EVENT-REHEAD: EVENT-AWARE HEAD RETRIEVAL AND MEMORY-AUGMENTED STREAMING VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Language Models (VLMs) have achieved strong performance in video understanding, while streaming video question answering remains challenging due to the need for accurate retrieval under strict memory and latency constraints. Existing KV-cache-based methods mainly organize video memory as independent visual tokens, which ignores underlying temporal semantics of videos and leads to fragmented retrieval, failing to capture continuous evidence . We propose Event-ReHead, a training-free framework that improves streaming video understanding by introducing event semantics into KV-cache retrieval. Our method preserves the efficiency of frame-level retrieval while reorganizing retrieved evidence into an *event skeleton*, a compact and coherent representation capturing the underlying temporal structure. Building on this, we introduce a query-specific attention head selection strategy to construct more discriminative retrieval representations, enabling precise alignment with question-relevant visual signals. We further incorporate a lightweight interaction memory that selectively reuses prior question-answer context, providing useful temporal priors without introducing noise. We evaluate Event-ReHead on four streaming and three offline benchmarks, demonstrating consistent improvements across diverse VLM backbones. In streaming settings, it achieves up to +4.1% gains, while also yielding consistent improvements in offline scenarios, with an average gain of around +2–3%.

1 INTRODUCTION

Recent advances in Vision-Language Models (VLMs) have significantly improved video understanding across tasks such as captioning, question answering, and retrieval Bai et al. (2025a;b); Li et al. (2024a). However, most existing approaches Tao et al. (2025b); Wang et al. (2025b); Yang et al. (2025b) operate in an offline setting, where the entire video is processed jointly with the query. This paradigm becomes increasingly impractical for long videos due to high latency and memory costs Huang et al. (2020), and fails to reflect real-world scenarios such as autonomous driving, embodied agents, and surveillance, where visual streams arrive continuously and queries may be issued at arbitrary times Xiong et al. (2025); Yang et al. (2025a); Wang et al. (2025a).

Streaming video understanding Xiong et al. (2025); Huang et al. (2020) introduces a fundamentally different requirement: models must incrementally encode incoming frames, maintain a compact memory of past observations, and efficiently retrieve relevant evidence upon query Wang et al. (2025a). Recent KV-cache-based methods (e.g., ReKV Di et al. and HERMES Zhang et al. (2026)) address this challenge by compressing accumulated visual tokens and performing retrieval over the cached representations. While effective, these methods treat video memory as a collection of flat visual blocks, leading to several structural limitations.

(I) **Semantic Fragmentation.** Retrieval is typically performed at the frame or block level, ignoring the event-level structure of videos Lin et al. (2026). As many queries depend on temporally coherent events, retrieving isolated frames often produces fragmented evidence that fails to capture complete semantics.

(II) Uniform Head Aggregation. Existing methods Di et al.; Ning et al. (2025) aggregate representations across all attention heads when constructing retrieval keys. However, attention heads in

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.

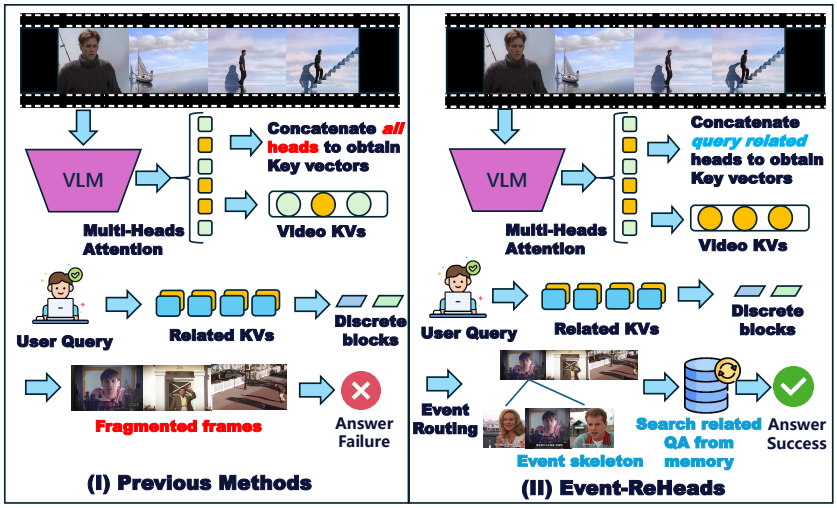

Figure 1: Compared with previous KV-cache retrieval methods that aggregate all heads and retrieve fragmented visual blocks, Event-ReHead selects query-related heads, routes retrieval to coherent event skeletons, and reuses relevant QA memory for more reliable streaming video understanding.

VLMs are highly heterogeneous, capturing different aspects such as objects, actions, or background patterns Xiao et al. (2024); Basile et al. (2025); Zheng et al. (2024). Uniform aggregation can dilute query-relevant signals and result in coarse or noisy retrieval.

(III) Lack of interaction-aware memory. Prior methods handle each query independently without leveraging historical interactions, ignoring the effective re-exploitation of the established context during streaming video understanding.

To address these limitations, we propose **Event-ReHead**, a training-free framework that enhances KV-cache retrieval with event-level semantics while preserving a high computation efficiency. (1) We introduce a novel event-based retrieval paradigm that reorganizes retrieved key evidence by extracting an **event skeleton** from both **visual** and **semantic** perspectives, thereby providing a coherent and compact structured representation of events. (2) we propose a **query-specific attention head selection strategy** that exploits head heterogeneity to construct more discriminative retrieval representations, enabling precise alignment between the query and relevant visual signals. (3) we incorporate a **compact interaction memory bank** that selectively reuses prior question-answer pairs, providing useful temporal context while avoiding prompt contamination.

Event-ReHead is fully training-free and can be seamlessly integrated with different video-LLM backbones. Experiments on multiple streaming and offline benchmarks demonstrate consistent improvements over strong KV-cache retrieval baselines, including gains of **up to $\uparrow$ 3% on StreamingBench** Lin et al. (2026) and **$\sim$ 4.5% on OVO-Bench**, as well as notable improvements on offline datasets such as EgoSchema Mangalam et al. (2023) and VideoMME Fu et al. (2025).

Our contributions are summarized as follows.

- We introduce an **event-aware KV-cache retrieval paradigm** for streaming video understanding, which enhances frame-level retrieval with semantically coherent event structures.
- We develop a **query-adaptive retrieval mechanism** that combines attention head selection and event-aware context construction to improve the precision and completeness of retrieved evidence.
- We propose a **lightweight interaction memory** that reuses historical question-answer context in a controlled manner, improving efficiency and temporal consistency in streaming scenarios.
- We evaluate Event-ReHead on multiple streaming and offline video understanding benchmarks with different backbone models, demonstrating consistent improvements over strong KV-cache retrieval baselines.

2 Related Work

Video-LLMs and efficient long-video modeling. Recent Video-LLMs, such as Video-LLaMA Cheng et al. (2024), VideoChat Li et al. (2025), LLaVA-NeXT-Video Liu et al. (2024), LongVA Zhang et al. (2024), InternVL Chen et al. (2024c), and Qwen-VL Bai et al. (2025a), extend image-language models to video understanding by encoding sampled frames and aligning visual features with LLMs. They achieve strong results on offline video QA, captioning, and temporal reasoning, but usually assume that the full video is available before inference. To reduce long-video cost, methods such as FastV Chen et al. (2024b), BOLT Liu et al. (2025), and TimeViper Xu et al. (2025) perform frame selection, token pruning, token merging, or efficient temporal modeling. Architectural approaches such as InfiniteVL Tao et al. (2025a) further improve scalability. However, these methods either rely on input-level compression that may lose temporal details, or require model-specific architectural changes and training Ning et al. (2025).

