EFFICIENT MULTIMODAL MEMORY CONSTRUCTION FOR LONG-FORM VIDEO RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-form video question answering requires retaining sparse evidence across extended temporal contexts. Retrieval-based systems make this setting tractable, but they can only retrieve evidence that was preserved when the video memory was written. Under a lightweight memory builder, compressing each clip into one query-agnostic caption is a difficult target: a generic narrative can omit fine-grained cues that later become relevant. We therefore propose **`MVP-Mem`**, a retrieval-oriented framework that represents each clip as short phrases extracted under complementary semantic views. The resulting phrases form compact, independently retrievable evidence anchors, accessed through a simple phrase-wise maximum scoring rule. Our primary comparison uses Qwen3-VL-4B for both caption and phrase memory construction, with the same downstream answer generator. Across four long-form video QA benchmarks, **`MVP-Mem`** improves over this caption baseline on every benchmark, raising average accuracy from 44.5% to 54.0% (+9.5 percentage points). On EgoLifeQA, it also nearly halves memory construction time relative to caption generation with the same Qwen3-VL-4B builder. These results support multi-view phrase memory as an effective construction target for lightweight VLMs.

1 INTRODUCTION

Video-language systems are increasingly expected to operate over long-form video rather than short clips (Li et al., 2026; Mangalam et al., 2023; Fang et al., 2024; Fu et al., 2025). This setting arises naturally in wearable platforms such as AI glasses (Yang et al., 2025b; Alam et al., 2026) and in assistive memory applications (Grauman et al., 2022; Tian et al., 2025), where a user may later ask a recall query such as "Where did I leave my keys?" about an event that occurred tens of minutes or hours earlier. Since processing an entire long video with a large video-language model at query time is often impractical due to context-length limits, compute cost, latency, and privacy concerns, practical systems typically partition the video into clips, construct offline memories, retrieve candidate clips for a query, and generate an answer grounded in the retrieved evidence, as illustrated in Figure 1. In this retrieval-based setting, video memory construction is a central bottleneck: retrieval can only surface evidence that has been preserved in the clip-level memories.

This paper studies memory construction under a practical constraint: per-clip memory should be constructible by using a *lightweight* VLM memory builder (*e.g.,* Qwen3-VL-4B (Bai et al., 2025)), rather than by repeatedly invoking a cloud-scale model for every clip. This constraint matters because, in long-form video, retrieving sparse evidence across many clips makes performance heavily dependent on memory quality. Under this constraint, lightweight VLMs often produce incomplete or wrong descriptions, making it difficult to preserve fine-grained cues that may later become query-relevant, such as a brief object relocation, a location transition, or a state change. On the other hand, cloud-scale per-clip captioning increases offline cost, latency, and the risk of privacy breach. Thus, the central question is not only which retriever to use, but also *what kind of memory a lightweight VLM can reliably construct so that relevant evidence*

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.

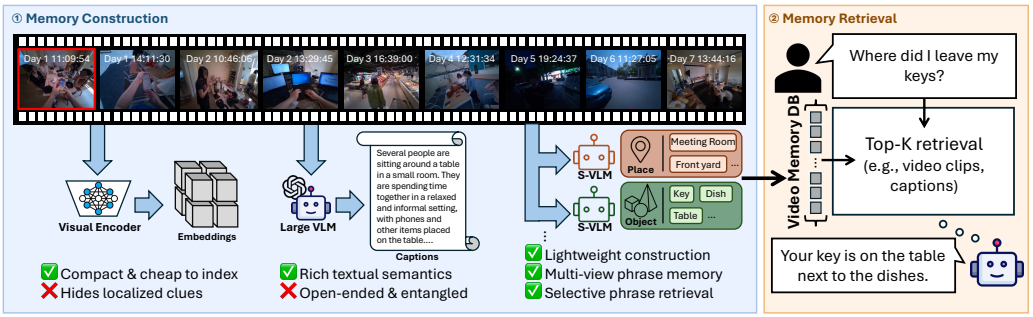


Figure 1: **Overview of long-form video QA with various memory construction strategies.** A long video is first converted into clip memories; at query time, the system uses the constructed memories to retrieve top-K clips to answer the user query. Visual-embedding memories are compact but coarse, while caption-based memories expose rich semantics but can be open-ended and entangled. **MVP-Mem** constructs multi-view phrase memories with a lightweight VLM and retrieves clips through the strongest supporting phrase.

remains retrievable in long-form video QA. Our thesis is that, under a constrained memory builder, a primary bottleneck is not only how to retrieve from memory but how to write the memory in the first place. We therefore optimize the memory-writing target rather than introduce a more complex retriever.

Many practical long-form retrieval pipelines instantiate clip memory either with a single visual embedding (Tschannen et al., 2025; Bolya et al., 2025; Cherti et al., 2023; Vasu et al., 2024; Wu et al., 2026) or with caption-style textual memory (Wang et al., 2021; Yang et al., 2023; Luo et al., 2025; Yang et al., 2025b). Single-vector embeddings are compact and cheap to construct the memory, but they collapse heterogeneous clip content into one similarity score; dominant scene factors can overshadow subtle but query-critical cues. Caption-style memories expose visual evidence in language by describing each clip in natural language. However, for a lightweight VLM (*e.g.,* Qwen3-VL-4B), a single all-purpose caption is a difficult memory-construction target: without knowing future queries, the model must compress many possible retrieval cues into one generic description. In our analyses, such captions sometimes omit distinctive retrieval anchors or introduce unsupported details (See Appendix B.1). Thus, the issue is not that captions are inherently inferior, but that caption generation is a difficult target for lightweight retrieval-oriented memory construction.

We argue that video memory constructed by lightweight VLMs should store *compact* retrieval cues rather than fluent all-purpose descriptions. A generic sentence-level caption tends to package visual evidence into a coherent narrative, spending part of its limited token budget on full-sentence narration and often merging multiple cues into one description. For retrieval, however, relevant evidence often appears as localized anchors, such as an object being moved, a room transition, or a visible state change. *Phrase*-level memory represents these anchors more directly: it avoids much of the overhead of full-sentence narration while still preserving local relational context, unlike isolated object labels. We further condition phrase extraction on multiple semantic views so that the lightweight VLM allocates memory to complementary types of evidence rather than only the most salient scene content. Thus, *multi-view phrase* memory is not intended to be universally more expressive than a strong caption; it is a higher-density and more controllable construction target for lightweight VLMs, preserving diverse evidence as compact, independently retrievable anchors.

To this end, we propose **MVP-Mem** (Multi-View Phrase-Level Memory), which co-designs memory construction and retrieval around a shared principle: preserve diverse evidence anchors when writing memory and select the strongest supporting anchor at query time. During construction, a cloud LLM is used once per benchmark domain to induce complementary semantic views from a high-level benchmark description; these views are fixed before retrieval and generated without access to individual test questions, answer distributions, or ground-truth annotations. A lightweight VLM then extracts focused phrases for each view and clip, yielding high-coverage memories that inevitably include phrases irrelevant to a particular query. Dur-

ing retrieval, a simple phrase-wise maximum rule scores a clip by its most query-aligned phrase instead of averaging all phrases into one representation. This scoring rule is intentionally simple: our contribution is the retrieval-oriented multi-vector memory and its alignment with the scoring rule, not the maximum operator itself. Our primary comparison is against captions written by the same Qwen3-VL-4B memory builder, testing whether a retrieval-oriented memory design improves performance at fixed builder capacity. With the same downstream answer generator, **MVP-Mem** improves accuracy on all four long-form video QA benchmarks, raising the average from 44.5% to 54.0% (+9.5 percentage points). On EgoLifeQA, it uses roughly half as many generated memory tokens as the same-model caption baseline.

In summary, our contributions are as follows. **(1)** We identify the query-agnostic compression target imposed by single-caption memory as a key bottleneck for lightweight per-clip memory builders in long-form video QA. **(2)** We introduce multi-view phrase memory, which decomposes memory writing into focused evidence-preservation tasks and stores the resulting cues as compact, independently retrievable anchors. **(3)** We align this representation with a simple phrase-wise maximum scoring rule so that the expanded memory is not averaged into a single vector and its most relevant local anchor remains accessible. **(4)** We demonstrate a 9.5-percentage-point average QA improvement over captions from the same Qwen3-VL-4B builder across four benchmarks, and use controlled ablations to examine the contributions of phrase representation, semantic views, and retrieval granularity.

2 MVP-MEM

2.1 PROBLEM FORMULATION

Let V be a long-form video and let $\mathcal{V} = \{v_j\}_{j\in[N]}$ denote its partition into N clips. For each clip v_j, we construct a compact memory representation that should preserve the evidence needed for downstream retrieval. Given a natural-language query q, the goal is to retrieve the top-K clips whose memories contain the evidence required to answer the query q:

$$\mathrm{TopK}(q) = \underset{j\in[N]}{\arg\,\mathrm{topK}}\ s(q, v_j), \tag{1}$$

where $s(q, v_j)$ is a clip scoring function which measures the relevance of the memory of clip v_j to the query q. The retrieved clips are then passed to a downstream answer generator.

The core challenge is that a clip memory must preserve localized cues while remaining discriminative for retrieval. Adding more descriptive content can improve coverage, but it can also introduce query-irrelevant text that dilutes matching. **MVP-Mem** addresses this tension by constructing each clip as phrase memories from multiple semantic views and retrieving clips with a phrase-level rule.

2.2 OVERVIEW

As shown in Figure 2, **MVP-Mem** consists of two stages: memory construction and memory retrieval. The core idea is to *co-design how a video clip is represented in memory and how that memory is accessed at retrieval time*. This design deliberately separates coverage from matching: memory construction expands each clip into multiple focused evidence anchors, while retrieval lets the query select the anchor that provides the strongest support.

