EVENTMEM: ADAPTIVE EVENT-CENTRIC MEMORY FOR STREAMING VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

MLLMs often struggle with real-time reasoning over streaming videos as visual information continuously accumulates over time. A key bottleneck is efficient memory management: preserving long-term context within a compact representation while retaining transient and fine-grained visual information. To address these challenges, we introduce EventMem, an event-centric memory framework without additional training: (1) EventMem discovers structured event units from incoming video streams by temporal-aware adaptive K-means event segmentation **(TAKC)**. (2) it performs temporal segmentation and aligned frame compression **(TSAFM)** to each event, reducing redundant visual information while retaining temporally informative frames. (3) it builds **complementary event-residual memory** to preserve both long-term context and recent fine-grained evidence, and retrieves relevant historical events and fuses them with short-term evidence for efficient streaming reasoning. Extensive experiments show that EventMem consistently improves streaming video understanding. It improves the accuracy on object search, long-term and short-term questions by **11.7, 6.7, and 3.1** points, respectively, validating the effectiveness of event-aware compression and complementary memory for streaming reasoning.

1 INTRODUCTION

Multimodal large language models (MLLMs) have achieved remarkable progress in video question answering, event understanding, and temporal reasoning (Bai et al. (2025a;b)). However, the majority of video MLLMs require availability of the full video locally and perform inference over a fixed set of sampled frames or visual tokens. Real-world applications, such as embodied agents, intelligent assistants, live-stream analysis, and first-person interaction, instead require models to continuously process incoming visual streams and answer queries at arbitrary times (Di et al. (2025); Qian et al. (2025); Wang et al. (2026)). As visual information continuously accumulates, directly retaining historical observations leads to rapidly growing memory and computation, making efficient memory management a central bottleneck for streaming video understanding.

The key difficulty is not merely storing more visual information, but preserving the right information under a constrained memory budget. A streaming model must retain long-term context for reasoning over distant events, while simultaneously preserving transient actions, state transitions, and fine-grained recent details. These two requirements are inherently difficult to satisfy with a single representation. Aggressive compression improves efficiency but may discard short-lived yet critical evidence, whereas retaining dense visual observations preserves details at the cost of increasing memory and inference overhead. Therefore, effective streaming memory requires jointly modeling temporal event structure, visual redundancy, and the complementary roles of historical context and recent evidence. Most of existing approaches reduce temporal redundancy directly from streaming visual inputs to control token accumulation Yao et al. (2025); Xie

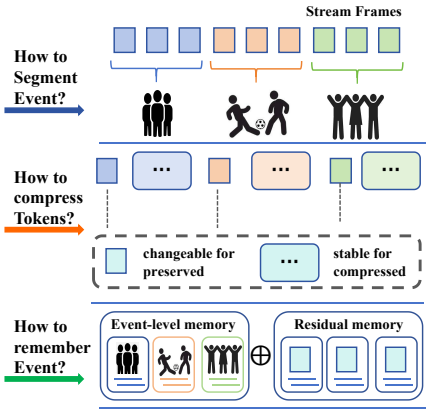


Figure 1: The key challenges in video understanding under the stream inputs

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.

et al. (2026); Chen et al. (2026). However, compression is often performed independently of event structure, making it difficult to distinguish redundant stable observations from temporally informative changes. More recent methods have begun to organize streaming history at the event level Zeng et al. (2026); Guo et al. (2026); Liang et al. (2026). Nevertheless, these strategies typically apply single spatial or temporal compression to frames within an event, overlooking inter-frame semantic changes and thus failing to model how an event evolves over time. Meanwhile, the complementary between short- and long-term memory has been further explored to provide more information for MLLMs reasoning Xiong et al. (2025); Li et al. (2026). More generally, streaming memory involves a compromise between compact long-term information and detailed recent evidence, as stronger historical compression may reduce local details while short-term memory necessarily covers a more limited temporal range.

These observations lead to three tightly coupled questions for streaming video memory, as shown in Figure 1: how to discover meaningful event units from an evolving video stream, how to remove redundant visual information without suppressing temporal changes, and how to preserve both long-term event context and recent fine-grained evidence? Addressing them independently is insufficient. Event segmentation determines the semantic units to be memorized; compression determines which visual evidence survives within each event; and memory organization determines whether historical and recent evidence can be efficiently retrieved at inference time. This motivates a unified event-centric memory mechanism that jointly models adaptive event discovery, change-aware visual compression, and complementary long–short-term evidence preservation.

