving causal construction.

# F PROMPT TEMPLATES AND ANSWER-TIME EVIDENCE

## F.1 ANSWER GENERATION PROMPTS

Table 9 presents the answer-generation prompt from WorldMM (Yeo et al., 2026b), used unchanged for Ego-LifeQA and Ego-R1. Each evidence card contains the segment time, dialogue, dataset-provided description, and video clip. The same description is used for a given segment across retrieval methods; neither captions generated by the caption-style baseline nor **LoreMem**'s ordinal lexicon records are inserted into this field. The prompt also provides the question time and answer options, and asks the model to return only the letter of the selected option. Its explicit day-offset rules support questions involving relative expressions such as "yesterday" and "the day before yesterday."

Table 10 presents the answer-generation prompt used for SuperMemory-VQA and TeleEgo. Each evidence card instead contains a segment identifier and time range, the aligned transcript, and sampled video frames. After reading all retrieved evidence cards, the model selects one of the provided options and returns only its letter. When the choices include an explicit insufficient-evidence option, the prompt instructs the model to select it if the retrieved evidence does not support an answer.

## F.2 CAPTION GENERATION PROMPT

Table 11 presents the caption-generation prompt shared across all four datasets. Given the video and its dialogue, the model is instructed to describe the observed events over time from the camera wearer's first-person perspective. The resulting description is stored by the caption-style memory for later retrieval.

**System prompt.** You are an AI assistant that answers questions based on video content, dialogue, and segment descriptions.

**User prompt template.** You will receive a series of video evidence cards. Each evidence card provides information in the following format:

**Evidence card format.**

- **time**: `{{time}}`
- **dialogue**: `{{dialogue}}`
- **description**: `{{description}}`
- **video**: `{{video clip}}`

The **time** field is written as `DAYX_HHMMSSFF_HHMMSSFF`, where `DAYX` indicates the day on which the events happened, and each `HHMMSSFF` specifies a timestamp with hours, minutes, seconds, and frame index.

**Video evidences.**

```
- time: {{time}}
- dialogue: {{dialogue}}
- description: {{description}}
- video: {{video clip}}

- time: {{time}}
- dialogue: {{dialogue}}
- description: {{description}}
- video: {{video clip}}

...
```

Using the information provided in the evidence cards, analyze the events and actions to answer the following question:

**Question:** `{question}`
**Question time:** `{question_time}`
**Options:** `{formatted_options}`

**Instructions.**

1. Use the evidence cards to determine the most appropriate answer.
2. Provide your answer as the letter corresponding to the correct option from the given choices.
3. To calculate time differences, assume that:
   - if the question refers to `yesterday`, the target time is one day before the `question_time`;
   - if the question refers to `the day before yesterday`, the target time is two days before the `question_time`;
   - if the question refers to `the day before the last event`, compute the difference between the `question_time` and the time of the last event;
   - pay attention to the `DAYX` field and calculate the correct day difference.
4. Respond with the letter only, for example `A`, `B`, `C`, or `D`.

Table 9: **EgoLifeQA and Ego-R1 answer-generation prompt.** Repeat the evidence-card block once per retrieved segment.

**System prompt.** You are an AI assistant that answers questions based on egocentric video memory.

**User prompt template.** You will receive a series of video evidences from an egocentric (first-person) recording. Each evidence card provides:

- **time**: the recording segment (video id and time range)
- **transcript**: speech transcribed from that segment
- **video**: sampled frames from that segment

**Video evidences:**

```
- time: {chunk_key}
- transcript: {transcript}
- video: {sampled_frames}

...
```

Based on the video evidences above, answer the following memory question.

**Question:** `{question}`

**Options:**

```
A. {choice_0}
B. {choice_1}
C. {choice_2}
D. {choice_3}
```

**Instructions.**

- Use the evidence cards to determine the most appropriate answer.
- If the evidence is insufficient to answer, choose the option stating the question cannot be answered.
- Respond with the letter only (e.g., `A`, `B`, `C`, or `D`).

Table 10: **SuperMemory-VQA and TeleEgo answer-generation prompt.** Repeat the evidence-card block once per retrieved segment.

You are an egocentric agent. Take yourself as the main character of the video.

- video: {video}
- dialogue: {dialogue}

Imagine you are the character in the video, describing from a first-person perspective what you saw, listen(dialogue) and everything that happened over time. Use I as the subject.

Table 11: **caption-generation prompt for all the datasets.**