Figure 2(left) shows memory construction. Each clip is processed under multiple complementary *semantic views* so that different aspects of the same clip, such as actions, objects, locations, and interactions, are extracted as separate phrases. At retrieval time, shown in Figure 2(right), **MVP-Mem** embeds the query and phrase memories using a text encoder and scores each clip with a phrase-wise maximum rule that selects the strongest supporting phrase. This construction-retrieval alignment is central to the method: multi-view

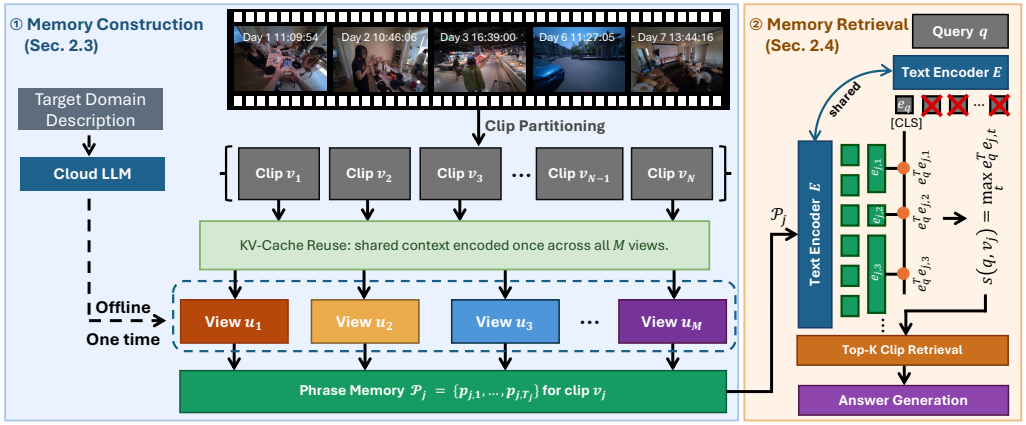


Figure 2: **Overview of MVP-Mem. Left:** During memory construction, **MVP-Mem** first defines a small set of complementary semantic views for the target domain in a one-time offline initialization step. It then prompts a lightweight VLM to extract focused phrases from each clip under these view-specific prompts, while reusing a shared KV-cache across views. **Right:** During memory retrieval, the query and phrase memories are embedded by the text encoder, and a simple phrase-wise maximum rule ranks clips by the phrase that best supports the query before answer generation.

phrase expansion increases the chance of preserving the needed evidence, while the set-based scoring rule prevents query-irrelevant phrases from diluting the relevant signal.

2.3 Memory Construction

MVP-Mem constructs each clip as a *multi-view phrase memory*: a compact set of phrase memories drawn from complementary semantic views.

Benchmark-level view induction. To adapt memory construction to different video environments, **MVP-Mem** uses a one-time view-induction step. For each benchmark domain, a cloud LLM receives the high-level benchmark description and defines a semantic schema consisting of complementary views and view-specific prompts. This is benchmark-conditioned adaptation rather than query-conditioned prompting: the view set is fixed before clip memories are constructed and is generated without access to individual test questions, answer distributions, or ground-truth annotations. For example, for EgoLife, the LLM induces views related to plan-level transitions, object traces, social interactions, and distinctive entities. The same schema is then applied to every clip and query within that benchmark. Full benchmark descriptions and the fixed generated view sets are provided in Appendix E.1.

Phrase memory construction. Let $\mathcal{U} = \{u_i\}_{i=1}^{M}$ denote the induced semantic views, and M is the number of views. For each view u_i and clip v_j, a lightweight VLM (*i.e.,* memory builder) constructs a small list of short phrases conditioned on the view-specific prompt and, when specified by the view, auxiliary text such as dialogue or subtitles (See Appendix E.2 for the prompt template). We denote the resulting phrase set by $K_{i,j} = \{k_{i,j}^{(1)}, \ldots, k_{i,j}^{(L_{i,j})}\}$, where $L_{i,j}$ is the number of phrases for view u_i in clip v_j. After processing all views, we concatenate the outputs into a unified phrase pool for clip v_j,

$$\mathcal{P}_j = \mathrm{Concat}(K_{1,j}, \ldots, K_{M,j}) = \{p_{j,t}\}_{t=1}^{T_j}, \tag{2}$$

where $T_j = \sum_{i=1}^{M} L_{i,j}$. This phrase pool is the memory for clip v_j.

Amortized multi-view construction. A natural concern is that applying M view-specific prompts may multiply the cost of per-clip memory construction. In practice, across repeated per-view calls, the domi-

nant repeated cost lies in prefilling the shared visual-text clip context, which is identical across views. We therefore prefill this shared context once per clip and reuse its KV-cache when generating the phrase list for each view. For the repeated shared-context prefill stage, this optimization changes the cost from $\mathcal{O}(M\ell_j)$ to $\mathcal{O}(\ell_j)+\mathcal{O}(Mb)$, where ℓ_j denotes the shared visual-text context length and b abstracts the much smaller per-view suffix processing and short generation overhead after cache reuse. Because ℓ_j is dominated by video tokens and the phrase outputs are short, multi-view construction adds mainly lightweight view-specific decoding on top of a single shared prefill. The optimization affects only inference efficiency; the resulting memory representation and retrieval rule remain unchanged.

2.4 Memory Retrieval

Once memory is constructed as a pool of phrases from Section 2.3, most phrases from a clip may be unrelated to any given query. Retrieval should therefore reward the phrase that best supports the query rather than average over unrelated evidence. We use a simple maximum over query–phrase similarities. We refer to this scoring rule descriptively as *asymmetric MaxSim* because the clip, but not the query, is represented by multiple vectors.

Phrase encoding. At retrieval time, a shared text encoder E maps both the query and each memory phrase into a common embedding space. For each clip-level phrase set $\mathcal{P}_j$, we encode the full set jointly and derive a separate embedding $\mathbf{e}_{j,t} \in \mathbb{R}^d$ for each phrase $p_{j,t} \in \mathcal{P}_j$ by mean-pooling over the token embeddings of its span. Thus, each phrase remains the retrieval anchor, but its representation is contextualized by the other phrases in the same set. We adopt this design because individual phrases are often short and semantically incomplete in isolation, whereas the co-occurring phrases provide useful local context for disambiguation and semantic calibration. The query embedding $\mathbf{e}_q = E(q) \in \mathbb{R}^d$ is obtained from the [CLS] token, and retrieval is still performed by scoring each phrase embedding independently against $\mathbf{e}_q$.

Phrase-wise maximum scoring. We first compute the similarity between the query embedding and each phrase embedding in clip v_j,

$$r(q, p_{j,t}) = \mathbf{e}_q^\top \mathbf{e}_{j,t}. \tag{3}$$

We then score the clip by the maximum similarity over its phrase memory:

$$s(q, v_j) = \max_{p_{j,t} \in \mathcal{P}_j} r(q, p_{j,t}). \tag{4}$$

The intuition is that a clip should be retrieved if at least one phrase in its memory strongly supports the query. Averaging over all phrase matches can dilute the signal from the most informative phrase with unrelated or weakly relevant phrases. The maximum operator is intentionally simple and is not itself our primary methodological contribution. Rather, the contribution lies in constructing retrieval-oriented, semantically decomposed phrase memories whose independently addressable units make this set-based scoring rule effective. This scoring form is related to established multi-vector and late-interaction retrieval (Khattab & Zaharia, 2020; Santhanam et al., 2022). Its role here is to match the structure of VLM-generated phrase memory: unlike token-level ColBERT, the query is a single vector and the document-side units are generated multimodal evidence phrases.

Final retrieval. After precomputing phrase embeddings offline, all clips are ranked by $s(q, v_j)$, and the top-K clips are returned. K is determined by a user in consideration of trade-off between evidence coverage and computational cost. The selected clips and their associated answer-time evidence are then passed to a downstream answer generator (*e.g.,* Qwen3-VL-8B). Phrase memories are used exclusively for retrieval and are not included in the answer-generation input. See Appendix E.3 for the benchmark-specific input formats.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Datasets. We evaluate on four long-video QA benchmarks. **Video-MME (Long)** (Fu et al., 2025) covers open-domain perception and reasoning. **EgoLife** (Yang et al., 2025b) tests multi-day egocentric recall of activities, objects, and interactions. **Ego-R1** (Tian et al., 2025) emphasizes temporal and causal reasoning in egocentric video. **Supermemory-VQA** (Alam et al., 2026) evaluates practical long-horizon memory across personal, spatial, conversational, and temporal tasks. See Appendix A.1 for details.

Baselines. Our primary baseline is single-caption memory constructed by Qwen3-VL-4B, the same per-clip builder used by **`MVP-Mem`**. This comparison holds builder capacity and the downstream answer generator fixed while changing memory construction and retrieval. We also evaluate the following baseline families to contextualize the results. *No-memory methods* provide reference systems without external memory and include both general video LLMs, such as GPT-5.4-mini (Singh et al., 2025) and Qwen3-VL-8B-Instruct (Bai et al., 2025), and long-video models, including Time-R1 (Wang et al., 2025a) and Video-RTS (Wang et al., 2025b). *Visual embedding memory methods* test clip retrieval from single visual embeddings using OpenCLIP (Cherti et al., 2023), Perception Encoder (Bolya et al., 2025), MobileCLIP (Vasu et al., 2024), and SigLIP 2 (Tschannen et al., 2025). *Caption-style memory methods* test retrieval over textual clip memories: direct caption retrieval ranks clip captions by query similarity. We adopt the caption-generation prompt from EgoLife (Yang et al., 2025b) and vary only the captioning model. These model-generated captions serve as retrieval memories; in the EgoLifeQA experiments, they are not passed to the answer generator, which instead receives the dataset-provided description associated with each retrieved segment. EgoRAG (Yang et al., 2025b) organizes caption-derived information into hierarchical textual memory, while LightRAG (Guo et al., 2025) builds a graph from the provided captions. Video-RAG (Luo et al., 2025) provides a complementary transcript-based retrieval baseline: it retrieves directly from speech transcripts and other extracted video text, such as OCR, without rewriting the transcripts into view-conditioned phrase memories. Its inclusion tests retrieval over source text alongside caption-based and multi-view memory construction.