To this end, we introduce **EventMem**, an event-centric memory framework without additional training for understanding streaming video. EventMem progressively converts incoming observations into compact and retrievable event-level memory. First, **Temporal-aware Adaptive K-Means Event Segmentation (TAKC)** dynamically discovers event structure from incoming frames, adaptively determines the number of sub-events, and localizes temporal boundaries. Second, **Temporal Segmentation and Aligned Frame Compression (TSAFM)** distinguishes temporally changing frames from stable observations, preserving the former while merging redundant stable-frame tokens through aligned fusion. Third, **Complementary Event-Residual Streaming Memory** stores compressed historical events as long-term memory while maintaining residual short-term visual evidence for recent details, transitions, and information insufficiently represented by historical summaries. Finally, question-relevant historical events are retrieved and fused with short-term evidence, enabling the frozen MLLM to reason over both compact long-range context and fine-grained recent observations without additional training.

We evaluate EventMem on multiple streaming video understanding benchmarks and demonstrate its superiority among previous methods. Our work offers the following key contributions:

- We provide **EventMem**, an event-centric memory architecture for understanding streaming video with adaptive event discovery, compression, and complementary memory construction.
- We introduce an **event-aware compression strategy** that detects temporal boundaries and compresses stable observations while preserving frames with significant temporal changes.
- We develop a **complementary event-residual memory** that combines compact long-term event presentation with granular short-term residuals for efficient historical and recent retrieval.
- Extensive experiments across streaming video benchmarks demonstrate consistent improvements, particularly on object search (by 11.7 points), long-term memory (by 6.7 points), and short-term memory tasks (by 3.1 points).

2 Related Work

2.1 Streaming Video Understanding.

The rising adoption of MLLMs in real-world applications has fueled interest in streaming video understanding, where models must perceive evolving visual scenes and respond in real time (Chen et al. (2024a); Qian et al. (2024); Zhang et al. (2024a)). Despite MLLMs have succeeded in offline video understanding tasks where full video sequences are accessible (Wang et al. (2024); Zeng et al. (2025)), they still fall short in streaming scenarios that must efficiently manage historical information while focusing on the present context for timely and accurate responses. To this end,

recent research has mainly proceeded along three lines.. A first approach transforms the existing offline MLLMs into a streaming assistant by adding extra modules, such as a memory buffer, round-decayed compression, and a proactive response model (Wang et al. (2025); Jin et al. (2024)). A second strategy concentrates on memory construction, which aims to effectively memorize and efficiently retrieve historical information, thereby improving real-time response speed(Xiong et al. (2025); Xie et al. (2026); Zeng et al. (2026)). A third method focuses on the visual token compression to boost computational efficiency while maintaining semantic faithfulness (Chen et al. (2026); Xu et al. (2025)).

2.2 Streaming Video Memory Construction

To overcome the limited context window of MLLMs, recent works construct external or recurrent memories for long and streaming videos. Current approaches can be broadly grouped into three families. One group builds persistent memory banks or recurrent states, where historical clips are compressed and propagated across time, as in MA-LMM, VideoStreaming, and Flash-VStream (He et al. (2024); Qian et al. (2024); Zhang et al. (2024a)). Another group extends long-video context through dense-frame or hierarchical abstraction, enabling models such as StreamChat to access long temporal histories or support multi-round streaming dialogue (Xiong et al. (2025)). The other group organizes streaming history at the event level, as in streamforest, Event-VStream, and EventMemAgent (Liang et al. (2026); Zeng et al. (2026); Guo et al. (2026); Wen et al. (2026)). However, those architectures rely on hierarchical tree clustering or uniform intra-event frame downsampling, EventMem differentiates itself through change-aware stable-frame merging (TSAFM) coupled with an anti-summary residual visual buffer.

