VISUAL AGENTIC MEMORY: ENABLING ONLINE LONG VIDEO UNDERSTANDING VIA ONLINE INDEXING, HIERARCHICAL MEMORY, AND AGENTIC RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Long video understanding requires more than large context windows. It also needs a memory mechanism that decides what visual evidence to retain, keeps it searchable over long horizons, and grounds later reasoning in recoverable observations rather than compressed latent state alone. We propose Visual Agentic Memory (VAM), a training-free framework with three components. *Online Indexing* supports selective evidence retention under streaming constraints. *Hierarchical Memory* organises retained evidence in a parallel temporal-spatial representation, with temporal event summaries for long-horizon localisation and spatial observations for direct visual verification. *Agentic Retrieval* searches, inspects, and verifies candidate evidence before producing a grounded answer. On OVO-Bench, VAM reaches a competitive RT+BT average of 68.41, improving over the 67.46 end-to-end Gemini 3 Flash baseline and ranking fourth in the updated comparison. On the month-scale split of MM-Lifelong *train@month* (105.6 hours over 51 days), VAM reaches 17.11%, second only to ReMA with GPT-5 (17.62%). These results suggest that long-horizon video understanding benefits from treating visual memory as an explicit, inspectable, and queryable substrate. Code will be released.

1 INTRODUCTION

A central challenge in long video understanding is that high-bandwidth video capture is now practical, yet turning unbounded streams into a stable, queryable, and verifiable memory substrate remains difficult. Modern multimodal systems can process substantial visual context, but they still struggle to maintain a usable account of what happened, where it happened, and how scenes evolved over time Grauman et al. (2022); Yang et al. (2025b); Chen et al. (2026b); Du et al. (2024). A system cannot reason about the future or explain the past if earlier observations cannot be retained, localised, and revisited reliably Ha & Schmidhuber (2018); Zhu et al. (2024); DataFlow Team et al. (2026); Hafner et al. (2023); NVIDIA et al. (2025). The practical question is how archives spanning hours, days, or months can remain searchable, auditable, and useful for downstream reasoning under real-world constraints Lin et al. (2026); Niu et al. (2025); Song et al. (2024); Wu et al. (2024); Zhou et al. (2025).

Existing paradigms address only part of this problem. Direct long-context models remain vulnerable to attention dilution as duration grows Zou et al. (2024); Xue et al. (2025a). Compact memory schemes improve tractability but often weaken recoverability once fine-grained evidence has been discarded. This creates a recoverability collapse in which compressed memory preserves semantic gist while discarding the visual evidence needed for later verification Xie et al. (2026); Song et al. (2025); Ma et al. (2025). Recursive agentic approaches can also accumulate memory drift when intermediate reasoning cannot be checked against preserved observations Chen et al. (2026b). Online-tuned designs often compress aggressively or maintain only shallow recent memory Liu et al. (2026b); Wang et al. (2025), whereas long-range-tuned designs often rely on offline preprocessing or memory structures that are costly to update continuously Rege et al. (2026); Wang et al. (2024b). These difficulties motivate a unified view of the problem as the orchestration of indexing, storage, and retrieval rather than longer context alone.

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.

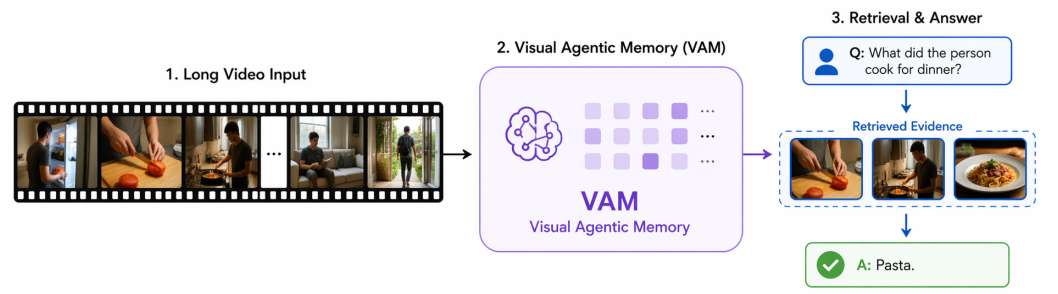


Figure 1: VAM turns continuous video streams into searchable long-horizon memory. Unlike prior offline methods that operate on finite clips and suffer from memory explosion, VAM supports online indexing over effectively unbounded video while keeping each retained frame directly retrievable and visually faithful through adaptive temporal-visual compression, hierarchical memory, and parallel temporal-spatial representations.

Strong pretrained multimodal models offer broad semantic priors but do not by themselves solve evidence retention, temporal disambiguation, provenance tracking, or long-horizon retrieval Alayrac et al. (2022); Li et al. (2023); Bai et al. (2025a); Qwen Team, Alibaba (2026). We therefore formulate long video understanding around three coupled requirements. *Online Indexing* retains visual evidence across temporal and spatial dimensions under streaming constraints Xie et al. (2026); Zeng et al. (2025). *Hierarchical Memory* keeps a temporal layer of event summaries aligned with a spatial layer of recoverable observations rather than only compressed summaries Song et al. (2024); Rege et al. (2026). *Agentic Retrieval* moves between these layers to perform task-relevant search and visual inspection over that memory Yuan et al. (2025); Pan et al. (2025); Fan et al. (2024); Wang et al. (2024b). The central distinction is whether a long-video system can preserve verifiable visual evidence while making it available to the model throughout an extended stream. This framing is deliberately broader than any single benchmark: the same memory substrate can support retrospective question answering, timeline search, event lookup, and later re-inspection whenever a continuous visual stream must remain usable after capture.

To address these constraints, we propose *Visual Agentic Memory* (VAM), a training-free framework that transforms raw video streams into searchable long-horizon memory while preserving the original frames as first-class evidence. In our MM-Lifelong evaluation, this design scales to month-long streams (51 days), enabling the model to operate over durations far beyond ordinary context windows. VAM updates memory incrementally as frames arrive and avoids expensive global preprocessing such as optical flow Xie et al. (2026). It organises retained moments into events through a parallel temporal-spatial representation. Each event maintains a temporal representation for timeline reasoning and a spatial representation containing every retained raw frame and its embedding, so each stored observation remains directly retrievable without replacing the visual evidence with a lossy surrogate. The configured multimodal model writes event summaries from the temporally bounded evidence, capturing the salient activity, entities, and temporal cues within each event. At query time, it also serves as the retrieval agent, moving from temporal search to spatial inspection over the retained frames until it can answer with grounded evidence and cited intervals Pan et al. (2025); Ye et al. (2025).

Our contributions are as follows. (1) We formulate online long-video understanding as an evidence-recoverable visual memory problem, shifting the focus from longer context windows to searchable and inspectable memory. (2) We propose VAM, a training-free framework that incrementally builds memory from streaming video and decouples continuous indexing from query-time reasoning. (3) We introduce a hierarchical parallel temporal-spatial representation that aligns event-level summaries with retained frames, allowing temporal localisation and direct visual verification within the same memory. (4) We show how bounded agentic retrieval can move between these representations to support retrospective question answering, temporal localisation, and direct visual re-inspection without changing the indexed memory. (5) We validate the method on complementary online and long-horizon benchmarks: OVO-Bench tests stream-time responsiveness and backward retrieval, while MM-Lifelong *train@month* tests month-scale access to evidence (105.6 hours, 51 days). The results show that VAM remains effective while storing only a small fraction of the raw stream.

2 Related Work

Online long video understanding must update memory continuously under streaming input and keep it retrievable over long horizons. As duration extends from minutes to months, evidence becomes sparse, repeated events become harder to disambiguate, and retrieval errors propagate more severely Wu et al. (2024); Zhou et al. (2025); Song et al. (2025).

2.1 Methods

Building on increasingly capable video-native MLLMs Maaz et al. (2024); Wang et al. (2024c); Google Developers (2025); Google (2025); Bai et al. (2025a); Qwen Team, Alibaba (2026), current systems fall into three macro-paradigms: direct generative inference, compressed memory modelling, and agentic reasoning.

Direct generative inference. Systems decode answers from in-context visual tokens Lin et al. (2024); Maaz et al. (2024); Chen et al. (2024b). They are typically paired with fixed-rate sampling or bounded-context prompting, which saturates context as duration grows and remains vulnerable to attention dilution Zou et al. (2024). Streaming refinements such as ViSpeak Fu et al. (2025b) and StreamAgent Yang et al. (2025a) strengthen real-time interaction or task anticipation, but stop short of a persistent evidence-preserving architecture.

Compressed memory modelling. These methods retain compact memory rather than all recoverable evidence. Representative examples include MovieChat(+) with sparse retention and summarisation Song et al. (2024; 2025), StreamForest with a progressively compressed hierarchy Zeng et al. (2025), Infinity-Video with continuous-time consolidation Santos et al. (2025), FluxMem with Otsu-derived adaptive thresholds for spatial–temporal redundancy removal Xie et al. (2026), and StreamMind with online memory compression Ding et al. (2025); Liu et al. (2026b). They trade evidence fidelity for bounded storage. Once fine-grained details are discarded, later reasoning cannot revisit the original observations, weakening citation fidelity and retrospective inspection.