Streaming video understanding and memory. Streaming VideoQA requires continuous frame encoding, compact historical memory, and low-latency query response. Existing methods mainly address this through memory design. ReKV Di et al. retrieves query-relevant video KV caches; LiveVLM Ning et al. (2025) builds a streaming-oriented KV cache; HERMES treats KV cache as hierarchical memory; FreshMem Li et al. (2026) uses frequency-space hybrid memory; and StreamMem Yang et al. (2025c), Flash-VStream Zhang et al. (2025), MovieChat Song et al. (2024), and MA-LMM He et al. (2024) summarize or store historical visual content with different memory structures. These methods show that memory organization is crucial for streaming video understanding, especially under strict latency and memory constraints.

In our work, we organize streaming memory around event-level semantics rather than isolated frames. Additionally, we construct query-adaptive event skeletons to provide structured and coherent retrieval context.

3 Method

We propose **Event-ReHead**, a training-free framework that introduces event-aware retrieval into streaming video understanding. Given a video stream $\mathcal{V} = \{x_t\}_{t=1}^{T}$ and a user query q, the model incrementally encodes incoming frames into a KV cache and retrieves relevant historical evidence when the query arrives. Unlike prior KV-cache retrieval methods that treat the cache as a flat sequence of visual blocks, Event-ReHead organizes retrieval around event semantics. It identifies query-relevant attention heads, constructs question-aware event boundaries, retrieves the most relevant event skeleton, and reuses compact interaction history through a lightweight memory bank.

Let the VLM contain L transformer layers. For layer $\ell \in \{1, \dots, L\}$, the cached keys and values are denoted as $\mathbf{K}_\ell, \mathbf{V}_\ell \in \mathbb{R}^{H_\ell \times N \times d}$, where H_ℓ is the number of attention heads, N is the current cache length, and d is the head dimension. We denote by $\mathcal{I}_v$ the index set of visual tokens in the cache. The goal of retrieval is to select a compact subset $\mathcal{R}_\ell \subseteq \{1, \dots, N\}$ for each layer while preserving the visual evidence needed to answer q.

3.1 Query-Specific Attention Head Retrieval

Attention heads in VLMs are heterogeneous: different heads may focus on objects, actions, spatial relations, or background patterns. Aggregating all heads uniformly can weaken the heads that are most relevant to the current question. Event-ReHead therefore performs query-aware head retrieval before computing visual token scores.

For query q, let $\mathbf{A}_{\ell,h} \in \mathbb{R}^{M \times N}$ be the attention map from M query tokens to N cached tokens at layer ℓ and head h. We estimate the relevance of head h by its strongest visual response to the cached visual tokens:

$$s_{\ell,h}(q) = \frac{1}{M} \sum_{m=1}^{M} \max_{i \in \mathcal{I}_v} \mathbf{A}_{\ell,h}[m, i]. \tag{1}$$

Here, $\mathcal{I}_v$ denotes the indices of visual tokens in the cache, and $s_{\ell,h}(q)$ measures whether head h can focus on salient visual evidence for the current query. We then select the top-K heads:

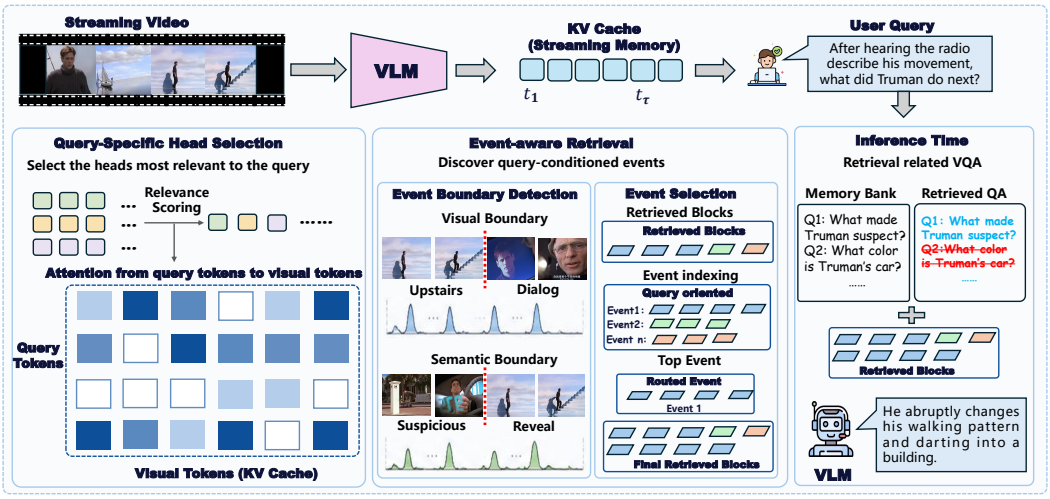

Figure 2: Overview of Event-ReHead. Given a streaming video and a user query, Event-ReHead selects query-relevant attention heads, discovers query-conditioned events, retrieves coherent event-level evidence from the KV cache, and combines it with useful historical QA context for final response generation.

$$\mathcal{H}_\ell(q) = \mathrm{TopK}_h\left(s_{\ell,h}(q)\right). \tag{2}$$

Given the selected heads $\mathcal{H}_\ell(q)$, the relevance score of a visual token i is computed as

$$a_{\ell,i}(q) = \frac{1}{|\mathcal{H}_\ell(q)|M} \sum_{h\in\mathcal{H}_\ell(q)} \sum_{m=1}^{M} \mathbf{A}_{\ell,h}[m,i], \quad i \in \mathcal{I}_v. \tag{3}$$

This produces a more discriminative retrieval score than all-head aggregation, allowing the model to focus on heads that encode question-specific information. We further follow the HERMES retrieval strategy and combine attention relevance with a temporal recency prior:

$$r_{\ell,i} = (1-\alpha_\ell)\,\mathrm{Norm}(a_{\ell,i}) + \alpha_\ell\,\mathrm{Norm}\left(\exp(-\beta_\ell d_i)\right), \tag{4}$$

where d_i is the normalized temporal distance from token i to the most recent visual token, and $\mathrm{Norm}(\cdot)$ denotes min-max normalization over visual tokens. The coefficients α_ℓ and β_ℓ are inherited from HERMES: α_ℓ controls the layer-wise weight of the recency prior, while β_ℓ controls how fast the recency score decays for older tokens. The resulting score $r_{\ell,i}$ is used to select the main retrieved token set $\mathcal{R}_\ell^{\text{base}}$, which serves as the dominant evidence source in Event-ReHead.

3.2 Question-Conditioned Event Discovery

Videos are naturally structured as events rather than isolated frames. Event-ReHead maintains a frame feature sequence $\mathbf{F} = \{\mathbf{f}_t\}_{t=1}^{T}$, where $\mathbf{f}_t \in \mathbb{R}^D$ is the frame-level feature of frame x_t. These features are taken from the visual representations used to build the KV cache, without an extra visual encoding pass. For the current user query q, we obtain a query representation $\mathbf{e}_q$ from the backbone text embedding module. We use $\mathbf{e}_q$ to make event discovery question-conditioned, allowing different queries to produce different event boundaries. The visual transition score between adjacent frames is

$$c_t^v = 1 - \cos(\mathbf{f}_t, \mathbf{f}_{t+1}), \quad t = 1, \ldots, T-1, \tag{5}$$

where $\cos(\cdot,\cdot)$ is cosine similarity. A higher value indicates a stronger visual change. To make event discovery sensitive to the question, we compute a query feature $\mathbf{u}_q \in \mathbb{R}^D$ and measure the relevance between each frame j and the query:

$$\rho_j(q) = \cos(\mathbf{f}_j, \mathbf{u}_q). \tag{6}$$

For a candidate boundary after frame t, where t denotes the boundary position, we compare the query relevance before and after this boundary. The semantic transition score is

$$c_t^s(q) = \left| \frac{1}{|\mathcal{L}_t|} \sum_{j \in \mathcal{L}_t} \rho_j(q) - \frac{1}{|\mathcal{R}_t|} \sum_{j \in \mathcal{R}_t} \rho_j(q) \right|, \tag{7}$$

where $\mathcal{L}_t$ and $\mathcal{R}_t$ are short temporal windows before and after the candidate boundary, and j indexes frames within these windows. This term captures whether the query-relevant content changes across the boundary. The final boundary score is

$$c_t(q) = c_t^v + \lambda c_t^s(q), \tag{8}$$

where λ controls the contribution of semantic information. We select high-scoring boundaries under a minimum event-length constraint and obtain

$$0 = b_0 < b_1 < \cdots < b_J = T. \tag{9}$$

These boundaries define events

$$\mathcal{E}_j = \{x_t \mid b_{j-1} \le t < b_j\}, \quad j = 1, \ldots, J. \tag{10}$$

Thus, event boundaries are not fixed solely by visual changes; they are adapted to the semantic focus of the current query.