Implementation details. For clip-level retrieval, we use `bge-large` (Xiao et al., 2024) as the shared text encoder and retrieve the top-K=20 clips per query. Memory embeddings are precomputed offline. For a controlled comparison, all four visual embedding models, OpenCLIP, SigLIP 2, Perception Encoder, and MobileCLIP, use Base/16-sized backbones. Following (Yang et al., 2025b), we partition each video into 30-second clips and sample frames at 0.5 fps; with K=20 retrieved clips, retrieval-based methods pass roughly 300 frames to the answer generator, matching the frame budget used by the no-retrieval baselines. All methods that answer from retrieved video clips, including caption-retrieval baselines and **`MVP-Mem`**, use Qwen3-VL-8B-Instruct (Bai et al., 2025) as the answer generator. The memory builder in **`MVP-Mem`** is Qwen3-VL-4B. Metric definitions are provided in Appendix A.2. All experiments are conducted on a single NVIDIA H100 GPU. Since inference is performed using deterministic greedy decoding, results do not vary across random seeds; thus, all reported results are obtained from a single run.

3.2 MAIN RESULTS

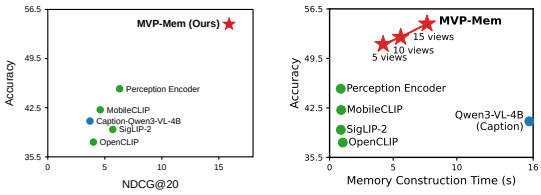
Improving memory with the same lightweight builder. Table 1 first tests whether changing the memory-writing target improves QA with model capacity held fixed. With the same Qwen3-VL-4B builder and Qwen3-VL-8B answer generator, **`MVP-Mem`** improves over single-caption memory on all four benchmarks: by 3.0 percentage points (pp) on Video-MME (Long), 13.8 pp on EgoLifeQA, 19.3 pp on Ego-R1, and 1.9 pp on SuperMemory-VQA. Its average accuracy rises from 44.5 to **54.0**, a **9.5 pp** gain over the same-model caption baseline. This primary comparison evaluates the combined construction and retrieval design; Table 2 examines the contributions of its individual components.

Approach	Memory Builder	Generator	MME(Long)	EgoLife	Ego-R1	SM-VQA	Avg.
No Memory	-	GPT-5.4-mini	59.7	35.8	37.3	41.9	43.7
	-	Qwen3-VL-8B	53.6	31.2	27.7	37.1	37.4
	-	Time-R1	33.1	34.2	32.7	37.5	34.4
	-	Video-RTS	33.3	35.8	33.3	38.6	35.3
Visual Embedding	OpenCLIP	Qwen3-VL-8B	57.7	37.6	35.0	37.1	41.9
	Perception Encoder (PE)	Qwen3-VL-8B	56.9	45.2	40.0	38.3	45.1
	MobileCLIP	Qwen3-VL-8B	57.9	42.2	32.7	37.9	42.7
	SigLIP 2	Qwen3-VL-8B	55.9	39.4	50.3	37.9	45.9
EgoRAG	EgoGPT-7B	Qwen3-VL-8B	55.0	45.2	49.0	38.8	47.0
Video-RAG	-	Qwen3-VL-8B	58.0	48.0	38.0	41.8	46.5
LightRAG	GPT-5.2	Qwen3-VL-8B	**60.4**	37.1	39.7	42.9	45.0
Caption	Qwen3-VL-4B	Qwen3-VL-8B	56.2	40.6	37.0	44.2	44.5
MVP-Mem (Ours)	Qwen3-VL-4B	Qwen3-VL-8B	59.2	**54.4**	**56.3**	**46.1**	**54.0**

Table 1: **QA accuracy (%) on Video-MME (Long), EgoLifeQA, Ego-R1 Bench, and SuperMemory-VQA (SM-VQA).** The final two rows give the primary comparison: caption memory and **MVP-Mem** use the same Qwen3-VL-4B builder and Qwen3-VL-8B answer generator. All memory-based methods share the answer generator. **Bold** indicates the best score among all methods.

Memory	Multi view	Max sim	Tokens /clip	Vectors /clip	Acc.
Caption	✗	✗	410	1.0	40.6
Caption‡	✗	✓	410	20.9	38.2
Phrases	✗	✗	47	1.0	48.0
Phrases (long)	✗	✗	254	1.0	49.8
Phrases (long)	✗	✓	254	69.3	51.2
Phrases	✓	✗	213	1.0	52.0
Phrases	✓	✓	213	49.0	**54.4**

Table 2: **Controlled ablation on EgoLifeQA.** With the same maximum scoring rule, multi-view phrases outperform the expanded single-view baseline (54.4 vs. 51.2) using fewer tokens and vectors. "Long" expands the single-view output budget; ‡ splits captions into sentence vectors before maximum scoring.



(a) Retrieval quality vs. QA (b) Memory construction

Figure 3: **Retrieval quality and memory construction efficiency on EgoLifeQA. (a)** Multi-view phrase memory improves retrieval NDCG@20 and QA accuracy over captions from the same Qwen3-VL-4B builder and visual-memory baselines. **(b)** It achieves higher accuracy while nearly halving construction time relative to Qwen3-VL-4B caption.

Other memory and retrieval baselines. **MVP-Mem** exceeds SigLIP 2, the strongest visual-memory baseline on average, by 8.1 pp, and Video-RAG, LightRAG, and EgoRAG by 7.5, 9.0, and 7.0 pp, respectively. Video-RAG provides a complementary comparison to retrieval over transcripts and extracted video text. Broader comparisons with systems using different backbones and iterative inference are reported separately in Appendix B.2.

3.3 What Makes Phrase Memory Effective?

Table 2 separates the effects of phrase representation, semantic views, and retrieval granularity on EgoLifeQA, using the same Qwen3-VL-4B builder and Qwen3-VL-8B answer generator.

Phrase representation improves the lightweight baseline. With single-vector retrieval fixed, replacing a 410-token caption with 47-token single-view phrases raises accuracy from 40.6 to 48.0 (+7.4 pp). Expanding the single-view phrase output to 254 tokens raises accuracy further to 49.8. Thus, the initial phrase-memory gain is already present without multi-view prompting or maximum scoring; the expanded variant provides a stronger baseline for testing semantic decomposition.

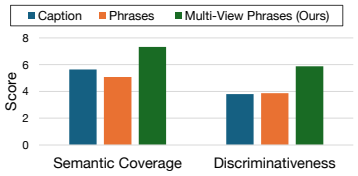


Figure 4: **Auxiliary memory-quality analysis.** LLM-judge scores for coverage and discriminativeness with the same 4B builder.

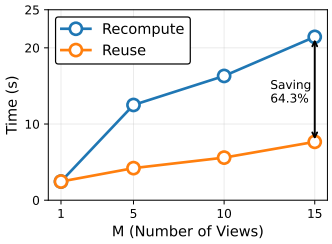


Figure 5: **KV-cache reuse on EgoLifeQA.** Sharing the clip context reduces construction time compared with recomputing it for every view.

Retrieval	Acc.	Time (s)
Single-vector (PE)	45.2	0.12
Multi-view phrase memory		
Single-vector	52.0	**0.08**
ColBERT	51.2	0.41
Phrase-wise max	**54.4**	0.26

Table 3: **Retrieval cost on EgoLifeQA.** The phrase-memory rows use the same constructed memory and vary the retrieval rule. PE is a visual embedding reference.

Multi-view gains persist against a larger single-view memory. The key comparison holds the maximum scoring rule fixed: multi-view phrases reach **54.4**, versus **51.2** for expanded single-view phrases, while using **fewer tokens** (213 vs. 254) and **fewer vectors** (49.0 vs. 69.3). The same ordering holds with single-vector retrieval (52.0 vs. 49.8). On EgoLifeQA, these controls support a benefit from semantic view decomposition beyond increasing output length or the number of indexed vectors.

Maximum scoring benefits from phrases. Keeping the multi-view memory fixed, phrase-wise maximum scoring improves accuracy from 52.0 to 54.4 (+2.4 pp); it also improves expanded single-view phrases from 49.8 to 51.2. In contrast, splitting captions into sentence vectors and taking their maximum similarity reduces accuracy from 40.6 to 38.2. The results support aligning retrieval with the constructed phrases: adding maximum scoring alone does not explain the gain over captions.

Retrieval evidence and memory diagnostics. Figure 3a relates retrieval NDCG@20 to QA accuracy on EgoLifeQA. For a given retrieved segment, all methods supply the same dataset-provided description, dialogue, and video to the answer generator; generated memories are used only for retrieval. The comparison therefore reflects which evidence is retrieved. In retrieval-only evaluation, **`MVP-Mem`** improves NDCG@20 over Qwen3-VL-4B captions from 3.7 to 15.9 on EgoLifeQA, from 2.3 to 22.9 on Ego-R1, and from 17.1 to 22.8 on SuperMemory-VQA (Appendix B.5). Figure 4 provides an auxiliary diagnostic: GPT-5.2 serves as an LLM judge (Gu et al., 2026), scoring memories from the same 4B builder for semantic coverage and discriminativeness. Multi-view phrases score higher on both axes. These scores are descriptive support, while the controlled ablation and retrieval metrics provide the primary evidence (judge protocol in Appendix C).