Agentic reasoning. These systems interleave planning, search, inspection, and recursive refinement through iterative tool use Rege et al. (2026); Wang et al. (2024b); Yuan et al. (2025); Chen et al. (2026b); Arnab et al. (2025); Yin et al. (2026); Yeo et al. (2026); Chen et al. (2026a); Wang et al. (2023). They differ markedly in their primary substrate. Examples include temporally annotated entity scene graphs (EGAgent Rege et al. (2026)), multi-agent collaboration (LVAgent Chen et al. (2025)), trained "thinking with video" policies (VideoExplorer Yuan et al. (2025)), adaptive reasoning over dynamically constructed hierarchical memory (VideoARM Yin et al. (2026)), multimodal memory combining episodic, semantic, and visual representations (WorldMM Yeo et al. (2026)), long-term multimodal memory for agentic systems (TeleMem Chen et al. (2026a)), LLM-based reasoning over compressed textual descriptions (LifelongMemory Wang et al. (2023)), and recursive belief updating (ReMA, Temporal Chain of Thought Chen et al. (2026b); Arnab et al. (2025)). Tool use does not imply evidence preservation. Most agents do not organise raw RGB observations as a first-class queryable substrate, exposing them to global localisation collapse as the searchable pool grows Chen et al. (2026b). Recursive variants are also sensitive to early hypothesis errors that are inherited rather than corrected through direct lookup. Our evidence-preserving design explicitly retains raw frames alongside event summaries, enabling direct visual verification rather than relying solely on compressed or textual representations.

Retrieval-augmented competitors. Video-RAG and Omni-AdaVideoRAG use visually aligned or adaptive multimodal indexing for long-video retrieval Luo et al. (2024); Xue et al. (2025b). For continuous streams, LiveVLM and StreamMem construct compact KV-cache memories for future-query answering Ning et al. (2025); Yang et al. (2025c).

2.2 Benchmarks

Long video benchmarks fall into *online* benchmarks (timestamp-conditioned, sequential input) and *long-horizon* benchmarks (retrieval and reasoning over extended context). Beyond short-clip predecessors like MovieQA and TVQA Tapaswi et al. (2016); Lei et al. (2018), recent coverage spans StreamingBench, LongVideoBench, MLVU, EgoLifeQA, Video-MME, OVO-Bench, and MM-Lifelong Lin et al. (2026); Wu et al. (2024); Zhou et al. (2025); Yang et al. (2025b); Fu et al. (2025a); Niu et al. (2025); Chen et al. (2026b). We use *OVO-Bench* Niu et al. (2025) for strict online evaluation and *MM-Lifelong train@month* Chen et al. (2026b) for the long-horizon setting. The latter spans 105.6 hours over 51 days with a human–model gap exceeding 70 percentage points,

explicitly exposing the working-memory bottleneck of end-to-end MLLMs and global localisation collapse in agentic baselines.

3 METHODOLOGY

The VAM framework transforms a continuous video stream into a searchable memory trace for retrieval, inspection, and citation. The objective is to retain recoverable visual evidence while keeping long-horizon search tractable. Its temporal and spatial representations provide complementary levels of access: event-level localisation narrows the search, while retained frames support direct verification. Continuous memory construction is decoupled from query-time reasoning, so indexing proceeds as the stream evolves.

3.1 SYSTEM ARCHITECTURE

The framework consists of three coupled components: online indexing (Section 3.2), a hierarchical memory substrate (Section 3.3), and query-time agentic retrieval (Section 3.4). These components are joined by asynchronous orchestration (Section 3.5). Online indexing filters incoming frames and retains visually significant transitions as *moments*; each moment is a frame-level unit that stores a raw observation and its embedding. Contiguous moments are grouped into *events* carrying event summaries and temporal metadata. Event summaries support coarse retrieval, while spatial evidence remains available for direct inspection. At query time, the retrieval agent moves from event-level temporal hypotheses to frame-level spatial evidence and synthesises answers anchored to recoverable visual intervals. The configured multimodal model writes event summaries during indexing and supports visual inspection at query time, providing a shared interface for memory construction and grounded retrieval.

3.2 ONLINE INDEXING

Online indexing transforms the dense raw stream into a sparser yet evidence-preserving memory trace through three consecutive functions. The first filters observations that are unlikely to support later reasoning. The second encodes retained frames for retrieval. The third commits memory boundaries that preserve transitions while compacting quasi-static intervals. Given $\mathcal{F} = \{F_1, \ldots, F_N\}$ with relative and absolute timestamps t_i, T_i, the resulting frame-level entries are *moments*, each storing a retained frame with embedding and timestamps. Contiguous moments later define *events*.

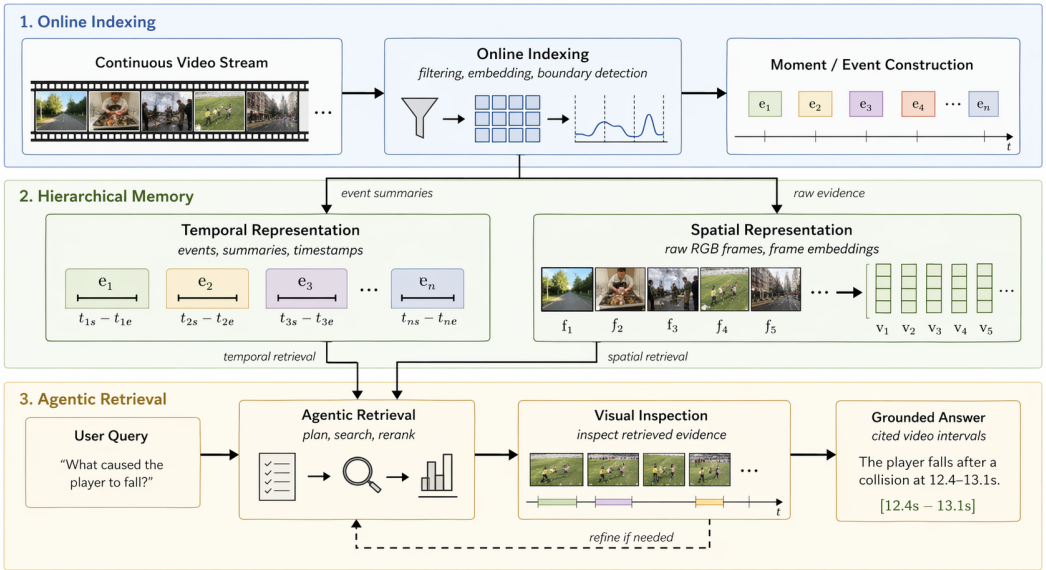


Figure 2: System architecture. Three coupled layers (Online Indexing, Hierarchical Memory, and Agentic Retrieval) feed a parallel temporal/spatial memory. At query time the agent retrieves from both representations and inspects raw frames before answering. The dashed link shows feedback driving additional retrieval.

3.2.1 Stage 1: Frame Filtering

The filtering stage removes observations that are visually unreliable or insufficiently novel before any high-cost neural processing.

Blur Detection (Laplacian Variance). Sharpness is the variance of the Laplacian operator:

$$\text{Sharpness}(I_t) = \text{Var}(\nabla^2 I_{t,\text{gray}}) \tag{1}$$

A blurry image yields lower variance. If $\text{Sharpness}(I_t) < \tau_{blur}$, the frame is discarded.

Redundancy Filtering. Many streams remain dominated by visually near-identical frames even when image quality is acceptable. The system performs pre-embedding redundancy suppression based on visual similarity, placed before dense embedding so repeated content is discarded early. Only frames passing both gates proceed (Figure 3, defaults in Appendix B).

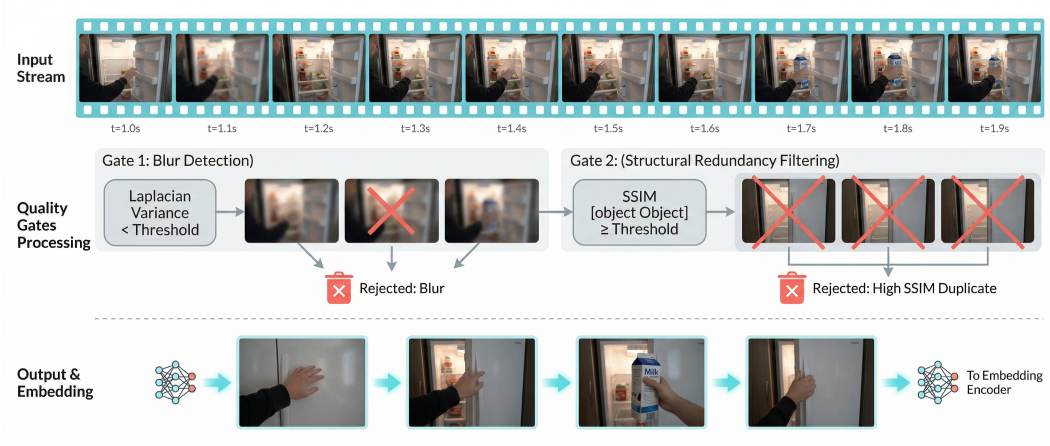


Figure 3: Frame filtering. Top: raw stream. Middle: frames rejected for blur ($< \tau_{blur}$) or redundancy. Bottom: retained keyframes proceeding to embedding.