2.3 Streaming Video compression

Reducing visual redundancy is essential for scaling video MLLMs to long or streaming inputs. Existing methods mainly fall into three categories. Frame- and spatiotemporal-level methods reduce redundant frames or visual regions before reasoning; LongVU performs query-aware spatiotemporal compression, while VideoStreaming jointly compresses video states during memory propagation (Shen et al. (2024); Qian et al. (2024)). Methods of token pruning remove less informative visual tokens inside the vision encoder or LLM, including FastV, SparseVLM, MustDrop, VisionZip, and LLaVA-Scissor (Chen et al. (2024b); Zhang et al. (2024c); Liu et al. (2024); Yang et al. (2025); Sun et al. (2025)). Token merging methods preserve information by aggregating redundant tokens, as in FrameFusion and recent unified spatiotemporal compression approaches (Fu et al. (2025b); Du et al. (2026)). Streaming-specific methods such as StreamingTOM further impose causal token reduction and bounded online memory (Chen et al. (2026)). Different from token-centric compression, our method compresses videos at the event-memory level, producing compact but temporally structured evidence for streaming understanding.

3 Method

Given a video stream and user questions with timestamps as input, EventMem aims to support efficient online video reasoning under restricted context. As illustrated in Figure 2 Our framework maintains a compact yet informative memory state without any additional training and can adapt to different MLLMs, such as LongVA, Qwen-VL. EventMem consists of three key components: an adaptive event partitioning and token compression module (§3.1), which dynamically segments the incoming video stream into event-level units and compresses redundant stable frames while preserving visually changing frames; a complementary event-residual streaming memory system (§3.2), which stores compressed historical events in long-term memory and retains recent, detailed, and under-summarized raw visual tokens in short-term memory; and a memory-augmented inference strategy (§3.3), which retrieves question-relevant long-term events and fuses them with short-term visual evidence and textual captions.

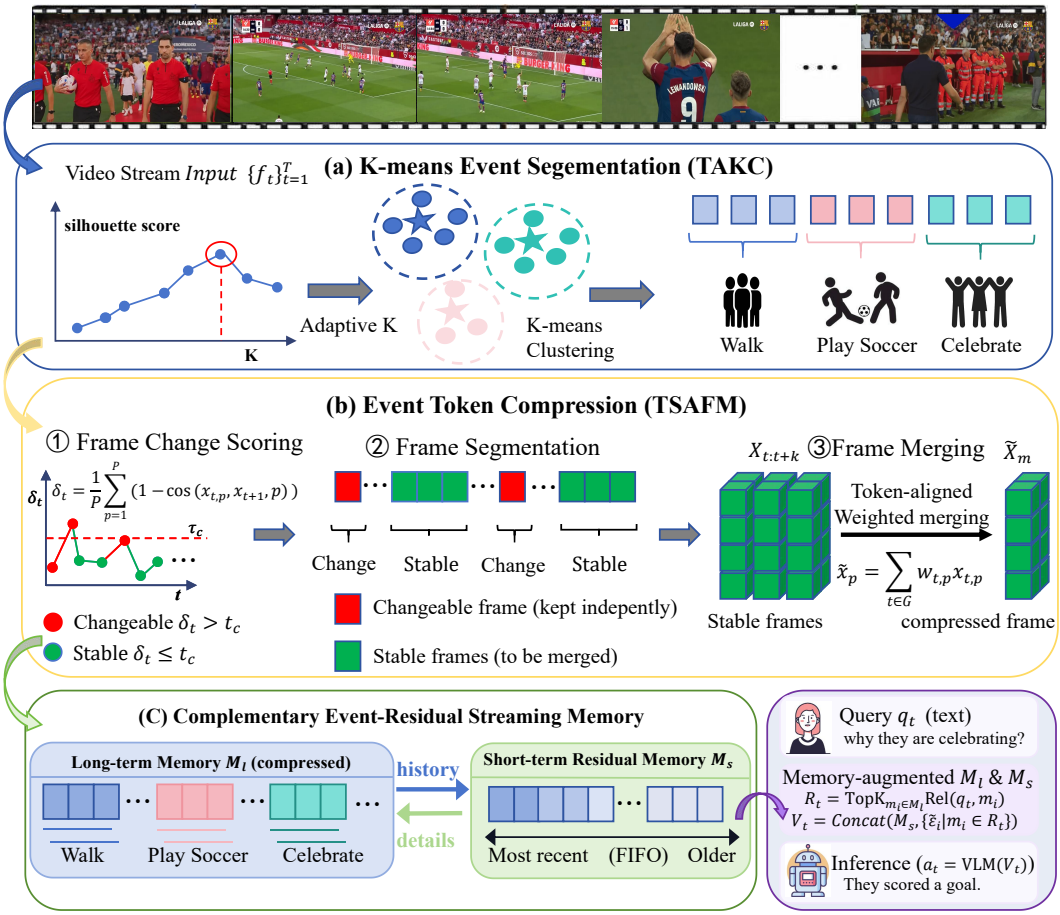


Figure 2: The overall structure of EventMem. (a) is the fisrt step for Event Segmentaion, (b) is the second step for memory compression, (c) is the third step for memory construction