3.3 Event-Centric KV Retrieval

After obtaining token relevance scores from query-aware heads, Event-ReHead retrieves visual evidence in an event-centric manner. Let $\mathcal{F}_q$ be the set of frame indices associated with the top-ranked visual tokens. To avoid favoring longer events, we select the most relevant event by measuring the density of retrieved frames within each event:

$$j^* = \arg\max_j \frac{|\mathcal{F}_q \cap [b_{j-1}, b_j)|}{b_j - b_{j-1}}. \tag{11}$$

The normalization by event length reduces the bias toward long events and selects the event with the highest concentration of query-relevant frames. The selected event $\mathcal{E}_{j^*}$ therefore provides coherent temporal context for answering q.

Directly retaining all frames in $\mathcal{E}_{j^*}$ may introduce unnecessary computation, especially for long events. We therefore construct a compact *event skeleton* $\mathcal{S}_{j^*}$ using some representative anchors from the beginning, middle, and ending stages of the event, respectively:

$$\mathcal{S}_{j^*} = \{b_{j^*-1},\ \lfloor (b_{j^*-1} + b_{j^*} - 1)/2 \rfloor,\ b_{j^*} - 1\}. \tag{12}$$

The event skeleton preserves the coarse compact temporal progression of the event, providing coherent event-level context while maintaining efficiency.

Before adding skeleton frames, we check whether they are already covered by the base retrieval. The coverage ratio is

$$\eta(q) = \frac{|\mathcal{S}_{j^*} \cap \mathcal{F}_q|}{|\mathcal{S}_{j^*}|}. \tag{13}$$

If $\eta(q) \ge \tau_c$, where τ_c is the coverage threshold, no event frames are added. Otherwise, we retrieve the KV positions corresponding to the missing skeleton frames and augment the base retrieval set:

$$\mathcal{R}_\ell = \mathcal{R}_\ell^{\text{base}} \cup \mathcal{P}_\ell(\mathcal{S}_{j^*}), \tag{14}$$

This formulation combines fine-grained token retrieval with event-level continuity. The query-aware heads identify precise evidence, while the event skeleton preserves the semantic structure that frame-level retrieval may miss.

3.4 LIGHTWEIGHT INTERACTION MEMORY

Streaming video QA often contains correlated questions Li et al. (2025); Song et al. (2024). A previous answer may identify a person, object, place, or action that is useful for later queries. However, injecting all previous interactions into the prompt would increase context length and may introduce irrelevant information. We therefore maintain a compact memory bank:

$$\mathcal{M} = \{(q_i, a_i)\}_{i=1}^{B}, \tag{15}$$

where q_i and a_i are the i-th historical question and answer, and B is the maximum memory size. When the memory exceeds B, we keep the most recent entries. For a new query q, we compute the similarity between q and each historical question q_i:

$$m_i(q) = \cos(\phi(q), \phi(q_i)), \tag{16}$$

where $\phi(\cdot)$ is a lightweight text feature. In implementation, ϕ can be instantiated as a sparse normalized token vector after stop-word removal, which keeps memory retrieval cheap and model-independent. We retrieve only the top-K_m entries whose similarity exceeds a threshold τ_m:

$$\mathcal{M}_q = \mathrm{TopK}_{(q_i,a_i)\in\mathcal{M}}\big(m_i(q)\big), \quad m_i(q) > \tau_m. \tag{17}$$

The retrieved entries $\mathcal{M}_q$ are prepended to the prompt as short historical context. After answering the current query, the new pair (q, a) is added to the memory bank. This design intentionally keeps memory small. It provides temporal priors from previous interactions while avoiding prompt contamination from unrelated history.

3.5 INFERENCE

Event-ReHead is applied entirely at inference time and requires no model training. For each video stream, the model incrementally updates the KV cache and frame feature sequence. When a query arrives, Event-ReHead first performs query-aware head retrieval, then discovers question-conditioned events, selects the most relevant event skeleton, and retrieves compact interaction memory. The retained KV cache is finally used to generate the answer. By coupling precise head-level retrieval with event-level organization, Event-ReHead converts flat KV-cache memory into a semantically structured representation. This enables more coherent evidence retrieval under streaming memory constraints while remaining compatible with different VLM backbones.

4 EXPERIMENTS

In this section, we perform a wide range of experiments to answer the following questions.

RQ1: How does EVENT-ReHead **perform** on both streaming video understanding and general offline video understanding tasks compared to baselines.

RQ2: How **efficient** is EVENT-ReHead?

RQ3: How does the **modular design** of EVENT-ReHead enhance the model's capabilities?

RQ4: How sensitive is the model to **hyper-parameters**?

4.1 MAIN RESULTS (RQ1)

Table 1 reports the results on streaming and offline video understanding benchmarks. Overall, Event-ReHead consistently improves three backbone models in the training-free offline-to-online setting. Compared with HERMES, Event-ReHead brings average gains of 1.63, 2.11, and 1.50 points on LLaVA-OV-7B Li et al. (2024a), Qwen2.5-VL-7B Bai et al. (2025b), and Qwen2.5-VL-32B Bai et al. (2025b), respectively, showing that the method generalizes across different model architectures.

The gains are clear on streaming benchmarks. On LLaVA-OV-7B, Event-ReHead improves StreamingBench, OVO-Bench-real-time, and OVO-Bench-Backward by 2.07, 1.35, and 1.38 points over HERMES. On Qwen2.5-VL-7B, the improvements are larger on OVO-Bench, with gains of 4.09 and 2.09 points on the real-time and backward splits. On Qwen2.5-VL-32B, Event-ReHead further

improves OVO-Bench-real-time and OVO-Bench-Backward by 1.94 and 1.57 points. These results show that event-level memory provides more coherent evidence for streaming video queries.

Event-ReHead is performed entirely at inference time and does not require any additional model training. For example, it improves LLaVA-OV-7B by 1.84, 1.55, and 1.59 points on MVBench, EgoSchema, and VideoMME, respectively. With Qwen2.5-VL-7B, it improves the same datasets by 1.25, 2.20, and 2.46 points. For Qwen2.5-VL-32B, it achieves clear gains on MVBench and EgoSchema, while remaining comparable on VideoMME. These results indicate that event-aware KV-cache retrieval benefits both streaming interaction and general long-video reasoning.

Table 1: Main results of Event-ReHead in five datasets against baseline models and methods (%).