3.2.2 Stage 2: Embedding Extraction

For each retained frame, the system extracts a dense embedding $\mathbf{e}_t \in \mathbb{R}^d$ using a pre-trained multimodal embedding model Radford et al. (2021); Bain et al. (2021). The specific instantiation is an experimental choice. This embedding model serves as the retrieval-facing representation layer. It supports boundary-sensitive change detection, frame-level retrieval, and query–evidence alignment. The MLLM is reserved for event summarisation, visual inspection, and answer synthesis.

3.2.3 Stage 3: Memory Update and Boundary-Preserving Moment Commit

The system decides whether a new frame provides genuinely new information by comparing each incoming embedding to the current reference:

$$d_c(\mathbf{e}_t, \mathbf{e}_{ref}) = 1 - \frac{\mathbf{e}_t \cdot \mathbf{e}_{ref}}{\|\mathbf{e}_t\|\|\mathbf{e}_{ref}\|} \tag{2}$$

A fixed threshold is brittle because embedding variation depends on scene dynamics, camera motion, and environmental stability. VAM therefore maintains a sliding distance history $\mathcal{H}_d$ and derives an adaptive deduplication gate from recent statistics. With sufficient samples, Otsu partitioning Otsu (1979) is applied over $\mathcal{H}_d$. Otherwise, a configured similarity-derived default is used. This is methodologically close to FluxMem Xie et al. (2026), which applies Otsu thresholds to token reduction in temporal adjacency and spatial domain consolidation rather than to evidence-preserving indexing:

$$\tau_{dedup} \approx \text{AdaptiveThreshold}(\mathcal{H}_d). \tag{3}$$

Frames with $d_c \leq \tau_{dedup}$ are buffered as redundant context. When $d_c > \tau_{dedup}$, the buffered endpoint is committed as a moment and the new frame becomes the reference Rege et al. (2026); Yuan et al. (2025). Figure 4 illustrates this update rule. The same Otsu-based change analysis defines event boundaries: distances between consecutive moments are scanned for local peaks above a relaxed threshold, with a duration cap preventing unbounded events.

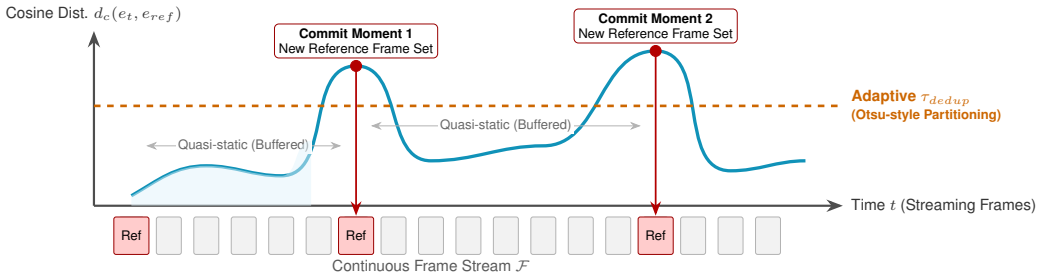


Figure 4: Adaptive deduplication. When d_c exceeds the adaptive τ_{dedup}, buffered moments are committed and the reference is reset. The same distance spikes drive event formation.

At runtime each frame undergoes (1) encoding to $\mathbf{e}_t$, (2) computing d_c against the reference, (3) comparing with τ_{dedup}, and (4) buffering or committing. The process is strictly online, with no future-frame access Rege et al. (2026); Yuan et al. (2025).

3.3 Hierarchical Memory

Embeddings alone are insufficient for episodic retrieval over long timelines. Precise frame matching and abstract event recall fail under different conditions Song et al. (2024); Rege et al. (2026); Yuan et al. (2025). Hierarchical memory organises moments into events so spatial and temporal representations stay synchronised through shared anchors.

Age-aware storage. Retained items occupy *recent*, *mid*, and *long* tiers by age Δt. Compression strength and minimum spacing increase with age: recent windows retain dense full-frame payloads, older segments are consolidated into compact event-level cues that remain retrievable Rege et al. (2026); Yuan et al. (2025).

Event formation. Boundaries come from the same Otsu-derived statistics applied to the moment-distance sequence, with a maximum event duration. Letting m_j denote the j-th moment, consecutive moments between two boundaries form an event:

$$E_i = \{m_{s_i}, \ldots, m_{e_i}\} \tag{4}$$

The MLLM generates an event summary from the same temporally bounded evidence (a clip cropped over the event interval when available, otherwise representative retained frames with metadata):

$$C_i = \text{MLLM}(E_i) \tag{5}$$

Each event maintains (i) retained moments, (ii) embeddings $\mathbf{e}_t$, (iii) summary C_i, and (iv) metadata (t, T). The spatial representation comprises moments and embeddings. The temporal representation consists of event summaries and time-aware retrieval cues. The two pathways share metadata, so retrieval can move from event-level temporal hypotheses to frame-level spatial evidence without alignment loss. Broader *summary documents* are materialised at query time by aggregating event summaries rather than replacing them.

3.4 Agentic Retrieval

Agentic retrieval is a bounded multi-turn loop. At each turn k, the MLLM receives the query Q and history $\mathcal{C}_{k-1}$, and selects an action:

$$a_k \in \{\texttt{search}, \texttt{inspect}, \texttt{summarize}, \texttt{answer}\}.$$

The loop follows the observe–hypothesise–inspect paradigm Rege et al. (2026); Yuan et al. (2025). This paradigm constrains inference to interpretable primitives, which preserves auditability and stabilises long-horizon search.

Intent parsing. Each turn converts the query into an operational specification. For complex requests, the MLLM decomposes user intent into executable sub-queries with explicit temporal constraints, target entities, operator selection, and source prioritisation.

Hybrid temporal-spatial retrieval. `search` consults two pathways within a hybrid retrieval step. Let $\mathbf{q}_k$ denote the turn-k retrieval embedding and $D_{TEMP} = \{d_i\}$ the set of event summaries (or materialised summary documents) with embeddings $\mathbf{v}_{d_i}$. *Temporal retrieval* scores summaries via cosine similarity, $S_{TEMP}(d_i) = \cos(\mathbf{q}_k, \mathbf{v}_{d_i})$, prioritising situational context. Summary documents aggregate overlapping evidence into a single retrieval-oriented description because retrieval over local event summaries alone does not guarantee cross-event consolidation. *Spatial retrieval* compares the query against frame embeddings, $S_{VIS}(F_t) = \cos(\mathbf{q}_k, \mathbf{e}_t)$, pinpointing object-level details. The agent can also use `visual_ref` for appearance-based retrieval. In addition, `inspect_k` enables look-ahead by automatically passing top candidates to inspection.

Grounded inspection. Retrieved candidates remain provisional. The MLLM invokes `inspect` (optionally as `joint_inspection` when comparing temporally related candidates) to evaluate evidence directly:

$$O_k = \text{Inspector}(\mathcal{X}_k^{vis}, \mathcal{X}_k^{temp}, Q) \tag{6}$$

where $\mathcal{X}_k^{vis}, \mathcal{X}_k^{temp}$ are the candidates selected at turn k. Inspection resolves ambiguities, adjudicates conflicts between retrieval signals, and is the primary safeguard against unsupported answers. `summarize` synthesises localised event summaries and frames into a reusable summary document persisted in memory.

Answer synthesis and termination. Termination is an explicit planner action. `answer` is emitted once the context contains sufficient evidence, the turn budget grows small, or successive turns cease to improve results. The answer must be associated with a `best_ref`, typically a `(turn_idx, result_idx)` pointer to the most probative candidate, resolved against the stored turn context so the response remains coupled to a concrete evidence instance. When `best_ref` is absent or invalid, the controller falls back to the highest-scoring retrieved candidate. If progress stagnates or the maximum turn budget K is reached, termination is forced. The output preserves the answer text, evidence sufficiency, principal supporting candidate, the full multi-turn trace, and total search duration.

3.5 Asynchronous Orchestration

Online indexing and agentic retrieval are decoupled but coordinated processes over the shared memory. Indexing runs in the background, filtering frames and writing moments and events. Agentic retrieval is instantiated on demand and interacts with memory only through bounded retrieval operators Lin et al. (2026). Explicit resource bounds (turn budgets, candidate set sizes, and inspection batch limits) keep retrieval cost predictable while cited evidence stays grounded in indexed memory entries.

A practical consequence is that long-horizon retrieval often requires passing multiple candidate intervals or frame sets to the MLLM for direct inspection. This makes evaluation substantially more expensive than retrieval-only or one-pass decoding settings. As in many recent long-video studies, the empirical scope is further constrained by inference budget and serving access for repeated multi-image MLLM calls. The study therefore adopts a targeted rather than exhaustive evaluation design, focusing on representative benchmark subsets that are most informative for online responsiveness, retrospective evidence recovery, and grounded answer generation.

4 Evaluation

VAM is evaluated along two directions: online performance under strict streaming constraints (*OVO-Bench*) and long-horizon performance over extended temporal spans (*MM-Lifelong*). These benchmarks are used as complementary tests of one general memory substrate rather than as isolated leaderboard targets: OVO-Bench probes responsiveness during a live stream, whereas MM-Lifelong probes whether the same design remains usable when evidence is distributed across many days. The Gemini VAM results use Gemini Embedding 2 for memory indexing and retrieval and Gemini 3 Flash for event summarisation, visual inspection, agentic retrieval, and final answer generation. The initial sampling rate is 0.5 fps; retained memory is then formed adaptively rather than stored at a constant fps. The central goal is to extend the model's usable visual reasoning horizon to streams far longer than its native context while keeping every retained frame directly retrievable and visually faithful for later inspection.