3.4 Construction and Retrieval Efficiency

Construction quality per token and per second. Figure 3b shows that multi-view phrase construction achieves higher accuracy while nearly halving construction time relative to Qwen3-VL-4B caption construction on EgoLifeQA (7.7s versus approximately 16s per clip). Increasing the views from 5 to 15 improves accuracy while keeping construction time moderate. The full memory uses 213 generated tokens per clip, compared with 410 for Qwen3-VL-4B caption memory.

Shared prefill makes multiple views practical. Figure 5 isolates KV-cache reuse. At 15 views, sharing the clip context reduces construction time from 21.5s to 7.7s per clip, a 64.3% reduction. The benefit grows with the number of views because each additional call otherwise repeats the shared-context prefill. This optimization reduces the cost of constructing the same multi-view representation.

Retrieval accuracy comes with a larger index. Table 3 makes the retrieval trade-off explicit. Phrase-wise maximum scoring improves over a single vector for the same memory (54.4 vs. 52.0), but stores 49 phrase vectors per clip on average and takes 0.26s rather than 0.08s. Compared with the evaluated ColBERT

configuration, it is more accurate (54.4 vs. 51.2) and faster (0.26s vs. 0.41s). Thus, the efficiency advantage concerns memory construction; phrase-level retrieval incurs additional index storage and query-time computation. The scoring comparison supports a practical match to phrase memory.

4 Related Work

Memory construction. Contrastive vision-language pretraining has made visual retrieval practical by mapping visual content and text into a shared embedding space. CLIP (Radford et al., 2021) and OpenCLIP (Cherti et al., 2023) establish this paradigm and show how retrieval quality improves through large-scale language-image pretraining and open-data scaling. SigLIP (Zhai et al., 2023) replaces the standard softmax contrastive objective with pairwise sigmoid matching, and SigLIP 2 (Tschannen et al., 2025) extends this line with stronger semantic understanding, localization, and dense features. MobileCLIP (Vasu et al., 2024) pushes similar alignment into smaller and faster encoders for efficient deployment, while Perception Encoder (Bolya et al., 2025) studies how to obtain stronger visual embeddings from the vision backbone rather than only from its final output. Across these efforts, the goal is to produce a better *single* visual embedding through improved pretraining objectives, model scaling, efficiency, or feature extraction. Caption-based methods instead move clip memory into language. Video captioning (Wang et al., 2021) and sequence-generation models (Yang et al., 2023) produce temporally grounded sentence-level descriptions that make actions, objects, and events explicit in text. More recent retrieval-oriented systems (Jeong et al., 2025; Luo et al., 2025; Wu et al., 2026) index or reason over such textual memories for long-video understanding. Compared with single-vector visual embeddings, caption memory is more explicit but costlier and can dilute local retrieval cues, especially for a lightweight VLM, motivating **MVP-Mem** to build phrase memories from complementary views.

Long-form video question answering. Recent long-form video QA systems increasingly combine retrieval with hierarchy or agentic decomposition to manage extended temporal context. WorldMM Yeo et al. (2026) and HippoMM Lin et al. (2026) build multimodal memory hierarchies for long-horizon reasoning. EgoRAG Yang et al. (2025b) retrieves relevant egocentric clips before downstream answering, and M3-Agent Long et al. (2026) improves QA through specialized agents that observe, remember, and reason over video in multiple steps. In contrast, **MVP-Mem** focuses on retrieval-oriented memory construction and retrieval rather than inference-time orchestration for final QA.

Multi-vector and late-interaction retrieval. Multi-vector representations and maximum-similarity aggregation are established ideas in information retrieval. ColBERT and ColBERTv2 (Khattab & Zaharia, 2020; Santhanam et al., 2022) retain contextualized query and document token vectors and aggregate token-level interactions; COIL (Gao et al., 2021) stores contextualized token representations in an inverted index; and ColPali (Faysse et al., 2025) extends late interaction to multi-vector visual-document retrieval.

5 Conclusion

We presented **MVP-Mem**, a retrieval-oriented memory framework for long-video QA that enables a lightweight VLM to preserve evidence as compact, independently retrievable phrases. With the same Qwen3-VL-4B builder and downstream answer generator, it improves over caption memory on all four benchmarks, raising average QA accuracy by 9.5 percentage points. Controlled ablations on EgoLifeQA support the contributions of phrase representation, semantic views, and phrase-wise retrieval; the full memory uses roughly half as many generated tokens as the same-model caption baseline. These findings support retrieval-oriented memory construction with lightweight VLMs, with lower construction cost but a larger retrieval index than single-vector memory. See Appendix D for limitations.

6 AI USE STATEMENT

The authors used generative AI tools only to assist with editorial revisions, such as correcting grammar and refining English expressions. No AI tools were used for the conception of the research, methodological development, experimental design or execution, data analysis, interpretation of results, or generation of technical content. All AI-assisted edits were carefully reviewed and approved by the authors.

7 ETHICS STATEMENT

Long-form video memory systems can support useful applications such as assistive recall, personal productivity, and accessibility by helping users recover relevant events from extended recordings. At the same time, they raise clear risks around privacy, surveillance, and erroneous retrieval, especially in egocentric settings where the stored memory may expose sensitive routines, locations, and social interactions. Practical deployment should therefore require strong access control, user consent, and careful auditing of stored and retrieved memories.

8 REPRODUCIBILITY STATEMENT

To support reproducibility under the double-blind review setting, we provide full details of the model architecture, training and inference procedures, datasets, evaluation protocol, and hyperparameters in the paper and appendix. Due to institutional constraints, the source code cannot be publicly released at the time of submission; however, this is independent of the technical contributions, and the implementation will be released after acceptance. We believe the provided descriptions are sufficient to enable independent reproduction of the reported results.

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, Hanoona Bangalath, 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.

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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv Preprint*, 2025.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. MMBench-Video: A long-form multi-shot benchmark for holistic video understanding. In *Proceedings of Advances in Neural Information Processing Systems*, 2024.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *Proceedings of International Conference on Learning Representations*, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2025.

Luyu Gao, Zhuyun Dai, and Jamie Callan. Coil: Revisit exact lexical match in information retrieval with contextualized inverted list. In *Proceedings of North American Chapter of the Association for Computational Linguistics*, 2021.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2022.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 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.

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.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. In *Proceedings of International Conference on Learning Representations*, 2026.

Yueqian Lin, Jingyang Zhang, Qinsi Wang, Hancheng Ye, Yuzhe Fu, Yudong Liu, Hai Helen Li, and Yiran Chen. Hippomm: Hippocampal-inspired multimodal memory for long audiovisual event understanding. In *Proceedings of Conference on Computer Vision and Pattern Recognition (Findings)*, 2026.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In *Proceedings of International Conference on Learning Representations*, 2026.

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

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Proceedings of Advances in Neural Information Processing Systems*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of International Conference on Machine Learning*, 2021.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of North American Chapter of the Association for Computational Linguistics*, 2022.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *ArXiv Preprint*, 2025.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *ArXiv Preprint*, 2025.

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.

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *Proceedings of International Conference on Computer Vision*, 2021.

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.

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*, 2025b.

Peiran Wu, Zhuorui Yu, Yunze Liu, Chi-Hao Wu, Enmin Zhou, and Junxiao Shen. Marc: Memory-augmented rl token compression for efficient video understanding. In *Proceedings of International Conference on Learning Representations*, 2026.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *ArXiv Preprint*, 2025a.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2Seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2023.

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*, 2025b.

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*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of International Conference on Computer Vision*, 2023.

A Additional Experimental Details

A.1 Datasets

The four benchmarks used in our experiments cover complementary long-video settings: open-domain long-form video understanding and long-horizon egocentric memory reasoning. Table 4 summarizes the evaluation scale for each benchmark setting used in this paper.

Dataset	Num. Queries	Video Length
Video-MME (Long)	900	0.64 h
EgoLifeQA	500	44.3 h
Ego-R1 Bench	300	44.3 h
SuperMemory-VQA	4,853	52.9 h

Table 4: **Dataset statistics for the four evaluation benchmarks.** We report the number of multiple-choice queries used in our experiments and the corresponding video duration for each benchmark setting.

Video-MME (Long). Video-MME (Fu et al., 2025) is a comprehensive benchmark for evaluating multimodal LLMs on video analysis across a broad range of capabilities, including perception, OCR, counting, temporal reasoning, and higher-level semantic understanding. The full benchmark contains short-, medium-, and long-duration subsets; in this paper we use only the long split, which contains 900 multiple-choice questions over videos longer than 30 minutes. This split is the most relevant to our setting because it requires long-context evidence retrieval rather than short-window recognition. Video-MME also provides subtitle text, which we use as the auxiliary text modality during memory construction.

EgoLifeQA. EgoLife (Yang et al., 2025b) is a large-scale multimodal egocentric dataset collected from six participants who continuously recorded one week of daily life with wearable AI glasses, yielding roughly 300 hours of video over activities such as discussions, shopping, cooking, socializing, and entertainment. EgoLifeQA is the associated long-context question-answering benchmark, designed to support life-assistant scenarios such as recalling past events, tracking object locations, monitoring habits, and reasoning over daily routines. Following the setting used in prior work, we evaluate the 500-question split for participant A1 (JAKE), whose stream spans 44.3 hours. This benchmark is particularly relevant to retrieval-based memory because many questions depend on sparse cues distributed across routine activities and spoken interactions.