Model	Frames	StreamingBench	OVO-Bench-real-time	OVO-Bench-Backward	MVBench	EgoSchema	VideoMME
Proprietary MLLMs							
Gemini-1.5-Pro	1 fps	75.69	69.32	62.54	75.69	69.32	66.41
GPT-4o	64	73.28	64.46	60.75	73.28	64.46	62.87
Claude-3.5-Sonnet	20	72.44	-	-	72.44	-	-
Open-source offline MLLMs							
Video-Llama-2-7B	32	49.52	-	-	49.52	-	-
VILA-1.5-8B	14	52.32	-	-	52.32	-	-
Video-CCAM-14B	96	53.96	-	-	53.96	-	-
LongVA-7B	128	59.96	-	-	59.96	-	-
Qwen2-VL-7B	64	69.04	60.65	48.58	67	66.7	63.3
Intern-VL-V2-8B	16	63.72	60.73	44	65.8	-	56.3
Llava-Next-video-32B	64	66.96	-	-	66.96	-	-
MiniCPM-2.6-8B	32	67.44	-	-	67.44	-	-
Open-source Online MLLMs							
Flash-Vstream-7B	-	23.23	29.86	25.35	-	-	-
VideoLLM-online-8B	2 fps	35.99	20.79	17.73	-	-	-
Dispider-7B	1 fps	67.63	54.55	36.06	-	55.6	-
TimeChat-Online-7B	1 fps	75.36	61.9	41.7	75.36	61.9	53.22
StreamForest-7B	1 fps	77.26	61.2	52.02	70.2	-	-
Training-free Offline-to-Online Methods							
Llava-OV-7B	64	71.34	63.06	43.64	57.02	59.93	57.67
with ReKV	0.5 fps	69.22	57.33	44.16	56.83	60.7	57.74
with HERMES	0.5 fps	73.23	66.34	50.2	56.92	60.29	58.44
with LiveVLM	0.5 fps	72.92	64.43	48.77	56.56	60.15	58.73
with streamKV	0.5 fps	68.8	63.21	46.17	57.3	60.07	57.96
with OURS	0.5 fps	**74.7**	**67.02**	**50.43**	**58.79**	**61.78**	**59**
Llava-OV-0.5B	64	59.24	49.63	34.88	44.73	27.4	42.26
with ReKV	0.5 fps	60.13	49.85	35.13	45.75	31	45.32
with HERMES	0.5 fps	62.04	50.34	34.8	**46.45**	31.39	47.13
with OURS	0.5 fps	**63.35**	**51.84**	**43.53**	46.35	**32.72**	**48.48**
Qwen2.5-VL-7B	1 fps	73.31	59.9	44.65	65	58.47	64.52
with ReKV	1 fps	75.53	63.37	45.14	66.17	59.72	**64.73**
with HERMES	1 fps	79.44	68.98	49.43	65.53	59.97	62
with OURS	1 fps	**79.96**	**72.51**	**50.19**	**66.65**	**61.67**	63.31
Qwen2.5-VL-32B	1 fps	74.27	64.4	50.33	67.78	59.32	66.85
with ReKV	1 fps	77.36	68.73	55.4	68.3	63.73	**66.9**
with HERMES	1 fps	80.08	72.37	55.42	68.93	65.85	65.44
with OURS	1 fps	**80.65**	**72.87**	**56.56**	**70.36**	**69.46**	63.11

We further evaluate Event-ReHead on RVS-Ego and RVS-Movie, which emphasize retrospective visual search over long video streams. As shown in Table 2, Event-ReHead remains effective on both 7B and 0.5B backbones. With LLaVA-OV-7B, Event-ReHead achieves 61.8% accuracy and 4.2 score on RVS-Ego, improving HERMES by 1.5 points in accuracy and 0.2 in score. On RVS-Movie, it reaches 55.2% accuracy and 3.7 score, outperforming HERMES by 0.8 points and 0.1 score. Compared with other streaming memory methods, Event-ReHead also obtains the best RVS-Movie accuracy and the highest scores on both datasets, indicating stronger ability to retrieve temporally relevant evidence from long-range context.

The improvement is also observed with the lightweight LLaVA-OV-0.5B backbone. Event-ReHead improves HERMES from 53.0 to 54.9 on RVS-Ego and from 42.5 to 45.1 on RVS-Movie, while matching or improving the corresponding score metrics. These results suggest that event-aware memory organization is particularly useful for retrospective search tasks: instead of relying only on scattered frame-level evidence, Event-ReHead retrieves temporally coherent event context, which provides more reliable support for locating and reasoning over past visual observations.

Table 2: Performance of our methods against baselines on RVS-Ego and RVS-Movie (%).

Model	RVS-Ego		RVS-Movie	
	Acc	Score	Acc	Score
LLaVA-OV-7B	56.2	3.7	43.0	3.3
+ReKV	63.7	4.0	54.4	3.6
+Flash-Vstream	57.0	4.0	53.1	3.3
+InfiniPot-V	57.9	3.5	51.4	3.5
+StreamMem	57.6	3.8	52.7	3.4
+StreamingTOM	58.3	3.9	53.2	3.5
+HERMES	60.3	4.0	54.4	3.6
+Event-ReHead	**61.8**	**4.2**	**55.2**	**3.7**
LLaVA-OV-0.5B	51.8	3.7	37.2	3.2
+ReKV	54.7	3.9	44.6	3.4
+HERMES	53.0	3.8	42.5	3.4
+Event-ReHead	**54.9**	**3.9**	**45.1**	**3.4**

Table 3: Efficiency across input frame numbers under two chunk sizes. "TTFT"denotes Time to First Token and "TPOT" denotes Time Per Output Token.

Metric	Frames			
	16	64	256	512
Chunk Size: 8				
GPU Mem. (GB)↓	16.80	16.83	16.83	16.83
TTFT (ms) ↓	27.52	28.71	28.71	28.71
TTOP (ms)↓	24.56	24.31	24.45	24.51
Chunk Size: 16				
GPU Mem.↓	17.83	17.92	17.92	17.92
TTFT↓	27.12	29.07	28.63	28.47
TTOP↓	24.73	24.02	23.90	23.93

4.2 Efficiency Analysis (RQ2)

We evaluate efficiency using GPU memory and TTFT, where TTFT denotes Time to First Token. As shown in Figure 3, EVENT-ReHead achieves lower memory usage than ReKV across all frame settings. It reduces memory by 1.54 GB, 3.12 GB, and 3.18 GB at 16, 64, and 256 frames, respectively. Its memory cost also remains nearly stable as the number of frames increases, indicating better scalability for long video streams.

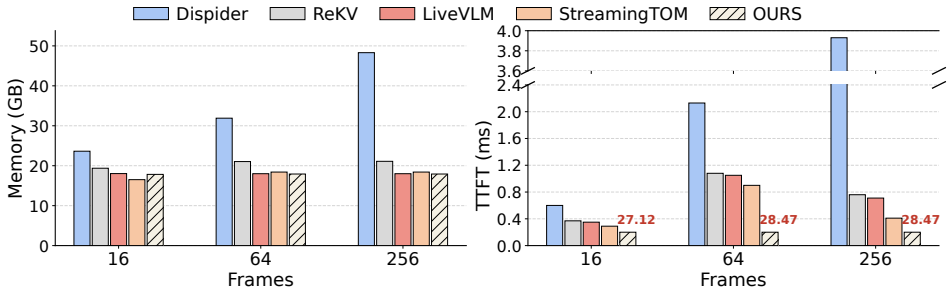

Figure 3: Efficiency analysis of Event-ReHead. Event-ReHead achieve the best performance in terms of efficiency.

EVENT-ReHead also reduces TTFT compared with ReKV by 0.17 ms, 0.88 ms, and 0.56 ms at 16, 64, and 256 frames, respectively. This shows that event-aware retrieval and query-adaptive head selection improve retrieval quality without adding noticeable response overhead. The advantage comes from the HERMES-style compact KV-cache design. Unlike ReKV, which performs query-time retrieval over visual blocks, EVENT-ReHead maintains a compact reusable cache and organizes it with event-level memory. This design improves semantic retrieval while keeping both memory usage and response latency low.