Model	Real-Time Visual Perception							Backward Tracing				Forward Active Responding				RT+BT Avg.	Overall
	OCR	ACR	ATR	STU	FPD	OJR	Avg.	EPM	ASI	HLD	Avg.	REC	SSR	CRR	Avg.		
Human Agents	93.96	92.57	94.83	92.70	91.09	94.02	93.20	92.59	93.02	91.37	92.33	95.48	89.67	93.56	92.90	92.77	92.81
Proprietary MLLMs																	
Gemini 3 Flash Google (2025)	86.58	68.81	79.31	58.99	70.30	70.65	72.44	58.92	**76.35**	52.15	**62.47**	35.53	**74.24**	61.67	57.15	67.46	**64.02**
GPT-4o OpenAI (2024)	69.80	64.22	71.55	51.12	70.30	59.78	64.46	57.91	75.68	48.66	60.75	27.58	73.21	59.40	53.40	62.61	59.54
Open-Source Offline MLLMs																	
Qwen2-VL-72B Wang et al. (2024a)	65.77	60.55	69.83	51.69	69.31	54.35	61.92	52.53	60.81	57.53	56.96	38.83	64.07	45.00	49.30	59.44	56.06
LLaVA-Video-7B Zhang et al. (2025b)	69.13	58.72	68.83	49.44	74.26	59.78	63.36	56.23	57.43	7.53	40.40	34.10	69.95	60.42	54.82	51.88	52.86
LLaVA-OneVision-7B Li et al. (2025)	66.44	57.80	73.28	53.37	71.29	61.96	64.02	54.21	55.41	21.51	43.71	25.64	67.09	58.75	50.49	53.86	52.74
Qwen2-VL-7B Wang et al. (2024a)	60.40	50.46	56.03	47.19	66.34	55.43	55.97	47.81	35.48	56.08	46.46	31.66	65.82	48.75	48.74	51.22	50.39
InternVL-V2-8B OpenGVLab Team (2024)	67.11	60.55	63.79	46.07	68.32	56.52	60.39	48.15	57.43	24.73	43.44	26.50	59.14	54.14	46.59	51.92	50.14
LongVU-7B Shen et al. (2025)	53.69	53.21	62.93	47.75	68.32	59.78	57.61	40.74	59.46	4.84	35.01	12.18	69.48	60.83	47.50	46.31	46.71
Qwen3-VL-2B-Instruct (63f batch) Bai et al. (2025a); Wang et al. (2026)	-	-	-	-	-	-	56.65	-	-	-	37.29	-	-	-	51.51	46.97	49.97
Qwen3-VL-4B-Instruct (63f batch) Bai et al. (2025a); Wang et al. (2026)	-	-	-	-	-	-	62.39	-	-	-	44.32	-	-	-	46.29	53.36	50.32
Qwen3-VL-8B-Instruct (63f batch) Bai et al. (2025a); Wang et al. (2026)	-	-	-	-	-	-	62.86	-	-	-	42.63	-	-	-	51.02	52.75	52.54
Qwen2.5-VL-7B (0.5fps)‡ Bai et al. (2025b); Liang et al. (2026)	76.51	57.80	68.10	46.63	66.34	56.52	60.93	49.83	58.11	44.62	50.24	-	-	-	-	55.59	-
Qwen3-VL-8B (0.5fps)‡ Bai et al. (2025a); Liang et al. (2026)	83.89	58.72	70.69	56.18	69.31	64.13	66.79	51.18	60.81	43.55	51.19	-	-	-	-	58.99	-
Open-Source Online MLLMs/Agents																	
Flash-VStream-7B Zhang et al. (2025a)	24.16	29.36	28.45	33.71	25.74	28.80	28.37	39.06	37.16	5.91	27.38	8.02	67.25	60.00	45.09	27.88	33.61
VideoLLM-online-8B Chen et al. (2024a)	8.05	23.85	12.07	14.04	45.54	21.20	20.79	22.22	18.80	12.18	17.73	-	-	-	-	19.26	-
Dispider Qian et al. (2025)	57.72	49.54	62.07	44.94	61.39	51.63	54.55	48.48	55.41	4.30	36.06	18.05	37.36	48.75	34.72	45.31	41.78
Streamo-7B (2fps*) Xia et al. (2026)	77.18	66.06	76.72	45.51	66.34	72.83	67.44	55.56	58.11	33.87	49.18	30.84	57.55	**82.50**	56.96	58.31	57.86
ViSpeak-7B Fu et al. (2025b)	75.17	58.72	71.55	51.12	74.26	66.85	66.28	59.93	48.65	**63.98**	57.52	33.81	68.52	60.42	54.25	61.90	59.35
FluxMem Xie et al. (2026)	81.20	59.60	70.70	53.40	75.20	63.00	67.18	-	-	-	-	-	-	-	-	-	-
ThinkStream-3B Liu et al. (2026b)	85.23	64.22	69.82	49.43	69.31	64.13	67.02	53.87	59.46	43.55	52.29	-	-	-	-	59.66	-
StreamForest-7B Zeng et al. (2025)	68.46	53.21	71.55	47.75	65.35	60.87	61.20	58.92	64.86	32.26	52.01	32.81	70.59	57.08	53.49	56.61	55.57
StreamAgent Yang et al. (2025a)	71.20	53.20	63.60	53.90	67.30	58.70	61.32	54.80	58.10	25.80	46.23	35.90	48.40	52.00	45.43	53.78	50.99
Qwen3-VL-8B + 4f Bai et al. (2025a); Shen et al. (2026)	**94.00**	**85.30**	**82.80**	65.70	**77.20**	**83.20**	**81.40**	51.90	58.10	52.10	54.00	-	-	-	-	67.70	-
OASIS-Qwen3-VL-8B Liang et al. (2026)	91.95	80.73	81.03	67.42	67.33	79.89	78.14	61.95	60.14	47.31	57.21	-	-	-	-	67.68	-
Response-G1-Qwen3-VL-8B Ma et al. (2026)	-	-	-	-	-	-	73.60	-	-	-	52.10	-	-	-	**58.20**	62.85	61.30
TimeChat-Online-7B Yao et al. (2025)	69.80	48.60	64.70	44.90	68.30	55.40	58.60	53.90	62.80	9.10	42.00	-	-	-	-	51.80	-
SAVEMem-Qwen2.5-VL-7B Wu et al. (2026)	91.95	68.81	81.03	65.73	70.30	71.74	74.93	50.84	62.84	37.63	50.44	-	-	-	-	62.69	-
R3-Streaming-7B|4B-Thinking Liu et al. (2026a)	87.92	70.64	73.28	59.55	69.32	70.65	71.89	51.18	65.54	37.10	51.27	**45.13**	57.09	49.58	50.60	61.58	57.92
Ours																	
VAM (Ours; Gemini)	85.23	76.15	81.90	**82.19**	74.26	73.91	78.94	**62.29**	66.22	45.16	57.89	22.78	58.82	65.42	49.01	**68.41**	61.95
VAM (Ours; Qwen)	70.47	56.88	63.79	52.25	66.34	58.70	61.40	50.84	46.62	59.14	52.20	16.05	55.80	54.17	42.01	56.80	51.87

Table 1: OVO-Bench Niu et al. (2025) results across all three modes. The Gemini 3 Flash and Qwen configurations are evaluated with the VAM pipeline under the OVO-Bench protocol; the Qwen configuration uses Qwen3.5-9B for planning and answer judgement, Qwen3-VL-8B-Instruct for multimodal inspection, and Qwen3-VL-Embedding-8B for memory indexing. "63f batch" denotes the raw Qwen3-VL-Instruct batch baselines reported by Think While Watching, whose single-turn evaluation averages 63.23 input frames Wang et al. (2026). ‡ denotes the uniform 0.5-fps Qwen2.5-VL-7B and Qwen3-VL-8B baselines reproduced by the OASIS authors Liang et al. (2026). "-" marks scores not reported by the source. "+4f" denotes the SimpleStream baseline that applies the off-the-shelf Qwen3-VL-8B-Instruct model to the four most recent frames sampled at 1 fps, without training, memory, or retrieval Shen et al. (2026). RT+BT values for the raw Qwen batch baselines and Response-G1 are arithmetic means of their reported group averages. Bold and underline mark the best and second-best reported values in each column, excluding the human row.

4.1 OVO-Bench Results

OVO-Bench Niu et al. (2025) contains 644 videos and 3100 queries, with sources ranging from minutes to roughly half an hour. It tests timestamp-conditioned operation in which each response is grounded only in observations available up to the query time. It decomposes online performance into *Real-Time Visual Perception* (RT), *Backward Tracing* (BT), and *Forward Active Responding* (Fwd). RT and BT are the most informative components for VAM, since they test whether online indexing preserves a usable memory substrate and whether agentic retrieval can localise earlier evidence within it.