Ego-R1 Bench. Ego-R1 Bench (Tian et al., 2025) was introduced to study reasoning over ultra-long egocentric video and accompanies a Chain-of-Tool-Thought framework for agentic, step-by-step video QA. Compared with EgoLifeQA, which is closely tied to daily-life assistance and episodic recall, Ego-R1 Bench places greater emphasis on temporal dependency, causal reasoning, and questions that require combining information across distant moments. The benchmark consists of a 300-question evaluation setting associated with the 44.3-hour stream summarized in Table 4. This makes Ego-R1 Bench complementary to EgoLifeQA: both are week-scale egocentric benchmarks, but Ego-R1 Bench stresses longer reasoning chains beyond direct event recall.

SuperMemory-VQA. SuperMemory-VQA (Alam et al., 2026) is an egocentric visual question-answering benchmark designed to evaluate AI assistants on practical, long-horizon memory tasks grounded in everyday experiences. The dataset contains 52.9 hours of multimodal recordings collected from ten participants wearing Meta Aria glasses, including synchronized RGB video, audio transcription, eye gaze, IMU measurements, and SLAM trajectories. Through a human-verified annotation pipeline, it provides 4,853 multiple-choice question-answer pairs spanning six categories: object and location memory, intent recall, visual scene recall, timeline reconstruction, conversational memory, and in-context retrieval. Each question includes an explicit unanswerable option, enabling the benchmark to assess whether models avoid hallucinating when the available evidence is insufficient. SuperMemory-VQA is particularly relevant to retrieval-based memory systems because its questions often require locating sparse evidence across recordings lasting several hours or multiple days and, in some cases, combining information from temporally distant moments.



Shure: If I roll down this slope, the time will be quicker.
Lucia: Ah, so... When the instruction is to calculate the fastest route...
Jake: Hahaha.
Lucia: It would roll.
Shure: Always thought...
Tasha: Ah, it would walk here. It took a shortcut.
Jake: Taking a shortcut and walking, at night, it would give me...
Alice: Gaode Maps doesn't care if I can pass. It just gives me an option to go through walls. Isn't it just like that?
Jake: Going through walls? That's not possible.
Lucia: That's probably because its data wasn't updated in time. It doesn't know that point.
Jake: Yeah, it's not smart enough. Rolling down is achievable. Going through walls is not.
Lucia: How do you know? Maybe it wants you to break down the wall.

It was Day 3, and I was walking along a paved path in a landscaped residential area under a bright blue sky. Trees lined the walkway
...
As we walked, **Shure** started imagining the **"fastest route"** in a very literal way and said, **"If I roll down this slope, the time will be quicker."**
...
"When the instruction is to calculate the fastest route... it would roll." I heard **Jake** crack up
...
Tasha chimed in: **"It took a shortcut."**
Then **Alice** jumped in: **"Gaode Maps** doesn't care if I can pass. It just gives me an option to **go through walls."**
Jake: "Going through walls? That's not possible."
Lucia: "That's probably because its data wasn't updated in time." Jake: "Rolling down is achievable. Going through walls is not." Lucia: "Maybe it wants you to break down the wall."

GPT-5.2 Caption (493 tokens)

I am walking along a paved path in a park, surrounded by trees and greenery. The sun is shining brightly, casting shadows on the ground. I am wearing a white shirt and carrying a backpack. I am walking with a group of people, including a woman in a white shirt and a man in a plaid shirt.
...
I am walking with purpose, and I am enjoying the journey. I am walking with confidence, and I am enjoying the experience. I am walking with joy, and I am enjoying the moment. I am walking with peace, and I am enjoying the tranquility. I am walking with love, and I am enjoying the connection. I am walking with gratitude, and I am enjoying the beauty. I am walking with hope, and I am enjoying the future. I am walking with joy, and I am enjoying the present. I am walking with peace, and I am enjoying the moment. I am walking with love, and I am enjoying the connection. ...

Qwen3-VL-4B Caption (874 tokens)

walking on sidewalk, pointing, talking, people walking on sidewalk, trees lining sidewalk, buildings in background, **Shure** asks question to **Jake, Lucia** responds to Shure, Jake laughs, **Tasha** suggests walking **shortcut**, **Alice** comments on **Gaode Maps**, Jake disagrees with Alice, Lucia suggests **wall-breaking possibility**, "always thought", **"it would roll", "took a shortcut"**, "at night", **"not possible"**, "not smart enough", **"achievable", "maybe it wants you to break down the wall"**, park, path, trees, buildings, group walking together, peer discussion, calculating **fastest route, rolling down slope, taking shortcut, going through walls, map data update, breaking down wall, Jake, Lucia, Shure, Tasha, Alice, Gaode Maps**, **'rolling down slope → quicker time', 'going through walls → not possible', 'outdated data → doesn't know point'**

MVP-Mem (173 tokens)

Figure 6: **Qualitative examples of memory construction under different representations.** Captions often mix multiple cues into a long description, while **MVP-Mem** organizes evidence into compact phrases from multiple semantic views. Yellow highlights mark words that are directly grounded in the video clip or dialogue and can serve as retrieval anchors, while Red highlights mark unsupported or hallucinated caption details, which can introduce misleading retrieval cues. In **MVP-Mem**, different text colors indicate phrase memories generated from different semantic views.

A.2 Metrics

We report both retrieval quality and end-task accuracy. For retrieval, we use $NDCG@K$, which evaluates how well relevant video clips are ranked within the top-K retrieved results, assigning higher scores when ground-truth clips appear closer to the top. We report the average NDCG@K over all queries. For end-to-end evaluation, we report *accuracy* (%). Since all benchmarks in our experiments are multiple-choice, accuracy measures whether the model selects the correct answer option among A/B/C/D. Reported differences are descriptive comparisons on the fixed evaluation sets; we do not estimate sampling confidence intervals or claim statistical significance or equivalence for small score differences.

B Additional Analyses

B.1 Qualitative comparison.

Figure 6 compares memories from the same Qwen3-VL-4B builder, with GPT-5.2 captions as a strong reference. The Qwen3-VL-4B caption baseline repeatedly uses generic phrases such as "I am walking," increasing semantic overlap across clips, and even introduces unsupported details (highlighted in red) such as describing the residential roadside scene as a "park." With the same lightweight builder, **MVP-Mem** organizes grounded retrieval anchors (highlighted in yellow) into compact phrase-level units aligned with the clip and dialogue. Overall, the examples illustrate that lightweight captioning models can produce diluted or misleading memories, whereas **MVP-Mem** maintains a compact structured representation of grounded cues such as "Gaode Maps," "Shure," and "Lucia."

Method	Memory Builder	Answer Generator	Max-Round	Cloud/clip	Video-MME (Long)	EgoLifeQA	Ego-R1 Bench	Avg.
Ego-R1	Gemini-1.5-Pro + GPT-4	Qwen2.5-3B-Instruct	7	Yes	42.7	53.0	52.0	49.2
HippoMM	Qwen2.5-VL	GPT-4o	2	No	41.6	54.6	53.0	49.7
M3-Agent	Qwen2.5-Omni (9B)	Qwen3 control (33B)	5	No	55.3	53.5	52.5	53.8
WorldMM-8B	GPT-5-mini	Qwen3-VL-8B	5	Yes	**66.0**	**56.4**	52.0	**58.1**
MVP-Mem	Qwen3-VL-4B	Qwen3-VL-8B	1	No	59.2	54.4	**56.3**	56.6

Table 5: **Broader comparison to recent memory-centric long-video QA systems.** "Memory" denotes the model used for memory construction, "Generator" the final answer model or controller, "Max-Round" the released benchmark setting for iterative retrieval or tool use, and "Cloud/clip" whether memory construction invokes a cloud model for each clip. These methods are not directly controlled against Table 1 because they differ jointly in memory, generator, and inference protocol. **MVP-Mem** achieves the best performance on Ego-R1 Bench and the second-best average with a local 4B memory builder and one retrieval round.

Memory	Builder	MME (Long)	EgoLifeQA	Ego-R1	SM-VQA	Avg.
Caption	Qwen3-VL-4B	56.2	40.6	37.0	44.2	44.5
Caption	GPT-5.2	**59.7**	46.2	55.0	**46.1**	51.8
MVP-Mem	Qwen3-VL-4B	59.2	**54.4**	**56.3**	**46.1**	**54.0**

Table 6: **Additional QA comparison with a stronger caption writer.** All rows use Qwen3-VL-8B for answer generation. The average covers all four benchmarks, whereas Table 5 averages three benchmarks.

B.2 Broader Comparison to Recent Memory Systems

Table 5 positions **MVP-Mem** against recent memory-centric long-video systems. We report these methods separately from the controlled baselines in Table 1 because they differ jointly in memory construction, answer backbones, and inference-time iteration. The comparison should therefore be interpreted as a broader system-level reference rather than a controlled architecture ablation.

Table 5 shows that **MVP-Mem** remains competitive with recent memory-oriented systems despite using a fixed single-pass pipeline. It outperforms Ego-R1, HippoMM, and M3-Agent on average and achieves the strongest result on Ego-R1 Bench. The comparison caveat is important: the systems differ jointly in memory construction, answer models, and inference protocols. Ego-R1 and M3-Agent rely on multi-round agentic reasoning, while HippoMM uses adaptive two-stage recall. WorldMM-8B attains the strongest average overall, but it combines per-clip GPT-5-mini memory construction with Qwen3-VL-8B and up to five rounds of adaptive retrieval. Accordingly, we claim a strong accuracy/compute operating point for inexpensive single-pass memory construction, not system-level state of the art.

B.3 Additional Comparison with GPT-5.2 Caption Memory

The main experiments evaluate memory construction with a fixed lightweight builder. Here, we additionally generate clip captions with GPT-5.2 to test whether **MVP-Mem** remains competitive when the caption writer is more capable. We use the same caption-generation prompt as the Qwen3-VL-4B caption baseline and retain the shared text encoder, top-$K{=}20$ retrieval budget, and Qwen3-VL-8B answer generator. GPT-5.2 is used to construct retrieval memory, not to generate the final answers. Unlike the broader system comparison above, this experiment retains the main paper's single-pass evaluation protocol.