4.3 Ablation Study (RQ3)

Figure 4 shows that each component contributes to Event-ReHead. Removing the event module causes the largest drop in most cases. For example, performance decreases by 1.73 points on StreamingBench with LLaVA-OV-7B and by 2.20 points with Qwen2.5-VL-7B. This confirms that event-level retrieval is important for preserving coherent temporal semantics beyond frame-level evidence.

The memory bank also brings consistent gains. Without it, the score drops by 1.20 points on StreamingBench for LLaVA-OV-7B and 1.18 points for Qwen2.5-VL-7B, showing that historical QA context provides useful temporal priors for streaming interaction. Removing head selection further

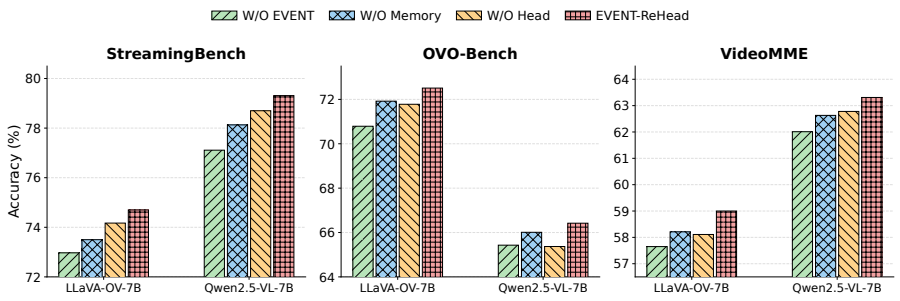

Figure 4: Ablation studies of Event-ReHead. The different modules of Event-ReHead exhibited different performance characteristics.

reduces performance, especially on OVO-Bench and VideoMME, indicating that query-relevant attention heads help build more discriminative retrieval features.

Overall, the full Event-ReHead achieves the best results across all datasets and backbones, demonstrating that event retrieval, memory reuse, and query-adaptive head selection are complementary.

4.4 Parameter analysis (RQ4)

Figure 5 studies three key hyperparameters: memory budget, number of selected heads, and event ratio. The results show that Event-ReHead is generally stable across different settings, while each parameter controls a different trade-off.

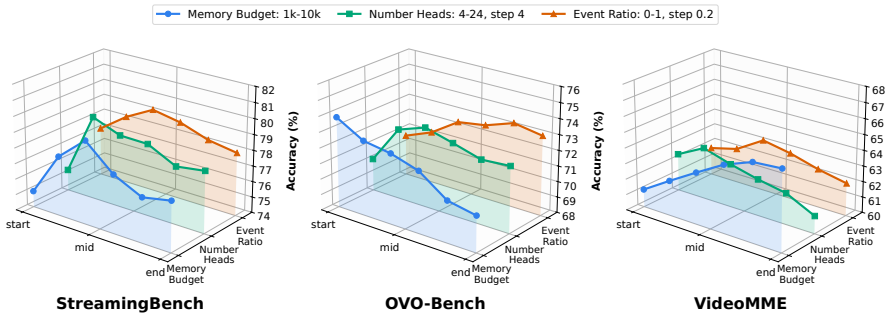

Figure 5: Performance Comparison of Qwen2.5-VL-7B across Different hyperparameter settings.

For memory budget, performance improves with a moderate cache size but saturates when the budget becomes too large, suggesting that excessive cache may introduce redundant context. For selected heads, using too few heads loses useful query-related signals, while too many heads can introduce irrelevant attention patterns. For event ratio, balanced fusion performs better than relying only on frame-level retrieval or event-level retrieval, showing that event semantics are most effective as complementary evidence. These trends support our design: compact memory storage, query-adaptive head selection, and event-aware retrieval each contribute to robust performance.

5 Conclusion

We propose Event-ReHead, a training-free framework for efficient streaming video understanding. Instead of treating video memory as isolated frames or fixed visual blocks, Event-ReHead introduces event-aware memory organization to preserve temporally coherent semantics. It further improves retrieval with query-adaptive attention head selection and a lightweight interaction memory bank for reusing relevant historical QA context. Experiments on streaming and offline video benchmarks show that Event-ReHead consistently improves different VLM backbones while maintaining low memory usage and fast response latency. These results demonstrate that event-level semantics provide an effective and practical direction for building more reliable streaming video understanding systems.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, proofreading, and presentation refinement. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the accuracy, originality, and scientific integrity of the manuscript. The pretrained vision-language models evaluated in this work are components of the experimental framework and are described separately in the methodology and experimental settings.

ETHICS STATEMENT

Event-ReHead is a training-free framework for streaming video understanding that combines event-aware retrieval, query-specific attention head selection, and interaction memory. Potential applications include assistive video analysis and interactive visual agents. However, processing video streams and retaining historical question–answer pairs may expose sensitive information about individuals, activities, and locations. The framework may also inherit biases and hallucinations from its pretrained backbones. Selective retrieval can omit relevant context, while reusing previous answers may propagate earlier errors across interactions. Improved retrieval performance therefore does not guarantee factual accuracy, and model outputs should not be treated as verified evidence or used as the sole basis for high-impact decisions. Responsible deployment should respect privacy, consent, and the usage terms of source videos and pretrained models, with appropriate access controls, memory-retention limits, and human oversight.

REPRODUCIBILITY STATEMENT

The Method section describes query-specific attention head selection, question-conditioned event discovery, event-centric KV-cache retrieval, interaction memory, and the inference procedure. Event-ReHead operates entirely at inference time without updating the pretrained model parameters. Appendix A describes the evaluation datasets and baseline methods. Appendix B reports the backbone models, hardware configuration, frame-sampling rates, KV-cache budget, principal retrieval settings, and evaluation metrics. Detailed task-level results are provided in Appendix C, with additional hyperparameter analyses in Appendix D. Component ablations and efficiency measurements are presented in Figure 4 and Table 3, respectively. These descriptions document the methodological choices and experimental settings used to obtain the reported results.

REFERENCES

Anthropic. Claude 3.5 sonnet. 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

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 arXiv:2511.21631*, 2025a.

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. *ArXiv*, abs/2502.13923, 2025b. URL https://api.semanticscholar.org/CorpusID:276449796.

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. Head pursuit: Probing attention specialization in multimodal transformers. *arXiv preprint arXiv:2510.21518*, 2025.

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*, pp. 18407–18418, 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, 2024b.

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. Streamingtom: Streaming token compression for efficient video understanding. *arXiv preprint arXiv:2510.18269*, 2025.

Yilong Chen, Xiang Bai, Zhibin Wang, Chengyu Bai, Yuhan Dai, and Ming Lu. Streamkv: Streaming video question-answering with segment-based kv cache retrieval and compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3120–3128, 2026.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024c.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

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 arXiv:2507.06261*, 2025.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval (2025). *URL https://arxiv. org/abs/2503.00540*.

Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024.

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 the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13504–13514, 2024.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *European conference on computer vision*, pp. 709–727. Springer, 2020.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. Infinipot-v: Memory-constrained kv cache compression for streaming video understanding. *arXiv preprint arXiv:2506.15745*, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Kangcong Li, Peng Ye, Lin Zhang, Chao Wang, Huafeng Qin, and Tao Chen. Freshmem: Brain-inspired frequency-space hybrid memory for streaming video understanding. *arXiv preprint arXiv:2602.01683*, 2026.

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

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025.

Yifei Li, Junbo Niu, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding?, 2025. *URL https://arxiv. org/abs/2501.05510*.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26689–26699, 2024.

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 *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12147–12151. IEEE, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3318–3327, 2025.

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

Zhenyu Ning, Guangda Liu, Qihao Jin, Wenchao Ding, Minyi Guo, and Jieru Zhao. Livevlm: Efficient online video understanding via streaming-oriented kv cache and retrieval. *arXiv preprint arXiv:2505.15269*, 2025.