As shown in Table 1, the Gemini-backed VAM configuration achieves an RT+BT average of 68.41 and an overall score of 61.95, compared with 67.46 and 64.02 for the end-to-end Gemini 3 Flash baseline, respectively. The Qwen-backed VAM configuration obtains an RT+BT average of 56.80 and an overall score of 51.87, with 59.14 on HLD. These results make the backbone-specific outcomes explicit: the Gemini configuration has the stronger aggregate scores, while the Qwen configuration also evaluates VAM's online memory and retrieval pipeline with a distinct model stack.

For the Gemini configuration, the 0.95-point RT+BT gain over its end-to-end baseline is accompanied by strong Spatial Understanding (82.19) and Episodic Memory (62.29) scores. Its RT gains are distributed across ACR (+7.34), ATR (+2.59), STU (+23.20), and OJR (+3.26), indicating improvements across several online perception capabilities rather than a single category.

Forward Active Responding interprets differently from the other modes. It tests calibration under partial evidence, namely whether the system can withhold commitment until decisive future evidence appears. VAM performs less strongly here because its controller is explicitly evidence-seeking and citation-oriented. Once retrieval surfaces observations that appear sufficient, the agent tends to commit even when the stream has not yet revealed the decisive later cue. The same evidence-first design that aids retrospective grounding can therefore become over-eager in forward tasks, consistent with the gap between VAM's RT+BT performance and its lower Forward Active average.

4.2 MM-Lifelong Results

The month-scale *Live Stream* portion of *MM-Lifelong* Chen et al. (2026b) spans 105.6 hours over 51 days, with evidence separated by large temporal gaps and substantial distractor content. We report *train@month* (266 questions) with Gemini. Following the benchmark protocol, end-to-end MLLMs use uniform sparse sampling up to their context capacity, while agentic baselines follow official defaults. *Ref@300* is scored at 5-minute temporal resolution. VAM makes the stream searchable and verifiable: the temporal representation narrows the search, and the spatial representation provides retained visual evidence for direct inspection. The month-scale storage comparison is given in Appendix Table 3.

Method	Acc.	Ref@300	Method	Acc.	Ref@300
End-to-End MLLMs			Agentic Methods		
GPT-5	10.15	1.39	LongVT-7B	5.83	1.71
Qwen3-VL-235B-A22B	9.09	0.39	ReMA with GPT-5	**17.62**	**9.91**
Qwen3-VL-30B-A3B	8.33	0.48	ReMA with Qwen3VL-A22B	14.23	<u>6.01</u>
Video-XL-2-8B	6.02	0.00	Human	82.5	31.2
Eagle-2.5-8B	3.76	1.59	Ours		
Nemotron-v2-12B	7.52	0.19	VAM (Ours; Gemini)	<u>17.11</u>	3.65
VideoMind-7B	5.26	-			
DeepVideoDiscovery	4.36	-			

Table 2: MM-Lifelong Chen et al. (2026b), with Acc. and Ref@300 reported as percentages. All rows use train@month. Bold/underline mark the best/second-best train@month model scores, excluding human.

The month-scale run retains 6,876 frames, or 0.06% of the raw 11,406,923-frame corpus. This compact representation is queryable through both temporal summaries and retained frames, supporting multiple downstream tasks without rebuilding a task-specific index.

5 Conclusion

We presented *Visual Agentic Memory* (VAM), a training-free framework for making month-long video searchable while keeping retained frames directly retrievable for visual verification. Its temporal representation narrows long-horizon search through event summaries and timeline cues; its spatial representation supports direct inspection. On OVO-Bench, VAM reaches an RT+BT average of 68.41, above the 67.46 end-to-end Gemini 3 Flash baseline. On MM-Lifelong *train@month*, it reaches 17.11% over 105.6 hours and 51 days, second only to ReMA with GPT-5, while retaining 6,876 frames (0.06% of the raw stream; Table 3). Future work includes audio-visual extension, lighter retrieval controllers, and longer-deployment evaluation.

Broader impacts. Long, searchable video streams can expose bystanders, private spaces, health information, routines, and relationships. Incidental capture complicates consent, while unrestricted retention creates risks of unauthorised access and retrospective profiling. In addition, incomplete summaries or retrievals may appear authoritative and cause harm in settings such as workplace monitoring or forensic review. Responsible deployment therefore requires privacy-preserving indexing, retention and deletion controls, and consent-aware access.

AI USE STATEMENT

In this work, we used generative AI tools for copy-editing, readability edits, LaTeX assistance, and preliminary organization of experiment logs and reported results. The authors reviewed the AI-assisted material and take responsibility for the final manuscript and reported results.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Temporal chain of thought: Long-video understanding by thinking in frames, 2025. URL https://arxiv.org/abs/2507.02001.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631. Technical report.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Chunliang Chen, Ming Guan, Xiao Lin, Jiaxu Li, Luxi Lin, Qiyi Wang, Xiangyu Chen, Jixiang Luo, Changzhi Sun, Dell Zhang, and Xuelong Li. Telemem: Building long-term and multimodal memory for agentic ai, 2026a. URL https://arxiv.org/abs/2601.06037.

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

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18407–18418, June 2024a.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision (ECCV)*, 2024b.

DataFlow Team, Bohan Zeng, Daili Hua, Kaixin Zhu, Yifan Dai, Bozhou Li, Yuran Wang, Chengzhuo Tong, Yifan Yang, Mingkun Chang, Jianbin Zhao, Zhou Liu, Hao Liang, Xiaochen Ma, Ruichuan An, Junbo Niu, Zimo Meng, Tianyi Bai, Meiyi Qiang, Huanyao Zhang, Zhiyou Xiao, Tianyu Guo, Qinhan Yu, Runhao Zhao, Zhengpin Li, Xinyi Huang, Yisheng Pan, Yiwen Tang, Juanxi Tian, Yang Shi, Yue Ding, Xinlong Chen, Hongcheng Gao, Minglei Shi, Jialong Wu, Zekun Wang, Yuanxing Zhang, Xintao Wang, Pengfei Wan, Yiren Song, Mike Zheng Shou, and Wentao Zhang. Openworldlib: A unified codebase and definition of advanced world models, 2026. URL https://arxiv.org/abs/2604.04707.

Xin Ding, Hao Wu, Yifan Yang, Shiqi Jiang, Qianxi Zhang, Donglin Bai, Zhibo Chen, and Ting Cao. Streammind: Unlocking full frame rate streaming video dialogue through event-gated cognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Yifan Du, Kun Zhou, Yuqi Huo, Yifan Li, Wayne Xin Zhao, Haoyu Lu, Zijia Zhao, Bingning Wang, Weipeng Chen, and Ji-Rong Wen. Towards event-oriented long video understanding, 2024. URL https://arxiv.org/abs/2406.14129.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision (ECCV)*, pp. 75–92, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24118–24130, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_Video_Analysis_CVPR_2025_paper.html.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Yi-Xing Peng, Kun-Yu Lin, Xihan Wei, Jian-Fang Hu, Xiaohua Xie, and Wei-Shi Zheng. Vispeak: Visual instruction feedback in streaming videos, 2025b. URL https://arxiv.org/abs/2503.12769.

Google. Gemini 3: Introducing the latest gemini ai model from google. https://blog.google/products/gemini/gemini-3/, November 2025. Official Google blog post, Accessed: 2026-04-21.

Google Developers. Advancing the frontier of video understanding with gemini 2.5. https://developers.googleblog.com/gemini-2-5-video-understanding/, May 2025. Google Developers Blog, Accessed: 2026-04-21.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems 31*, pp. 2451–2463. Curran Associates, Inc., 2018. URL https://papers.nips.cc/paper/7512-recurrent-world-models-facilitate-policy-evolution. https://worldmodels.github.io.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2023. URL https://arxiv.org/abs/2301.04104.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1369–1379, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1167. URL https://aclanthology.org/D18-1167/.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Zhijia Liang, Jiaming Li, Weikai Chen, Yanhao Zhang, Haonan Lu, and Guanbin Li. Oasis: On-demand hierarchical event memory for streaming video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Liang_OASIS_On-Demand_Hierarchical_Event_Memory_for_Streaming_Video_Reasoning_CVPR_2026_paper.pdf.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.342. URL https://aclanthology.org/2024.emnlp-main.342/.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12147–12151, 2026.

Jinming Liu, Jianguo Huang, Zhaoyang Jia, Jiahao Li, Xiaoyi Zhang, Zongyu Guo, Bin Li, Wenjun Zeng, Yan Lu, and Xin Jin. An efficient streaming video understanding framework with agentic control, 2026a. URL https://arxiv.org/abs/2605.17921.

Zikang Liu, Longteng Guo, Handong Li, Ru Zhen, Xingjian He, Ruyi Ji, Xiaoming Ren, Yanhao Zhang, Haonan Lu, and Jing Liu. Thinking in streaming video, 2026b. URL https://arxiv.org/abs/2603.12938.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-rag: Visually-aligned retrieval-augmented long video comprehension, 2024. URL https://arxiv.org/abs/2411.13093.

Ke Ma, Jiaqi Tang, Bin Guo, Xueting Han, Ruonan Xu, Qingfeng He, Ziheng Wang, Xu Wang, Qifeng Chen, Zhiwen Yu, and Yunhao Liu. Response-g1: Explicit scene graph modeling for proactive streaming video understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. URL https://aclanthology.org/2026.acl-long.2042/.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18936–18946, June 2025.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 12585–12602, 2024.