Table 6 shows that the lightweight **MVP-Mem** also performs well against GPT-5.2 caption memory. It achieves comparable accuracy on Video-MME (Long), Ego-R1, and SuperMemory-VQA, with differences of -0.5, +1.3, and 0.0 percentage points (pp), respectively, and improves accuracy by 8.2 pp on the evaluated EgoLifeQA stream. Its four-benchmark average is 54.0%, compared with 51.8% for GPT-5.2 captions.

Method	Memory	EgoLife	Ego-R1	SM-VQA
Visual RAG	OpenCLIP	4.0	2.0	9.2
Visual RAG	PE	6.3	3.4	14.2
Visual RAG	MobileCLIP	4.6	1.9	12.2
Visual RAG	SigLIP 2	5.7	2.5	10.8
Caption	Qwen3-VL-4B	3.7	2.3	17.1
Caption	GPT-5.2	12.3	**23.3**	22.7
MVP−Mem	Qwen3-VL-4B	**15.9**	22.9	**22.8**

Table 7: **Retrieval performance measured by NDCG@20 on EgoLife, Ego-R1, and SuperMemory-VQA (SM-VQA).** MVP-Mem substantially outperforms visual-memory baselines across all three benchmarks and achieves the best performance on EgoLife and SM-VQA, while remaining competitive with caption-based memory on Ego-R1.

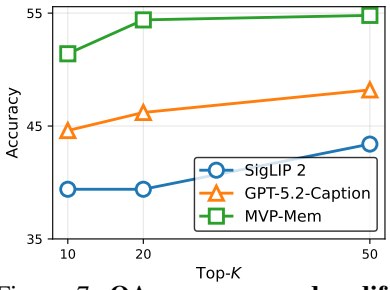


Figure 7: **QA accuracy under different retrieval budgets on EgoLifeQA. MVP−Mem** consistently outperforms various memories across all retrieval budgets.

Thus, the improvement over same-model captions also brings the lightweight phrase memory to a competitive accuracy level relative to a more capable caption writer. These observed scores do not establish statistical equivalence or a performance upper bound.

B.4 Memory Construction Cost

Figure 8 shows that **MVP−Mem** lies on a better write-time accuracy/token frontier than caption-based memory. The 15-view model reaches 54.4 accuracy with 213 generated tokens per clip, compared with 40.6 accuracy and 410 tokens for captions from the same Qwen3-VL-4B builder. As an additional reference, GPT-5.2 caption memory reaches 46.2 accuracy with about 410 tokens. Within **MVP−Mem**, increasing the number of semantic views from 5 to 10 to 15 raises QA accuracy from about 51.6 to 52.4 to 54.4 while increasing the generated token count from roughly 100 to 140 to 210. This trend supports the multi-view design at memory-writing time. The trade-off differs at retrieval time: the full system indexes 49 phrase vectors per clip on average and takes 0.26s per retrieval evaluation, compared with 0.08s when the same phrase memory is collapsed to one vector (Table 3). Thus, **MVP−Mem** reduces generated memory text and construction time but increases index size and query-time computation relative to a single-vector scoring rule.

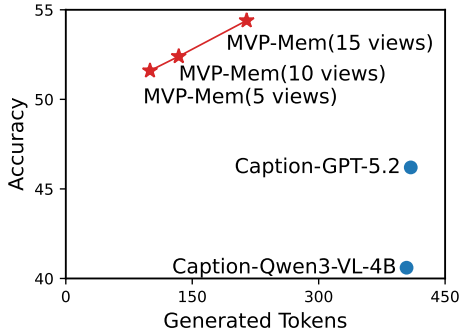


Figure 8: **Accuracy-efficiency comparison. MVP−Mem** attains higher QA accuracy than caption memory with roughly half as many tokens, and more semantic views further improve accuracy with controlled token growth.

B.5 Retrieval Results

We complement the end-to-end QA results in Table 1 with a retrieval-only evaluation on EgoLife, Ego-R1, and SuperMemory-VQA (SM-VQA). For each query, we rank the clip memories and report NDCG@20, a rank-sensitive metric that rewards retrieving relevant clips within the top 20 positions and assigns higher scores when they appear earlier in the ranked list. All methods use the same clip partition and retrieval scoring setup as in the main experiments, thereby isolating the quality of the memory representation and retrieval scoring from the downstream answer generator. We exclude Video-MME from this evaluation because many of its questions are not suitable for clip-level retrieval assessment. For example, questions about the overall

topic or global content of a video may treat most or all clips as relevant, making a localized retrieval ground truth ill-defined and the resulting ranking metric less informative.

Table 7 evaluates retrieval quality independently of answer generation. With the same Qwen3-VL-4B builder, **MVP-Mem** improves NDCG@20 over captions by 12.2 points on EgoLife, 20.6 points on Ego-R1, and 5.7 points on SM-VQA. These improvements complement the consistent QA gains over the same-model caption baseline. As a stronger reference, GPT-5.2 captions attain substantially higher retrieval scores than Qwen3-VL-4B captions. **MVP-Mem** reaches similar scores to this reference on Ego-R1 (-0.4 points) and SM-VQA (+0.1 points), and a higher score on EgoLife (+3.6 points).

Figure 7 further shows that the advantage of **MVP-Mem** becomes more pronounced in end-to-end QA. While SigLIP 2 visual memory shows limited gains as the retrieval budget increases, and GPT-5.2 caption memory improves moderately with larger top-K, **MVP-Mem** consistently achieves the highest accuracy across all retrieval budgets. Notably, the gap between **MVP-Mem** and GPT-5.2 caption memory in QA accuracy is larger than their NDCG difference in Table 7.

C LLM Judge Protocol

For the intrinsic memory-quality analysis in Figure 4, we use GPT-5.2 as an LLM judge on clips from the EgoLifeQA evaluation stream. For each clip and memory type, the judge independently scores the candidate memory along two retrieval-oriented axes: *semantic coverage* and *discriminativeness*. We report the average scores over the evaluated clips. The judge is instructed to evaluate retrieval utility. In particular, it is told that a candidate memory may be keyword phrases, a full caption, or a mixture of both, and that concise phrases should not be penalized if they capture retrieval-relevant evidence. Because the criteria and scores come from an LLM judge, this analysis is intended only as a format-neutral descriptive diagnostic; retrieval NDCG and end-task accuracy are the primary quantitative evaluations. The full prompt is shown in Table 8.

D Limitations and Future Work

While MVP-Mem generates fewer text tokens than caption-based memory, it still relies on autoregressive phrase construction and stores multiple phrase embeddings per clip. Its construction cost and index footprint therefore remain higher than those of single-vector embedding memories, reflecting a trade-off for preserving diverse and independently retrievable linguistic cues. Future work may investigate non-autoregressive phrase extraction, phrase deduplication, and more compact multi-vector indexing.

E Prompt Details

E.1 View Generation Prompts

This appendix details the benchmark-conditioned meta-prompts used by the cloud LLM (*e.g.,* GPT-5.2) to instantiate the semantic views introduced in Section 2.3. For each benchmark, we provide the high-level benchmark description given to the cloud model and the resulting fixed view-specific prompts used by the lightweight memory builder. The view list is fixed before clip memories and query retrieval are run, and no individual test questions, answers, or ground-truth annotations are supplied during induction. Across all settings, the prompts ask the small model to produce short phrases rather than full sentences so that the memory remains compact and compositional. The modality column indicates whether a view uses video only, text only, or both; for EgoLifeQA, `text` denotes dialogue, while for Video-MME (Long) it denotes subtitles. This protocol is benchmark-conditioned; we do not present it as evidence of universal-schema transfer.

You are an expert evaluator assessing how well a textual document serves as a retrieval index for a video clip. The document will be used in a retrieval system: given a natural language query about a video clip, the system searches documents to find the relevant clip. Your task is to judge how reliably this document would surface the correct clip when a user queries about its content. The document may be formatted as keyword phrases, full sentences (captions), or a mix; evaluate solely on retrieval effectiveness, not on writing style or format.

Video: `{video}`

The video frames above are sampled at 0.5 fps from a video clip. You are given one candidate document that will be used as a retrieval index for this clip.

Document: `{doc}`

Evaluate how well the document would enable a retrieval system to find this video clip when a user asks a natural language question about its content.

Important — Format Neutrality. The document may be formatted as keyword phrases, full sentences (captions), or a mix of both. Do not penalize or reward any particular format. A concise set of keyword phrases that covers the right concepts is equally valid as a fluent caption. Judge solely on whether the document's content would help a retrieval system surface the correct clip for a given user query.

Scoring criteria (each 0–10).
1. **Semantic Coverage**: Does the document capture the key actions, objects, people, and context that a user would naturally mention when querying about this clip?
2. **Discriminativeness**: Does the document contain specific, unique information that distinguishes this clip from other clips?

The model is asked to respond in JSON with the two scores for the document and a short retrieval-focused explanation.

Table 8: **Prompt template used for the LLM-as-a-Judge evaluation.** The candidate memory is inserted as Document, and the model returns structured JSON scores together with a retrieval-focused explanation.

EgoLifeQA & Ego-R1. The EgoLifeQA setting places more weight on long-horizon recall, routine behavior, social interaction, and object tracking over extended personal video streams. Accordingly, the meta-prompt emphasizes complementary semantic views that can capture both episodic events and persistent retrieval anchors from video and dialogue. Table 9 shows the dataset-conditioned meta-prompt used to induce the EgoLifeQA view set.