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 Computer Vision and Pattern Recognition Conference*, pp. 24045–24055, 2025.

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

Hongyuan Tao, Bencheng Liao, Shaoyu Chen, Haoran Yin, Qian Zhang, Wenyu Liu, and Xinggang Wang. Infinitevl: Synergizing linear and sparse attention for highly-efficient, unlimited-input vision-language models. *arXiv preprint arXiv:2512.08829*, 2025a.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18992–19001, 2025b.

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. *arXiv preprint arXiv:2505.05467*, 2025a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3272–3283, 2025b.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. Duoattention: Efficient long-context llm inference with retrieval and streaming heads. *arXiv preprint arXiv:2410.10819*, 2024.

Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. *arXiv preprint arXiv:2501.13468*, 2025.

Boshen Xu, Zihan Xiao, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Jian Luan, and Qin Jin. Timeviper: A hybrid mamba-transformer vision-language model for efficient long video understanding. *arXiv preprint arXiv:2511.16595*, 2025.

Haolin Yang, Feilong Tang, Lingxiao Zhao, Xiang An, Ming Hu, Huifa Li, Xinlin Zhuang, Yifan Lu, Xiaofeng Zhang, Abdalla Swikir, et al. Streamagent: Towards anticipatory agents for streaming video understanding. *arXiv preprint arXiv:2508.01875*, 2025a.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 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. *arXiv preprint arXiv:2508.15717*, 2025c.

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.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, et al. Streamforest: Efficient online video understanding with persistent event memory. *arXiv preprint arXiv:2509.24871*, 2025.

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*, pp. 21059–21069, 2025.

Haowei Zhang, Shudong Yang, Jinlan Fu, See-Kiong Ng, and Xipeng Qiu. Hermes: Kv cache as hierarchical memory for efficient streaming video understanding. *arXiv preprint arXiv:2601.14724*, 2026.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*, 2024.

A Datasets and Baselines

Datasets. We evaluate Event-ReHead on both streaming and offline video understanding benchmarks.

StreamingBench Lin et al. (2026): StreamingBench is a streaming video understanding benchmark where videos are observed progressively and questions are asked at different timestamps. It evaluates whether a model can answer queries using only the visual content available up to the current time.

OVO-Bench Li et al.: OVO-Bench evaluates online video understanding under real-time and backward reasoning settings. The real-time split focuses on questions about the current observed context, while the backward split requires models to retrieve and reason over previous video content.

RVS-Ego Zhang et al. (2025): RVS-Ego is a retrospective visual search benchmark built on egocentric videos. It requires the model to locate and reason about past observations in long first-person video streams.

RVS-Movie Zhang et al. (2025): RVS-Movie evaluates retrospective visual search in movie videos. Compared with egocentric videos, movie videos contain more diverse scenes, characters, and narrative changes, making long-range memory retrieval important.

MVBench Li et al. (2024b): MVBench is a comprehensive offline video understanding benchmark covering diverse temporal, action, and reasoning tasks. It is used to evaluate general video reasoning ability across multiple task types.

EgoSchema Mangalam et al. (2023): EgoSchema focuses on long-form egocentric video understanding. It requires models to answer questions that often depend on extended temporal context and human-object interactions.

VideoMME Fu et al. (2025): VideoMME is a multi-domain video understanding benchmark covering videos of different durations and categories. It evaluates both short-term perception and long-context reasoning in offline video QA.

Baselines. We compare Event-ReHead with the same categories of baselines used in HERMES. These include proprietary MLLMs such as Gemini-1.5-Pro Comanici et al. (2025), GPT-4o Hurst et al. (2024), and Claude-3.5-Sonnet Anthropic (2024); open-source offline Video-LLMs such as Video-LLaMA-2 Cheng et al. (2024), VILA-1.5 Lin et al. (2024), Video-CCAM Fei et al. (2024), LongVA Zhang et al. (2024), Qwen2-VL Wang et al. (2024), InternVL Chen et al. (2024c), LLaVA-NeXT-Video Liu et al. (2024), and MiniCPM Hu et al. (2024); and online or training-free offline-to-online methods such as Flash-VStream Zhang et al. (2025), VideoLLM-Online Chen et al. (2024a), Dispider Qian et al. (2025), TimeChat-Online Yao et al. (2025), StreamForest Zeng et al. (2025), ReKV Di et al., LiveVLM Ning et al. (2025), StreamKV Chen et al. (2026), and HERMES Zhang et al. (2026). For RVS-Ego and RVS-Movie, we additionally compare with ReKV, Flash-VStream, InfiniPot-V Kim et al. (2025), StreamMem Yang et al. (2025c), StreamingTOM Chen et al. (2025), and HERMES following their reported settings.

Among these methods, HERMES is the most direct baseline because it also follows a training-free KV-cache memory paradigm. We apply Event-ReHead to multiple backbones, including LLaVA-OV-7B, LLaVA-OV-0.5B, Qwen2.5-VL-7B, and Qwen2.5-VL-32B, and compare with the corresponding HERMES results under the same evaluation settings.

B EXPERIMENTAL DETAILS

All experiments are conducted in a training-free setting and using 8*NVIDIA H20 GPUs (96G). We directly apply Event-ReHead to pretrained VLM backbones without updating model parameters. For LLaVA-OV models, we use the official LLaVA-OneVision checkpoints. For Qwen-based models, we use Qwen2.5-VL and Qwen3-VL instruction-tuned checkpoints. Following the streaming evaluation protocol, videos are processed sequentially, and the model answers each query using only the visual content observed before the query timestamp.

For the main experiments, we sample videos at 0.5 FPS for LLaVA-OV and 1.0 FPS for Qwen-based models. Unless otherwise specified, the KV-cache memory budget is set to 4000 tokens. For query-adaptive head retrieval, we select 8 attention heads. For event-aware retrieval, we set the event ratio to 0.4 to balance the original token-level retrieval signal and the event-level evidence. These values are chosen according to the parameter analysis in Figure 5 and are kept fixed across datasets for fair comparison.

For event discovery, the semantic transition weight λ controls the contribution of query-conditioned semantic changes when detecting event boundaries. We use a fixed value for all experiments. In the HERMES-style base retrieval score, the layer-wise coefficients α_ℓ and β_ℓ control the recency prior

and its temporal decay, respectively. These coefficients are inherited from the HERMES retrieval strategy and are not tuned as additional hyperparameters in our method.

For evaluation, we use the official metrics of each benchmark. Multiple-choice datasets are evaluated by answer accuracy. For RVS-Ego and RVS-Movie, we report both accuracy and the official GPT-based score. We use the full evaluation split for all reported main results. For parallel inference, the dataset is evenly divided into chunks across available GPUs; Qwen2.5-VL-32B uses two GPUs per chunk, while smaller backbones are evaluated with one GPU per chunk.

C Full Performances

This section provides detailed results on StreamingBench and OVO-Bench. These fine-grained evaluations further show where Event-ReHead improves over existing training-free offline-to-online methods.

C.1 StreamingBench

Table 4 reports the category-level results on StreamingBench. Event-ReHead achieves the best average performance across all evaluated backbones. On LLaVA-OV-7B, our method improves the average score to 74.70, outperforming HERMES by 1.47 points. The gains are mainly from categories that require temporal evidence, such as Causal Reasoning, Event Understanding, Prospective Reasoning, Spatial Understanding, and Counting. This supports our motivation that event-aware retrieval can provide more coherent evidence than flat token or frame retrieval.