Zhenyu Ning, Guangda Liu, Qihao Jin, Chengwei Li, Wenchao Ding, Minyi Guo, and Jieru Zhao. Livevlm: Efficient online video understanding via streaming-oriented kv cache and retrieval, 2025. URL https://arxiv.org/abs/2505.15269.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. Ovo-bench: How far is your video-llms from real-world online video understanding? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18902–18913, 2025. doi: 10.1109/CVPR52734.2025.01761.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klar, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. URL https://arxiv.org/abs/2501.03575.

OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, May 2024. Official OpenAI blog post, Accessed: 2026-04-22.

OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024. URL https://internvl.github.io/blog/2024-07-02-InternVL-2.0/. Official InternVL2 blog post.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. Timesearch-r: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning, 2025. URL https://arxiv.org/abs/2511.05489.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Qwen Team, Alibaba. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026. Official blog post, Accessed: 2026-03-21.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. Agentic very long video understanding, 2026. URL https://arxiv.org/abs/2601.18157.

Saul Santos, Antonio Farinhas, Daniel C. McNamee, and Andre F. T. Martins. Infinity-video: A training-free approach to long video understanding via continuous-time memory consolidation, 2025. URL https://arxiv.org/abs/2501.19098.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599. PMLR, 2025. URL https://proceedings.mlr.press/v267/shen25j.html.

Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. A simple baseline for streaming video understanding. *arXiv preprint arXiv:2604.02317*, 2026. URL https://arxiv.org/abs/2604.02317.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 374–389, 2025. doi: 10.1109/TPAMI.2025.3604614.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. Streambridge: Turning your offline video large language model into a proactive streaming assistant. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Lu Wang, Zhuoran Jin, Yupu Hao, Yubo Chen, Kang Liu, Yulong Ao, and Jun Zhao. Think while watching: Online streaming segment-level memory for multi-turn video reasoning in multimodal large language models, 2026. URL https://arxiv.org/abs/2603.11896.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a. URL https://arxiv.org/abs/2409.12191.

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

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision – ECCV 2024*, pp. 396–416, 2024c. doi: 10.1007/978-3-031-73013-9_23. URL https://doi.org/10.1007/978-3-031-73013-9_23.

Ying Wang, Yanlai Yang, and Mengye Ren. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos, 2023. URL https://arxiv.org/abs/2312.05269.

Hang Wu, Sherin Mary Mathews, Yujun Cai, Ming-Hsuan Yang, and Yiwei Wang. Semantic-aware adaptive visual memory for streaming video understanding, 2026. URL https://arxiv.org/abs/2605.07897.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: a benchmark for long-context interleaved video-language understanding. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Jiaer Xia, Peixian Chen, Mengdan Zhang, Xing Sun, and Kaiyang Zhou. Streaming video instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yiweng Xie, Bo He, Junke Wang, Xiangyu Zheng, Ziyi Ye, and Zuxuan Wu. Fluxmem: Adaptive hierarchical memory for streaming video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Lijuan Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations (ICLR)*, 2025a.

Zhucun Xue, Jiangning Zhang, Xurong Xie, Yuxuan Cai, Yong Liu, Xiangtai Li, and Dacheng Tao. Adavideorag: Omni-contextual adaptive retrieval-augmented efficient long video understanding, 2025b. URL https://arxiv.org/abs/2506.13589.

Haolin Yang, Feilong Tang, Lingxiao Zhao, Xiang An, Ming Hu, Huifa Li, Xinlin Zhuang, Yifan Lu, Xiaofeng Zhang, Abdalla Swikir, Junjun He, Zongyuan Ge, and Imran Razzak. Streamagent: Towards anticipatory agents for streaming video understanding, 2025a. URL https://arxiv.org/abs/2508.01875.

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

Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. Streammem: Query-agnostic kv cache memory for streaming video understanding, 2025c. URL https://arxiv.org/abs/2508.15717.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, et al. Timechat-online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10807–10816, 2025. URL https://arxiv.org/abs/2504.17343.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. T*: Re-thinking temporal search for long-form video understanding, 2025. URL https://arxiv.org/abs/2504.02259.

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

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. Videoarm: Agentic reasoning over hierarchical memory for long-form video understanding, 2026. URL https://arxiv.org/abs/2512.12360.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. Videoexplorer: Think with videos for agentic long-video understanding, 2025. URL https://arxiv.org/abs/2506.10821.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, Yi Wang, and Limin Wang. Streamforest: Efficient online video understanding with persistent event memory, 2025. URL https://arxiv.org/abs/2509.24871.

Setting	# Stored Frames	Share of Raw Video	Multiple of VAM
Raw video frames	11,406,923	100.00%	1658.9× more
Uniform 1 fps sampling	380,304	3.33%	55.3× more
Uniform 0.5 fps sampling	190,152	1.67%	27.7× more
VAM output	6,876	0.06%	1.0×

Table 3: Retained-frame counts under different settings on the MM-Lifelong month-scale run. VAM keeps 27.7–55.3× fewer frames than uniform fixed-rate baselines and 1658.9× fewer than the raw stream.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025b.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701, 2025. doi: 10.1109/CVPR52734.2025.01278.

Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Bohan Li, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, Yang You, Zhaoxiang Zhang, Dawei Zhao, Liang Xiao, Jian Zhao, Jiwen Lu, and Guan Huang. Is sora a world simulator? a comprehensive survey on general world models and beyond, 2024. URL https://arxiv.org/abs/2405.03520.

Heqing Zou, Tianze Luo, Guiyang Xie, Victor (Xiao Jie) Zhang, Fengmao Lv, Guangcong Wang, Junyang Chen, Zhuochen Wang, Hansheng Zhang, and Huaijian Zhang. From seconds to hours: Reviewing multimodal large language models on comprehensive long video understanding, 2024. URL https://arxiv.org/abs/2409.18938.

A Limitations

Modality asymmetry. VAM retains selected frames for direct visual inspection, although older frames may be resized and compressed for storage. It does not preserve audio at comparable granularity, limiting verification of speech, ambient sound, and other non-visual cues.

Long-term scaling. Hierarchical organisation improves tractability, but the search space becomes increasingly crowded as archives grow from hours to months and potentially to year-long deployments. Coarse temporal localisation may remain feasible, while fine-grained disambiguation becomes harder. Storing more evidence improves recoverability, but it also increases the burden on ranking and verification at query time. Answer quality also depends on earlier indexing decisions. A missed boundary or unretained subtle state change cannot be recovered downstream.

Operational cost. The pipeline includes continuous indexing, embedding, hierarchical storage, and iterative retrieval and inspection. This introduces substantial inference latency and serving cost compared with simpler end-to-end or compressed alternatives, especially in long-horizon settings where each query may require broad search and repeated visual inspection. The reported results establish feasibility, but they do not yet resolve the engineering requirements of persistent, low-cost, year-scale multimodal memory.

B Implementation Defaults

B.1 Month-scale Retained-frame Count

B.2 Model Access and Serving Setup

The Gemini VAM results use Gemini 3 Flash and Gemini Embedding 2 through OpenRouter and Google Cloud APIs rather than self-hosted inference servers. Gemini Embedding 2 supplies frame, document, and query embeddings, while Gemini 3 Flash performs event and document generation, planning, visual inspection, and final answer generation. Reproduction therefore depends on comparable model access and serving configuration, together with the implementation defaults described below.

Table 4: Default online indexing parameters in our implementation.

Parameter	Default	Role in the pipeline
Base sampling rate	0.5 FPS	Initial extraction rate before filtering.
Maximum extracted frames	`None`	No global hard cap by default. Extraction is governed by the filtering stages.
τ_{blur}	20.0	Minimum Laplacian variance for sharpness filtering.
τ_{diff}	20.0	Mean absolute grayscale difference threshold for static-frame rejection.
τ_{ssim}	0.92	Structural redundancy threshold. Frames above this similarity are treated as near-duplicates.
τ_{hist}	0.0	Histogram-based redundancy gate, kept permissive by default.
Fallback τ_{sim}	0.88	Fallback embedding-space similarity threshold used when adaptive semantic deduplication cannot yet be estimated reliably.
Fallback τ_{event}	0.80	Fallback event-similarity threshold used when adaptive event-boundary estimation lacks sufficient history.
Maximum event duration	300.0 s	Limits how long a single event can grow before forced segmentation.

B.3 Online Indexing Parameters

Our implementation performs online indexing through an FFmpeg-based extraction stage, followed by lightweight heuristic filtering and embedding-level memory updates. In practice, the pipeline first removes obviously uninformative frames using blur and low-change gates. It then applies semantic comparison in the embedding space. Table 4 summarises the configuration defaults that initialise this process.

Blur and low-change gates run before embedding. Semantic deduplication and event segmentation use adaptive estimates over recent embedding distances, with the listed thresholds as fallbacks when estimates are unavailable.

B.4 Hierarchical Memory Defaults

Our implementation uses a three-tier hierarchical memory policy that trades spatial fidelity against temporal coverage as observations age. More recent frames are retained densely and without additional recompression, whereas older frames are thinned more aggressively and stored under stronger image compression. The default retention schedule is summarised in Table 5.

Table 5: Default hierarchical memory configuration.