The resulting fixed schema emphasizes temporal continuity, plans, task progress, social roles, and distinctive entities. Several views use `both` modalities so that evidence mentioned in dialogue can be aligned with what is visible in the scene (see Table 10).

Video-MME (Long). Video-MME (Long) covers broader open-domain video understanding, mixing perception, reasoning, OCR, counting, synopsis, and temporal reasoning across diverse scene types. Compared with the EgoLifeQA setting, its fixed views place more emphasis on motion, spatial layout, written text, quantitative cues, and salient open-world events. Table 11 shows the benchmark-conditioned meta-prompt used to induce the Video-MME view set.

The resulting Video-MME prompt set is broader and more open-domain than the EgoLifeQA view set. It includes OCR, counting, synopsis, anomaly, and spatial-layout views (see Table 12).

SuperMemory-VQA. SuperMemory-VQA focuses on practical memory assistance scenarios, evaluating whether models can retrieve and reason over information that users may forget over long periods of daily activities. Compared with the broader open-domain setting of Video-MME and the routine-behavior-oriented setting of EgoLifeQA, its fixed views place greater emphasis on episodic events, personal intentions, object

Dataset context. EgoLifeQA is designed to evaluate whether models can understand and remember everyday life from week-long video recordings. It includes questions that require recalling past events, tracking object locations, and reasoning over long-term activities. In our experiments, we use the questions from the perspective of a single participant (A1: JAKE) together with the corresponding video stream, which spans 44.3 hours.

Instruction. Video interpretation naturally varies depending on the observer's view. Focusing on a single aspect can cause important evidence from other views to be overlooked. Based on the dataset description above, design diverse perspectives that help small Vision-Language Models generate high-quality, retrieval-oriented textual memories from a single video segment and/or its corresponding dialogue. The views should improve downstream retrieval performance, and the small model should respond using a list of keywords or phrases rather than full sentences.

Required output format.

```
{"view_name": xxx,
"prompt to small model": yyy,
"Input": "video or dialogue or both"}
```

Table 9: **Dataset-conditioned meta-prompt given to the cloud LLM to generate candidate semantic views for EgoLifeQA and Ego-R1.**

locations, conversations, and temporal relations across distant moments. Table 13 shows the benchmark-conditioned meta-prompt used to induce the SuperMemory-VQA view set, and Table 14 lists the resulting schema.

E.2 Memory Construction Prompts

During memory construction, the lightweight memory builder receives the sampled video clip, the associated dialogue or subtitle text when available, and a view-specific instruction. We use a shared retrieval-oriented prompt template and replace the placeholder `{{view-wise prompt}}` with one of the "Prompt to Small Model" entries from Table 10 for EgoLifeQA and Ego-R1, Table 12 for Video-MME (Long), or Table 14 for SuperMemory-VQA. In the egocentric benchmarks, `{{dialogue}}` is instantiated with dialogue text, while in Video-MME (Long) it is instantiated with subtitle text. When the view-specific prompt imposes a narrower focus or a more specific output format, that instruction takes precedence over the generic template in Table 15.

E.3 Answer Generation Prompts

During answer generation on EgoLifeQA and Ego-R1, the answer model receives a sequence of evidence cards, one per retrieved segment. Each card contains the segment timestamp, associated dialogue, the dataset-provided segment description, and the retrieved video segment itself, as shown in Table 16. Importantly, `description` is the description distributed with each EgoLife segment, not a caption generated by a retrieval-memory builder. For example, the caption produced by the Caption–Qwen3-VL-4B baseline is used to index and retrieve segments but is not inserted into the answer-generation prompt; the same separation applies to other caption baselines and to the phrase memory produced by **`MVP-Mem`**. Thus, all EgoLifeQA and Ego-R1 retrieval methods provide the answer model with the same answer-time text for a given retrieved segment. This evidence-card block is repeated once per retrieved segment (up to $K=20$ in our experiments). The prompt then asks the model to answer a multiple-choice question using only the option letter.

For Video-MME (Long) and SuperMemory-VQA, we use the shared answer-generation prompt shown in Table 17. The answer model receives one evidence card per retrieved segment, containing the segment identifier and time range, its transcript, and sampled video frames. After the repeated evidence-card block, the model is given the question and four answer choices and is instructed to return only the letter of the most

View	Prompt to Small Model	Modality
Action Summary	Identify the main observable actions occurring in this segment. Output only a list of short action phrases (verbs plus objects when possible), such as 'cutting vegetables' or 'opening fridge'.	both
Object State and Location	Extract objects that appear or are mentioned, together with their state or location when available. Output only keywords or phrases in the form 'object - location/state'.	both
Actor Role Interaction	List the people involved and their interactions. Output only short relation phrases such as 'Jake gives tool to Alice' or 'Shure asks question to Jake'.	both
Event Boundary and Context	Identify whether a meaningful event starts, continues, or ends in this segment. Output only short phrases describing the event context, such as 'start cooking' or 'finish cleaning'.	both
Temporal Cues	Extract temporal indicators such as time references, sequence words, or routines. Output only keyword phrases such as 'after lunch', 'tomorrow plan', or 'last time mentioned tiramisu'.	text
Spatial Context	Identify the physical place or spatial setting that is visible or mentioned. Output only location keywords or short phrases such as 'kitchen' or 'outside near car'.	both
Object Interaction Trace	Track object usage transitions in this segment. Output only short phrases describing object handling or transfer, such as 'picked up screwdriver' or 'moved backpack to sofa'.	video
Dialogue Intention and Plans	Extract intentions, plans, commitments, or future tasks mentioned in speech. Output only keyword phrases such as 'need to buy milk' or 'finish editing later'.	text
Habit or Repetition Signals	Detect cues suggesting repeated behavior or preferences. Output only phrases indicating habits or recurring actions, such as 'usually drinks coffee' or 'daily workout'.	both
Social Relation Signals	Extract social dynamics or role relationships that are expressed or observed. Output only short phrases such as 'friend helping repair' or 'family conversation'.	both
Task Progress Status	Identify ongoing tasks and infer their progress state. Output only phrases such as 'assembling shelf - in progress' or 'packing bag - completed'.	both
Question or Problem Context	Extract explicit or implicit problems, questions, or goals being discussed or acted upon. Output only keyword phrases such as 'looking for keys' or 'fixing broken handle'.	both
Salient Changes	List notable changes within the segment, including object movement, scene shifts, or state changes. Output only short change phrases.	video
Memory Anchor Entities	Extract highly distinctive entities, such as objects, names, places, or unusual events, that could serve as retrieval anchors. Output only keywords or short noun phrases.	both
Cause-Effect Hints	Identify explicit or implied causal relations in actions or dialogue. Output only short phrases such as 'spilled water $\rightarrow$ cleaning floor' or 'late wake-up $\rightarrow$ skipped breakfast'.	both

Table 10: **Fixed EgoLifeQA and Ego-R1 view set used for multi-view phrase memory construction.**

appropriate option. If the evidence is insufficient, it selects the option indicating that the question cannot be answered.

Dataset context. Video-MME is a comprehensive video understanding benchmark with 2,700 questions spanning multiple video durations. In our experiments, we use only the long subset (longer than 30 minutes), which contains 900 questions and is specifically intended to assess long-video reasoning.

Instruction. Video interpretation naturally varies depending on the observer's perspective. Focusing on a single aspect can cause important information from other perspectives to be overlooked. Based on the dataset description above, design diverse views that help small Vision-Language Models generate high-quality, retrieval-oriented textual memories from a single video segment and/or its corresponding subtitle. The perspectives should improve downstream retrieval performance, and the small model should respond using a list of keywords or phrases rather than full sentences.

Required output format.

```
{"view_name": xxx,
"prompt to small model": yyy,
"Input": "video or subtitle or both"}
```

Table 11: **Dataset-conditioned meta-prompt given to the cloud LLM to generate candidate semantic views for Video-MME (Long).**

E.4 Single-view Phrase Memory Construction Prompt

For the single-view phrase memory baseline in Table 2 (*i.e.,* second row), we use a shared prompt template to construct phrase-level memory from each video clip and its accompanying dialogue or subtitle text. The prompt is designed to extract retrieval-oriented keyword phrases that capture the most salient actions, objects, characters, and character-specific actions in the scene. More specifically, it instructs the model to produce 8-12 concise keyword phrases, including action phrases, object keywords, character names, and character-action pairs, with an emphasis on specificity and retrieval effectiveness. The output is formatted as a comma-separated list without additional explanation. Here, the exact prompt template used for single-view phrase memory construction is shown in Table 18. This same template is applied to each clip independently, without multi-view decomposition. The resulting phrases are then concatenated into a single sequence and used as the memory representation for the single-view phrase baseline.

F Licenses and Terms of Use

Table 19 summarizes the main external assets used in our experiments together with their publicly documented licenses or usage terms. We use these assets under their original terms and do not redistribute any third-party data or model weights as part of this submission.