The improvement remains consistent on smaller and stronger backbones. With LLaVA-OV-0.5B, Event-ReHead improves the average score from 62.04 to 63.35 over HERMES. With Qwen2.5-VL-7B and Qwen2.5-VL-32B, our method further improves the average scores to 79.96 and 80.65, respectively. Although HERMES performs better on a few perception-heavy categories, Event-ReHead obtains stronger overall results by improving categories that benefit from event-level temporal organization.

C.2 OVOBench

Table 5 shows the detailed results on OVO-Bench, which includes both real-time visual perception and backward tracing tasks. Event-ReHead consistently improves the overall average across all backbones. On LLaVA-OV-7B, our method improves the real-time average from 66.34 to 67.02 and the backward average from 50.20 to 50.43, leading to an overall gain from 58.27 to 58.72. On LLaVA-OV-0.5B, the overall score also increases from 42.76 to 43.53.

The gains are more evident on Qwen backbones. With Qwen2.5-VL-7B, Event-ReHead improves the real-time average by 3.53 points and the backward average by 0.76 points over HERMES, increasing the overall score from 59.21 to 61.92. With Qwen2.5-VL-32B, our method improves the overall score from 63.90 to 65.60. These results indicate that Event-ReHead is especially effective for real-time perception and also benefits backward tracing, where temporally coherent event evidence is critical for locating relevant past information.

Overall, the fine-grained results demonstrate that Event-ReHead does not only improve aggregate performance, but also brings consistent benefits across diverse temporal reasoning and event-related categories. This verifies that introducing event-level memory organization and query-aware retrieval is effective for streaming video understanding.

D Analysis on More Hyper-parameters

We further analyze additional hyper-parameters related to event-aware retrieval on Qwen2.5-VL-7B. Figure 6 reports the results on StreamingBench, OVO-Bench, and VideoMME.

The left part studies the joint effect of event semantic weight and memory top-k. The best region appears around a moderate semantic weight and a small memory top-k, indicating that event semantics and historical interactions are both useful but should be controlled. Too weak semantic weighting

Table 4: Accuracy comparison (%) on StreamingBench focusing on Real-Time Visual Understanding tasks. Real-Time Visual Understanding tasks consists of Object Perception(OP), Causal Reasoning(CR),Clips Summarization(CS), Attribute Perception(ATP), Event Understanding(EU), Text-Rich Understanding(TR), Prospective Reasoning(PR), Spatial Understanding(SU), Action Perception(ACP), and Counting(CT).

Model	#Frames	OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	Avg.
Human	-	89.47	92	93.6	91.47	95.65	92.52	88	88.75	89.74	91.3	91.46
Proprietary MLLMs												
Gemini 1.5 pro	1 fps	79.02	80.47	83.54	79.67	80	84.74	77.78	64.23	71.95	48.7	75.69
GPT-4o	64	77.11	80.47	83.91	76.47	70.19	83.8	66.67	62.19	69.12	49.22	73.28
Claude-3.5-Sonnet	20	73.33	80.47	84.09	82.02	75.39	79.53	61.11	61.79	69.32	43.09	72.44
Open-source Offline-MLLMs												
Video-LLaMA2-7B	32	55.86	55.47	57.41	58.17	52.8	43.61	39.81	42.68	45.61	35.23	49.52
VILA-1.5-8B	14	53.68	49.22	70.98	56.86	53.42	53.89	54.63	48.78	50.14	17.62	52.32
Video-CCAM-14B	96	56.4	57.81	65.3	62.75	64.6	51.4	42.59	47.97	49.58	31.61	53.96
LongVL-7B	128	70.03	63.28	61.2	70.92	62.73	59.5	61.11	53.66	54.67	34.72	59.96
InternVL-V2-8B	16	68.12	60.94	69.4	77.12	67.7	62.93	59.26	53.25	54.96	56.48	63.72
Kangaroo-7B	64	71.12	84.38	70.66	73.2	67.08	61.68	56.48	55.69	62.04	38.86	64.6
LLaVA-NeXT-Video-32B	64	78.2	70.31	73.82	76.8	63.35	69.78	57.41	56.1	64.31	38.86	66.96
MiniCPM-V-2.6-8B	32	71.93	71.09	77.92	75.82	64.6	65.73	70.37	56.1	62.32	53.37	67.44
Open-source Online MLLMs												
Flash-VStream-7B	-	25.89	43.57	24.91	23.87	27.33	13.08	18.52	25.2	23.87	48.7	23.23
VideoLLM-online-8B	2 fps	39.07	40.06	34.49	31.05	45.96	32.4	31.48	34.16	42.49	27.89	35.99
Dispider-7B	1 fps	74.92	75.53	74.1	73.08	74.44	59.92	76.14	62.91	62.16	45.8	67.63
TimeChat-Online-7B	1 fps	80.22	82.03	79.5	83.33	76.1	78.5	78.7	64.63	69.6	57.98	75.36
StreamForest-7B	1 fps	83.11	82.81	82.65	84.26	77.5	78.19	76.85	69.11	75.64	54.4	77.26
Training-free Offline-to-Online Methods												
LLaVA-OV-7B	32	78.75	78.12	80.76	81.19	71.7	72.59	72.22	63.82	66.01	38.34	71.34
+ ReKV	0.5 fps	76.02	81.25	77.92	76.9	66.04	66.04	69.44	60.98	64.31	49.22	69.22
+ LiveVLM	0.5 fps	81.47	78.13	83.28	79.08	69.57	74.14	75	69.11	67.71	40.41	72.92
+ StreamKV	0.5 fps	73.8	77.3	85.9	77.5	73.3	63.9	69.4	61.4	63.2	35.8	68.8
+ HERMES	0.5 fps	**79.02**	81.25	**87.7**	80.2	69.18	**71.96**	73.15	66.26	**69.41**	43.52	73.23
+ OURS	0.5 fps	78.83	**81.97**	87.46	**80.52**	**71.73**	71.37	**74.13**	**67.01**	68.73	**44.69**	**74.7**
LLaVA-OV-0.5B	32	71.39	57.81	65.93	69.64	69.18	55.76	57.41	52.85	62.04	16.58	59.64
+ ReKV	0.5 fps	65.12	60.16	66.56	66.01	66.67	52.96	57.41	48.37	60.34	18.13	57.39
+ HERMES	0.5 fps	72.21	61.72	**70.98**	**72.94**	72.33	57.94	60.19	52.85	63.74	**19.17**	62.04
+ OURS	0.5 fps	**72.3**	**65.62**	69.72	72.63	**76.67**	**58.39**	**61.11**	**53.49**	**65.72**	18.13	**63.35**
Qwen2.5-VL-7B	1 fps	77.93	76.56	78.55	80.86	76.73	76.95	80.56	65.45	65.72	52.85	73.31
+ HERMES	0.5 fps	83.65	81.25	88.01	**87.46**	76.73	86.6	82.41	76.02	**73.94**	46.63	79.44
+ OURS	0.5 fps	**84.29**	**80.8**	**88.64**	85.15	**79.87**	**87.85**	**82.41**	**76.61**	73.37	**47.04**	**79.96**
Qwen2.5-VL-32B	1 fps	76.29	79.69	78.55	83.5	76.1	79.44	80.56	61.38	68.27	59.07	74.27
+ HERMES	0.5 fps	**83.92**	80.47	**87.7**	83.5	80.5	**88.16**	87.04	75.2	77.34	48.19	80.08
+ OURS	0.5 fps	82.02	**80.56**	87.07	**84.85**	**85.47**	86.92	**87.07**	**75.2**	**75.35**	**49.74**	**80.65**