Tier	Time Window	Min Gap (s)	Max Side (px)	JPEG Quality
recent	$t \leq 1$ hour	1.0	—	No additional recompression
mid	1 hour $< t \leq 24$ hours	20.0	768	70
long	$t > 24$ hours	120.0	512	45

The earliest and latest frames, along with event-linked long-term frames, are preserved explicitly. Compression is applied after frame selection.

B.5 Agentic Retrieval Constraints

The retrieval controller is implemented with strict schema validation rather than free-form prompting alone. Concretely, action payloads are parsed through bounded Pydantic schemas with extra keys forbidden, ensuring that the MLLM operates within an explicit control interface. At implementation level, these schemas inherit from a strict base model with extra keys forbidden globally. Controller outputs are therefore validated against an executable interface rather than loosely interpreted after generation. Table 6 lists the principal action- and field-level constraints exposed by the current controller.

The controller can alternate among `search`, `inspect`, and `summarize`; time ranges may be direct, anchor-relative, or bounded by two anchors.

B.6 Runtime Orchestration and Model Calls

The implementation instantiates three operational MLLM roles inside the same overall VAM architecture. A lightweight router handles top-level asynchronous orchestration, a main planner drives

Table 6: Key constraints in the current agentic retrieval schema.

Action / Field	Constraint	Meaning
search.top_k	1 to 200	Maximum number of candidates retrieved per query.
search.inspect_k	0 to 50	Number of top candidates passed to Visual Inspection, clipped to top_k.
search.threshold	≥ 0.0	Minimum retrieval threshold. The default schema value is 0.65.
inspect.max_frames	1 to 32	Maximum number of frames examined in a direct Visual Inspection step.
summarize. granularity_seconds	$(0, 86400]$	Resolution of reusable summary windows.
time_range.mode	relative, absolute, auto	Supports both uploaded videos and absolute live/history timestamps.
anchor.before_seconds anchor.after_seconds	$[0, 604800]$	Bounds event-relative windows to at most seven days on each side of the anchor.
anchor.top_k anchor.inspect_k	1 to 50 / 0 to 20	Controls anchor-candidate retrieval and subsequent inspection budget. inspect_k is clipped to top_k.
anchor. occurrence_index	1 to 100	Allows ordinal reference to repeated anchor events when such indexing is needed.
anchor. verification_prompt	optional, max 400 chars	Supports strict anchor confirmation without allowing arbitrarily long side instructions.
sources	subset of {frame, event, summary}	Allows the planner to explicitly target different memory structures.
summary_filter	optional	Restricts summary retrieval by structure label or granularity.

agentic retrieval, and a multimodal inspector/generator handles event-document writing, summary-document writing, direct visual inspection, and anchor verification. The Gemini experiments use Gemini 3 Flash and Gemini Embedding 2 through OpenRouter and Google Cloud. Multimodal messages are represented as a sequence of typed content blocks combining text, image URLs, and optional video URLs.

The route labels below identify logical call paths. In the released configuration they resolve to the configured provider model unless separate route-specific model identifiers are supplied.

Two implementation details are important for reproducing the implementation behaviour. First, multimodal inputs are normalised into a single OpenRouter-compatible content list containing one text block followed by image and optional video blocks. Second, images are automatically resized or JPEG-compressed before upload to satisfy configured limits of at most 5 MB per image and at most 1280×1280 pixels. When an event-level clip is attached, it is produced on the fly from the source video with FFmpeg using H.264 video, AAC audio, veryfast preset, and crf=28; if the provider rejects video input, the system retries with frames only.

B.7 Structured Interfaces Used by the Released System

The implementation does not treat prompts as unconstrained free-form text generation. Instead, every routing and planning decision is parsed against strict Pydantic schemas with extra="forbid", so additional keys are rejected rather than ignored.

Top-level routing interface. The router prompt requires exactly one JSON object and allows only three outputs: a final object with a non-empty text field; a retrieve tool call whose arguments contain exactly one field, question; and a summarize tool call whose arguments contain min_time, max_time, time_mode, granularity_seconds, and prompt, with an optional summary_structure. At this stage, retrieve is the default tool for ordinary question answering, while summarize is reserved for explicit timeline-style or reusable summary-document

Table 7: Runtime LLM calls in the implementation.

Function	Route	Decoding	Input and operational role
Top-level request router	light	$T = 0.0$, max tokens $= 300$	Receives the user transcript and chooses exactly one next step: `final`, `retrieve`, or `summarize`.
Agentic retrieval planner	main	$T = 0.0$, max tokens $= 1200$; retry $= 2400$	Receives goal, retrieved context, remaining-turn budget, and available summary structures; outputs one structured action among `answer/search/inspect/summarize`.
Event memory document writer	multimodal main	$T = 0.0$, max tokens $= 1000$	Receives a completed event with up to eight representative frames and, when available, an event video clip; writes a retrieval-oriented event document.
Summary-window writer	multimodal main	$T = 0.0$, max tokens $= 1000$	Receives a time window, overlapping event documents, and representative frames; writes a reusable summary document.
Multimodal inspector	multimodal main	$T = 0.0$, max tokens $= 1000$	Receives multiple frames and a direct visual question; returns evidence-grounded inspection text.
Anchor adjudicator	multimodal main	$T = 0.0$, max tokens $= 500$	Receives a candidate frame sequence and a target event specification; returns strict JSON with `match`, `confidence`, `observed_event`, and `reason`.
Session summariser	light	$T = 0.2$, max tokens $= 500$	Receives recent dialogue and returns a compact event document plus directly evidenced user habits.

requests. The summarization call enforces positive granularity, non-empty prompts, and either an open or bounded time interval.

Planner action interface. The agentic retrieval planner is limited to four actions: `answer`, `search`, `inspect`, and `summarize`. The `search` action contains a non-empty list of query objects, where each object has keys `q`, `top_k`, `inspect_k`, and `threshold`; `top_k` is capped at 200 and `inspect_k` is clipped not to exceed `top_k`. The `inspect` action requires either an explicit time range or a prior result reference, with `max_frames` bounded in $[1, 32]$. The `summarize` action requires a valid time range, a positive window size, and a non-empty prompt. Anchored time ranges support both single-anchor and between-anchor retrieval, but the schema forbids mixing them in one call. Verification prompts for anchor adjudication are capped at 400 characters, and anchor windows are bounded to at most seven days on each side.

Planner execution policy. If the planner output is invalid JSON on the first attempt, the implementation adds one repair instruction requesting "ONLY one valid JSON object" and re-queries the model once. If parsing still fails on the first turn, the controller falls back to a permissive `search` action rather than terminating the loop. This behaviour is used to keep the released system robust to occasional formatting drift while still preserving a strictly structured control surface.

B.8 Prompt Implementations by Agent

To make the appendix self-contained, this subsection organises the core prompt text and structured interfaces by agent role, with the main operational prompts reproduced in code-block form.

B.8.1 Routing Agent

The routing agent performs the first-stage decision in asynchronous orchestration. It chooses among returning a direct answer, invoking retrieval, or creating reusable summary documents.

```
You are the top-level routing module for Asynchronous Orchestration in a
    Visual
Agentic Memory system.
You receive one user request and must choose exactly one next step.
Reply with ONLY one JSON object.
Allowed outputs:
1) {"type":"final","text":"..."}
2) {"type":"tool","name":"retrieve","args":{"question":"..."}}
```

```
3) {"type":"tool","name":"summarize","args":{"min_time":0.0,"max_time
    ":1800.0,
     "time_mode":"relative","granularity_seconds":60.0,"prompt":"..."}}
Tool selection rules:
- Use 'retrieve' for direct questions about content, objects, actions,
    locations,
  or events stored in Hierarchical Memory.
- Use 'summarize' only when the user explicitly wants a timeline,
    segmented recap,
  interval-by-interval summary, or reusable summary documents over a
    stated time span.
- The summarize tool creates new searchable summary documents inside
    Hierarchical
  Memory for each time window. It is not just a one-off answer.
- If the user asks a normal question like 'what happened', 'where is X',
  'when did Y happen', or 'what is in the video', prefer 'retrieve', not
    'summarize'.
- Use 'final' only for requests that do not require memory retrieval or
    summarization.
Representative summarize requests include:
- 'summarize the first 30 minutes minute by minute'
- 'give me a 5-minute timeline of the first hour'
- 'summarize 0 to 1800 seconds so we can ask questions about that segment
     later'
When using 'retrieve':
- Rewrite the user's request into one retrieval-friendly English question
    .
- Preserve important objects, actions, attributes, and time clues,
    including
  relative-event phrases like 'before X', 'after Y', or 'around the
    moment when Z happened'.
- Put the rewrite in args.question.
When using 'summarize':
- Set min_time and max_time to the requested span.
- Use granularity_seconds for the requested window size, e.g. 60 for
    every minute,
  300 for every 5 minutes.
- You may optionally add summary_structure if you want to tag the
    summaries with
  a reusable structure label.
- Use time_mode='relative' for ordinary indexed videos unless the user
    clearly
  refers to absolute/live timestamps.
- Write the prompt as the summary objective, e.g. 'Summarize each minute
    for later QA'.
- Write the summary prompt in English.
- Do not use summarize without a meaningful time range and granularity.
Schema rules:
- 'type' must be either 'final' or 'tool'.
- If 'type' is 'tool', then 'name' must be either 'retrieve' or '
    summarize'.
- If 'name' is 'retrieve', args must contain exactly one key: 'question'.
- If 'name' is 'summarize', args must contain: min_time, max_time (
    optional),
  time_mode, granularity_seconds, prompt. summary_structure is optional.
- Do not return empty strings.
- Do not include any extra keys.



Top-level routing:
- final.text must be non-empty
- summarize.min_time >= 0
- summarize.max_time >= min_time when present
- summarize.granularity_seconds in (0, 86400]
```

```
- summary_structure length <= 64
- extra fields are forbidden
```