View	Prompt to Small Model	Modality
Human Participants and Roles	Identify people or agents and their roles in this segment. Output only keywords or short phrases that capture participant identity cues, social roles, and interaction roles, such as 'speaker at podium' or 'police officer'.	`both`
Interaction and Relationship Cues	Extract interaction-focused memory cues. Output only short phrases that capture who interacts with whom or with what, including social, physical, and task interactions.	`both`
Motion and Trajectory	Extract movement-related memory cues from this segment. Output only a list of short phrases that capture paths, direction, approach or departure, entering or exiting, and motion patterns.	`video`
Spatial Layout and Relations	Extract spatial cues and relations from this segment. Output only a list of short phrases that capture object positions, layout, left or right relations, inside or outside relations, and relative placement.	`video`
Scene and Environment Type	Extract phrases describing the scene type, environment, and situational context. Output only keywords or short phrases that capture where the event occurs and what type of setting it is.	`both`
Named Entities and Referential Anchors	Extract names, places, organizations, products, titles, and other specific referential anchors. Output only keywords or short phrases that are useful for exact-match retrieval.	`both`
OCR and Written Text	Extract visible or spoken text that may support retrieval. Output only keywords or short phrases, focusing on screen text, signs, labels, captions, numbers, and other exact strings.	`both`
Counting and Quantitative Cues	Extract count-related and quantity-related cues from this segment. Output only short phrases that capture numbers, plurality, repeated instances, and measurable amounts.	`both`
Causal and Explanatory Clues	Extract phrases suggesting cause, purpose, reaction, or explanation. Output only short phrases that capture why events happen, what triggers what, and what consequences follow.	`both`
Emotion and Social Atmosphere	Extract emotion-related and social-tone cues from the segment. Output only short phrases that capture visible mood, tone of speech, conflict, celebration, tension, or comfort.	`both`
High-Level Synopsis Cues	Generate a retrieval-oriented synopsis as a list of compact phrases rather than sentences. Capture the main gist of the segment using 5–10 phrase-level memory units.	`both`
Anomaly and Salient Change Detection	Extract unusual, surprising, or salient changes that would serve as strong retrieval anchors. Output only keywords or short phrases describing anomalies, unexpected actions, sudden transitions, or rare objects and events.	`both`

Table 12: **Fixed Video-MME (Long) view set used for multi-view phrase memory construction.**

Dataset context. SuperMemory-VQA is an egocentric visual question answering (VQA) dataset for evaluating AI assistants on practical, long-horizon memory tasks. It contains 52.9 hours of everyday activities recorded with AI glasses, including synchronized RGB video, audio transcription, eye gaze, IMU, and SLAM trajectories. Through a human-verified annotation pipeline, it provides 4,853 grounded question-answer pairs that span object and location memory, intent recall, visual scene recall, timeline reconstruction, conversational memory, and in-context retrieval.

Instruction. Video interpretation naturally varies depending on the observer's view. Focusing on a single aspect can cause important evidence from other views to be overlooked. Based on the dataset description above, design diverse perspectives that help small Vision-Language Models generate high-quality, retrieval-oriented textual memories from a single video segment and/or its corresponding dialogue. The views should improve downstream retrieval performance, and the small model should respond using a list of keywords or phrases rather than full sentences.

Required output format.

```
{"view_name": xxx,
"prompt to small model": yyy,
"Input": "video or dialogue or both"}
```

Table 13: **Dataset-conditioned meta-prompt given to the cloud LLM to generate candidate semantic views for SuperMemory-VQA.**

View	Prompt to Small Model	Modality
People and Identity	Extract names, relationships, roles, appearance cues, and social entities visible or mentioned. Output only keywords/phrases.	`both`
Location and Place	Extract places, rooms, landmarks, venues, spatial regions, and environment descriptors. Output only keywords/phrases.	`both`
Object Inventory	List salient objects, devices, personal belongings, tools, food, documents, and unique items. Output only keywords/phrases.	`video`
Actions and Activities	Extract ongoing, completed, or planned activities and interactions. Output activity-centric keywords/phrases only.	`both`
Conversation Topics	Identify discussion topics, decisions, requests, questions, agreements, and concerns. Output only keywords/phrases.	`dialogue`
Future Plans and Intentions	Extract future events, commitments, appointments, reminders, goals, and intentions mentioned or implied. Output only keywords/phrases.	`both`
Temporal Anchors	Extract dates, times, durations, ordering cues, recurring events, and temporal references. Output only keywords/phrases.	`both`
Multi-Step Event Chain	Identify causes, plans, outcomes, follow-up actions, and linked events that may connect to future moments. Output only keywords/phrases.	`both`
Social Interactions	Extract meetings, greetings, collaborations, calls, social activities, and participant relationships. Output only keywords/phrases.	`both`
Personal Memory Cues	Extract information likely useful for later personal-memory questions: belongings, purchases, gifts, tasks, reminders, and preferences. Output only keywords/phrases.	`both`
Spatial Relations	Describe spatial layouts, object locations, relative positions, movement paths, and navigation cues. Output only keywords/phrases.	`video`
Communication Artifacts	Extract messages, emails, notes, documents, screens, slides, forms, or written content references. Output only keywords/phrases.	`both`
Task and Responsibility	Identify assignments, ownership, deadlines, action items, responsibilities, and requested follow-ups. Output only keywords/phrases.	`both`
Emotion and Intent Signals	Extract explicit emotions, reactions, preferences, satisfaction, complaints, excitement, or concerns. Output only keywords/phrases.	`both`
Rare and Distinctive Details	Capture unusual objects, unique events, distinctive phrases, uncommon locations, and memorable details. Output only keywords/phrases.	`both`
Question-Centric Retrieval Tags	Generate retrieval-oriented tags answering who, what, where, when, why, and with whom. Output concise keywords/phrases.	`both`
Evidence-Linking Keys	Extract entities and concepts most likely to connect with events occurring before or after this segment. Output only keywords/phrases.	`both`
Named Entity Memory	Extract person names, organization names, product names, event names, locations, and project names. Output only keywords/phrases.	`both`

Table 14: **Fixed SuperMemory-VQA view set used for multi-view phrase memory construction.**

Shared prompt template. You should extract retrieval-oriented keywords from a video scene together with its accompanying dialogue or subtitle text.

Inputs.
Dialogue or subtitle: `{{dialogue}}`
Video: `{{video clips}}`

The goal is to extract comprehensive keywords from a video and dialogue or subtitle text that focus on view-wise prompt.

Example output. `wooden spoon, kitchen counter, smartphone screen, John, Sarah, John cooking pasta, Sarah reading recipe, following stove instructions, measuring flour, stainless steel pot`

View-wise prompt. `{{view-wise prompt}}`

Table 15: **Shared memory construction prompt template used by the lightweight memory builder.** The final field `{{view-wise prompt}}` is replaced with a benchmark-specific view instruction listed in Tables 10, 12, and 14.

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 16: **EgoLifeQA and Ego-R1 answer generation prompt template used by the answer model.** The evidence-card block is repeated once per retrieved segment. The `description` field contains EgoLife's dataset-provided segment description, which is held fixed across retrieval methods and is distinct from any model-generated caption or phrase memory used for retrieval.

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 17: **Video-MME (Long) and SuperMemory-VQA answer generation prompt template used by the answer model.** The evidence-card block is repeated once per retrieved segment and contains the segment identifier and time range, transcript, and sampled video frames.

Shared prompt template. You should extract retrieval-oriented keywords from a video scene together with its accompanying dialogue or subtitle text.

Inputs.
Dialogue or subtitle: `{{dialogue}}`
Video: `{{video clips}}`

The goal is to extract comprehensive keywords from a video that include actions, objects, and characters to maximize retrieval effectiveness. The keywords should capture the most important visual and contextual elements.

Task. Analyze the video and dialogue or subtitle text to generate 8–12 keyword phrases that include actions, objects, characters, and character-specific actions.

Instructions.

1. **Identify key actions.** Focus on the main tasks, events, or instructions shown in the video. Start each action phrase with a verb and keep action phrases 2–4 words long.
2. **Extract all objects.** Include visible and mentioned objects, tools, items, locations, and environmental elements. Use specific object names rather than generic terms, and include both foreground and background objects when useful.
3. **Identify characters.** Extract character names from dialogue and visual content. Use exact names when mentioned; otherwise use descriptive terms such as `man`, `woman`, or `child`.
4. **Generate character-action pairs.** For each identified character, create phrases that combine the character with their specific action in the form `[character] + [specific action]`. Examples include `John cooking`, `Sarah explaining`, and `teacher demonstrating`.
5. **Prioritize specificity.** Prefer concrete, specific terms such as `smartphone screen` or `wooden spoon`, and avoid generic phrases such as `doing something`, `device`, or `tool`.
6. **Output format.** Return a comma-separated list with no additional text or formatting. Aim to include 2–3 action phrases, 3–4 object keywords, 1–2 character names, and 2–3 character-action pairs, for a total of 8–12 keyword phrases.

Example output. `wooden spoon, kitchen counter, smartphone screen, John, Sarah, John cooking pasta, Sarah reading recipe, following stove instructions, measuring flour, stainless steel pot`

Table 18: **Prompt template used for single-view phrase memory construction by the lightweight memory builder.**

Asset	Role	License / Terms
Video-MME (Long) (Fu et al., 2025)	Benchmark	Academic-only custom terms
EgoLifeQA (Yang et al., 2025b)	Benchmark	MIT
Ego-R1 Bench (Tian et al., 2025)	Benchmark	MIT
Qwen3-VL-4B-Instruct (Bai et al., 2025)	Model	Apache-2.0
Qwen3-VL-8B-Instruct (Bai et al., 2025)	Model	Apache-2.0
Qwen3-8B (Yang et al., 2025a)	Model	Apache-2.0
`bge-large` (Xiao et al., 2024)	Model	MIT
OpenCLIP (Cherti et al., 2023)	Model	MIT
MobileCLIP (Vasu et al., 2024)	Model	MIT
SigLIP 2 (Tschannen et al., 2025)	Model	Apache-2.0
Perception Encoder (Bolya et al., 2025)	Model	Apache-2.0
GPT-5 / GPT-5.2 (Singh et al., 2025)	API model	OpenAI service terms
Gemini 2.5 Pro (Comanici et al., 2025)	API model	Google Gemini API terms

Table 19: **Main external assets used in this paper and their publicly documented licenses or usage terms.** When benchmark or model terms are not exposed as a standard open-source license, we report the public access restriction or service terms instead.