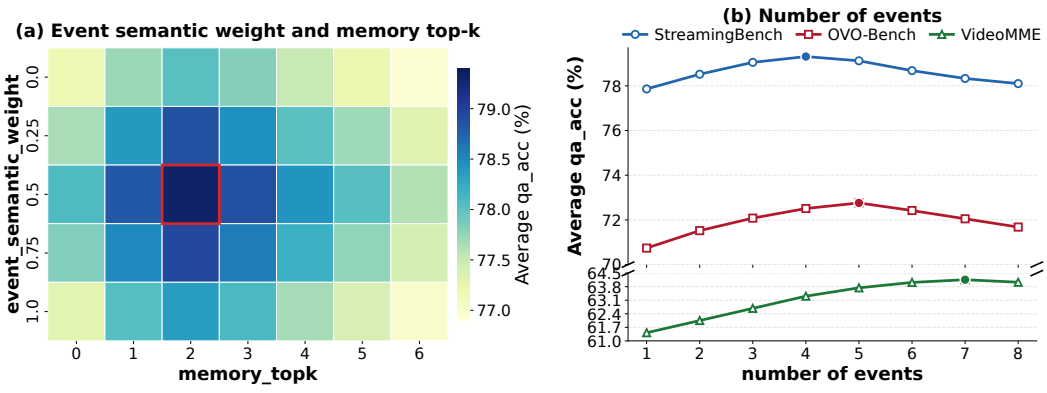

Figure 6: Analysis of additional hyper-parameters. Left: joint effect of event semantic weight and memory top-k on StreamingBench. Right: effect of the number of events on StreamingBench, OVO-Bench, and VideoMME.

cannot fully exploit event cues, while overly large memory top-k may introduce irrelevant historical context. The default setting, marked by the red box, provides a good balance between event grounding and memory reuse.

Table 5: Accuracy comparison (%) on OVO-Bench focusing on Real-Time Visual Perception and Backward Tracing tasks. Real-Time Visual Perception tasks consist of Optical Character Recognition (OCR), Action Recognition (ACR), Attribute Recognition (ATR), Spatial Understanding (STU), Future Prediction (FPD), Object Recognition (OJR). Backward Tracing tasks consists of Episodic Memory (EPM), Action Sequence Identification (ASI), Hallucination Detection (HLD).

Model	#Frames	Real-Time							Backward				Overall
		OCR	ACR	ATR	STU	FPD	OJR	Avg.	EPM	ASI	HLD	Avg.	Avg.
Human	-	93.96	92.57	94.83	92.7	91.09	94.02	93.2	92.59	93.02	91.37	92.33	92.77
Proprietary MLLMs													
Gemini 1.5 Pro	1fps	85.91	66.97	79.31	58.43	63.37	61.96	69.32	58.59	76.35	52.64	62.54	65.93
GPT-4o	64	69.8	64.22	71.55	51.12	70.3	59.78	64.46	57.91	75.68	48.66	60.75	62.61
Open-source Offline MLLMs													
LLaVA-Video-7B	64	69.8	59.63	66.38	50.56	72.28	61.41	63.34	51.18	64.19	9.68	41.68	52.51
Qwen2-VL-7B	64	69.13	53.21	63.79	50.56	66.34	60.87	60.65	44.44	66.89	34.41	48.58	54.62
InternVL2-8B	64	68.46	58.72	68.97	44.94	67.33	55.98	60.73	43.1	61.49	27.41	44	52.37
LongVU-7B	1fps	55.7	49.54	59.48	48.31	68.32	63.04	57.4	43.1	66.22	9.14	39.49	48.45
Open-source Online MLLMs													
VideoLLM-online-8B	2fps	8.05	23.85	12.07	14.04	45.54	21.2	20.79	22.22	18.8	12.18	17.73	19.26
Flash-VStream-7B	1fps	25.5	32.11	29.31	33.71	29.7	28.8	29.86	36.36	33.78	5.91	25.35	27.61
Dispider-7B	1fps	57.72	49.54	62.07	44.94	61.39	51.63	54.55	48.48	55.41	4.3	36.06	45.31
TimeChat-Online-7B	1fps	75.2	46.8	70.7	47.8	69.3	61.4	61.9	55.9	59.5	9.7	41.7	51.8
StreamForest-7B	1fps	68.46	53.21	71.55	47.75	65.35	60.87	61.2	58.92	64.86	32.26	52.02	56.61
Training-free Offline-to-Online Methods													
LLaVA-OV-7B	32	67.79	55.05	72.41	48.31	72.28	62.5	63.06	57.24	55.41	18.28	43.64	53.35
+ HERMES	0.5 fps	72.48	62.39	74.14	50.56	73.27	**65.22**	66.34	60.61	61.49	**28.49**	50.2	58.27
OURS	0.5 fps	**73.13**	**63.46**	**75.69**	**51.43**	**74.3**	64.13	**67.02**	**61.25**	**62.76**	27.27	**50.43**	**58.72**
LLaVA-OV-0.5B	32	53.69	53.21	48.28	33.71	60.4	48.91	49.7	46.13	45.27	12.37	34.59	42.15
+ HERMES	0.5 fps	56.38	47.71	56.9	**32.02**	62.38	**48.91**	50.72	47.81	47.97	**8.6**	34.8	42.76
OURS	0.5 fps	**58.02**	**51.79**	**60.17**	30.34	**62.38**	48.37	**51.84**	**48.14**	**49.59**	7.99	**35.24**	43.53
Qwen2.5-VL-7B	1fps	67.79	55.05	67.24	42.13	66.34	60.87	59.9	51.52	58.78	23.66	44.65	52.28
+ HERMES	1fps	85.23	64.22	71.55	53.37	**74.26**	65.22	68.98	**48.48**	62.16	37.63	49.43	59.21
OURS	1fps	**87.92**	**68.81**	**80.17**	**56.74**	71.29	**70.11**	**72.51**	46.46	**62.16**	**41.94**	**50.19**	**61.92**
Qwen2.5-VL-32B	1fps	77.18	58.72	68.1	50.56	74.26	57.61	64.4	58.59	62.84	29.57	50.33	57.37
+ HERMES	0.5 fps	88.59	65.14	74.14	**58.99**	**71.29**	76.09	72.37	52.19	**66.22**	47.85	55.42	63.9
OURS	0.5 fps	**88.59**	**67.89**	**75.86**	57.3	69.31	**78.26**	**72.87**	**55.89**	64.86	**48.92**	**56.56**	**65.6**

The right part analyzes the number of events. The optimal event granularity varies across datasets. StreamingBench reaches its best performance with fewer events, while OVO-Bench benefits from a slightly larger number of events. VideoMME, which contains longer videos, continues to improve when more events are used. This trend suggests that the proper event number should depend on video length: short videos require compact event structure, whereas long videos need finer event decomposition to cover long-range content.

Overall, these results show that Event-ReHead benefits from controlled event-aware memory organization.

E Case Studies

We provide three qualitative examples from StreamingBench, OVO-Bench, and VideoMME in Figure 7. Each case compares EVENT-ReHead with the original LLaVA-OV-0.5B baseline on the same video and question.

In the StreamingBench example, the question asks which car is leading at the current moment. The original model predicts the black car, while EVENT-ReHead correctly identifies the white car with a red roof. This shows that our method can retrieve the temporally relevant racing event rather than relying on visually salient but outdated objects.

In the OVO-Bench example, the model needs to answer where the book is. The original model predicts that the book is on a chair, whereas EVENT-ReHead correctly answers that it is on the table. This case requires locating the relevant past interaction, demonstrating the benefit of event-aware memory organization for retrospective reasoning.

In the VideoMME example, the question asks why Sadie has a countdown. The original model fails to produce the correct option, while EVENT-ReHead selects the answer related to the food preparer's judgment about Sadie's food choice. This indicates that event-level retrieval helps preserve the causal context needed for long-video reasoning.



Figure 7: Case Studies of our work in three benchmarks.

Overall, these cases show that EVENT-ReHead improves answer accuracy by retrieving more coherent event-level evidence. Compared with the original baseline, it better handles current-state recognition, retrospective localization, and causal reasoning across different video benchmarks.