B.8.2 PLANNING AGENT

The planning agent is the main controller for agentic retrieval. It decides whether to answer, continue searching, inspect known evidence directly, or materialise new summary documents.

```
You are the Main LLM for Agentic Retrieval in a Visual Agentic Memory
    system.
Your goal is to answer the user's question by identifying the most
    relevant visual
evidence across the Parallel Representation.
You have access to three tools:
1. 'search': Perform Agentic Retrieval over frame evidence and Temporal
   Representation documents.
2. 'inspect': Perform direct Visual Inspection on frames from a known
    time or
   from a previously found result.
3. 'summarize': Create reusable summary documents in Hierarchical Memory
    for a
   specific time range at a requested granularity.
Tool boundary rules:
- 'search' is the default tool for answering questions.
- Use 'inspect' when you already know the relevant time or result
    reference and
  want direct visual confirmation without another semantic retrieval step
    .
- 'summarize' is for creating persistent summary documents over a time
    range.
  Use it only when the user wants interval summaries or reusable summary
    documents
  in Hierarchical Memory.
- If the user asks a normal QA question, prefer 'search' first. Do not
    jump to
  'summarize' unless the user clearly wants timeline-style summaries.

Process:
1. Analyze the user's request and what you have found so far.
2. Decide whether you have enough information to answer.
3. If YES, output {"action":"answer","response":"...","best_ref":{"
    turn_idx":int,
   "result_idx":int},"thought":"..."}
4. If NO, choose a tool:
   - 'search': {"action":"search","queries":[{"q":"...","top_k":int,
     "inspect_k":int,"threshold":float}],"time_range":{...},"sources":
     ["frame"|"event"|"summary"],"summary_filter":{"summary_structure
    ":"...",
     "granularity_seconds":float},"visual_ref":{...},"joint_inspection":
    bool,
     "inspection_prompt":"...","thought":"..."}
   - 'inspect': {"action":"inspect","prompt":"...","time_range":{...},
     "ref":{"turn_idx":int,"result_idx":int},"max_frames":int,"thought
    ":"..."}
   - 'summarize': {"action":"summarize","time_range":{...},
     "granularity_seconds":float,"prompt":"...","summary_structure
    ":"...",
     "thought":"..."}

Schema rules:
- 'action' must be one of: 'answer', 'search', 'inspect', 'summarize'.
- For 'answer', 'response' must be non-empty.
```

```
- For 'search', 'queries' must be a non-empty list of objects with keys:
  q, top_k, inspect_k, threshold.
- For 'inspect', provide a non-empty 'prompt' and either 'time_range' or
    'ref'.
- For 'summarize', provide a non-empty 'prompt', a valid 'time_range',
    and a
  positive 'granularity_seconds'. summary_structure is optional.
- 'turn_idx' is 1-based. 'result_idx' is 0-based.
- Do not include unsupported tools or fields.
- Do not answer unless the current evidence is specific enough to support
    the claim.
- If evidence is weak, partial, or only loosely related, continue with '
    search'
  instead of forcing 'answer'.
- Use English wording for internal search queries, inspection prompts,
    and anchor
  descriptions.
- Keep the final answer in the user's language unless the user explicitly
    requests
  another language.

Tool details and strategy:
- queries: each query object contains q, top_k, inspect_k, threshold.
- sources: any subset of ['frame', 'event', 'summary'].
- summary_filter: optional filter for summary retrieval.
- time_range: supports direct intervals, a single anchor, or a pair of
  start_anchor and end_anchor.
- Do not guess first/second/last occurrences. First collect candidate
    events and
  inspect their returned times.
- Reuse prior result references when they are already available.
- Prefer inspect over search when the relevant time or reference is
    already known.
- If one method fails, switch retrieval method rather than repeating the
    same call.
Always decide the top_k, inspect_k, and which specific reference supports
    the
final answer.
Output ONLY valid JSON.




Goal: {goal}
Available summary structures:
{available_summary_structures}

Context (what we found so far):
{context}

Remaining turns: {remaining_turns}
If remaining_turns <= 2 and you already have useful evidence, prefer
    action=answer.
If results are not improving across turns, stop searching and answer with
    best
available evidence.

Reminder: if the goal is phrased relative to another event such as '
    before X'
or 'after Y', choose the method that best matches the uncertainty. If you
    still
need to figure out which occurrence matters, first search for that event,
    inspect
the returned times, and use those times or refs in the next turn. If
    context
```

```
already lists ordered candidate refs for that event, pick from those refs
     before
doing any broad inspection. If you already know the relevant time or
    reference,
prefer action=inspect instead of another vague search. If an anchor
    attempt failed
earlier, explicitly change the method instead of repeating the same
    source groups
and query.

What is the next step?



Planner constraints:
- search.top_k in [1, 200]
- search.inspect_k in [0, 50], clipped to top_k
- search.threshold >= 0
- inspect.max_frames in [1, 32]
- inspect requires either time_range or ref
- summarize.granularity_seconds in (0, 86400]
- time_range cannot mix anchor with start_anchor/end_anchor
- anchor.before_seconds and anchor.after_seconds in [0, 604800]
- anchor.top_k in [1, 50]
- anchor.inspect_k in [0, 20], clipped to top_k
- anchor.occurrence_index in [1, 100]
- anchor.verification_prompt length <= 400
- extra fields are forbidden
```

B.8.3 MEMORY WRITING AGENT

The memory-writing agent is responsible for converting bounded visual evidence into reusable textual memory objects at both the event level and the summary-window level.

```
System prompt:
You are building a retrieval document for visual memory.
Given a completed event, you may receive both a full event video clip and
representative frames. Use the full video clip as the primary source when
     available.
Write a detailed, factual memory document that remains easy to retrieve
    later.
Requirements:
- Include the time range explicitly.
- Describe people, objects, actions, scene changes, and on-screen text.
- Prefer a few short natural lines instead of one dense paragraph when
    there are
  multiple distinct observations or micro-steps.
- Each line can capture one salient action, object state, scene change,
    or visible text.
- Make the description rich enough for later retrieval by text.
- Stay objective. Keep it natural. No rigid schema or numbered template.

User payload:
{
  "task": "write_retrieval_memory_document",
  "relative_time": {"start_t": start_t, "end_t": end_t},
  "absolute_time": {"start_t": absolute_start_t, "end_t": absolute_end_t
    },
  "notes": [
    "Mention the time range in the description.",
    "Prefer short natural lines when there are multiple distinct
    observations.",
```

```
    "One line can describe one salient action, object state, scene change
    , or visible text."
  ]
}



System prompt:
You are building reusable timeline summaries for a video-memory system.
You will receive a time window, some event documents that overlap that
    window,
and representative frames.
Write a factual retrieval note for that window.
Requirements:
- Mention the time window clearly.
- Prefer concrete observations over speculation.
- Synthesize both the provided document context and the visible frames.
- Make the summary retrieval-friendly for later question answering.
- Prefer a few short natural lines over one dense paragraph when that
    helps
  preserve distinct moments.
- Keep it natural. No rigid schema or numbered template.

User payload:
{
  "task": "summarize_time_window_for_retrieval",
  "time_range": {"start": start_t, "end": end_t, "mode": time_mode},
  "granularity_seconds": granularity_s,
  "focus": prompt,
  "event_documents": event_documents,
  "summary_structure": "optional",
  "notes": [
    "Prefer a few concise natural lines when the window contains distinct
     moments.",
    "Ground the summary in the provided frames and event documents.",
    "This summary will become a searchable memory document for later QA
    .",
    "Follow the natural-language focus first; any structure label is only
     an optional tag."
  ]
}
```

B.8.4 INSPECTION AGENT

The inspection agent operates on already selected frames and performs direct visual verification. A separate anchor-adjudication prompt is used when the system must decide whether a candidate sequence truly contains the target event.

```
System prompt:
You are the Multimodal Inspector for a Visual Agentic Memory system.
You are performing Visual Inspection over a sequence of video frames.
Analyze the frames strictly with respect to the query.
Describe what happens across the frames when it is relevant to the query.
If the frames contain relevant visual evidence, describe it in detail.
If the frames are irrelevant to the query, clearly say so.

User prompt:
User Query: {query}

Please examine these {n} frames and answer the query.
```

```
System prompt:
You are validating whether a short sequence of frames contains a
    requested
anchor event.
Be strict. Do not mark a match unless the event itself is visually
    supported.
Reply with ONLY one JSON object using this schema:
{"match": true|false, "confidence": 0.0-1.0, "observed_event": "...",
 "reason": "..."}

User prompt template:
Target event: {target_event}
Verification focus: {verification_prompt}
Candidate hint: {candidate_hint}
Candidate window: {candidate_window}
Inspect the frames carefully. Set match=true only if the target event
    itself is
visible or strongly supported by the sequence.
```