3.1 Adaptive Event Partitioning and Token Compression

3.1.1 Temporal-aware Adaptive K-Means Event Segmentation

Given a newly arrived video segment, we first encode each sampled frame with the frozen encoder and obtain a sequence of frame-level visual tokens $X = \{X_1, X_2, ..., X_T\}, X_t \in R^{P\times d}$. T, P and d denote the count of frames in the current buffer, visual tokens per frame, and hidden dimension, respectively. We obtain a frame-level representation by average pooling $z_t = \frac{1}{P}\sum_{p=1}^{P} x_{t,p}$. To adaptively discover event structure, we perform K-means over $\{z_t\}_{t=1}^T$ with candidate cluster number $K \in [2, K_{max}]$:

$$\min_{\{\mu_k\},\{y_t\}} \sum_{t=1}^{T} \|\mathbf{z}_t - \mu_{y_t}\|_2^2 . \tag{1}$$

For each K, we evaluate clustering quality using the silhouette score:

$$s_t = \frac{b_t - a_t}{\max(a_t, b_t)}, \qquad S(K) = \frac{1}{T}\sum_{t=1}^{T} s_t. \tag{2}$$

where a_t is the average intra-cluster distance of frame t, and b_t is its minimum average distance to other clusters. The event number is selected as :

$$K^* = \begin{cases} \arg\max_K S(K), & \max_K S(K) \ge \tau_s, \\ 1, & \text{otherwise.} \end{cases} \tag{3}$$

After clustering, we choose one key frame from each cluster as the frame closest to its centroid:

$$r_c = \arg\min_{t:\, y_t = c} \|\mathbf{z}_t - \mu_c\|_2 . \tag{4}$$

Since clustering does not preserve temporal order, we sort key frames by timestamp. For two adjacent key frames r_i and r_{i+1}, the event boundary is located by

$$b_i = \arg\max_{t \in [r_i, r_{i+1}]} \left(\|\mathbf{z}_t - \mathbf{z}_{r_i}\|_2 + \left\|\mathbf{z}_t - \mathbf{z}_{r_{i+1}}\right\|_2 \right). \tag{5}$$

The resulting boundaries split the buffer into temporally ordered event segments $\{\mathcal{E}_i\}_{i=1}^{K^*}$.This adaptive partitioning avoids fixed-window segmentation and provides event-level units for subsequent token compression and memory construction.

3.1.2 Temporal Segmentation and Aligned Frame Compression

After adaptive event partitioning, each event segment contains a sequence of frame-level visual tokens $\mathcal{E} = \{\mathbf{X}_t\}_{t=1}^{L}, \mathbf{X}_t \in \mathbb{R}^{P \times d}$. L is the number of frames in the event. Since adjacent frames within the same event often contain substantial redundancy, we compress stable frames while preserving frames with obvious visual changes. We first compute the token-level temporal change between adjacent frames:

$$\delta_t = \frac{1}{P} \sum_{p=1}^{P} \left(1 - \cos\left(\mathbf{x}_{t,p}, \mathbf{x}_{t+1,p}\right)\right), \qquad t = 1, \ldots, L-1. \tag{6}$$

A frame is marked as a change frame if it has a large difference with its previous or next frame:

$$m_t = \mathbb{I}\left(\delta_{t-1} > \tau_c\right) \vee \mathbb{I}\left(\delta_t > \tau_c\right), \tag{7}$$

where τ_c is the change threshold. Change frames are directly retained to preserve transient visual evidence. For the remaining stable frames, we merge each maximal consecutive stable group G. Since frames inside G are temporally close and spatially aligned, we fuse tokens at the same spatial position. For token position p, the merged token is

$$\bar{\mathbf{x}}_p = \sum_{t \in G} \alpha_{t,p} \mathbf{x}_{t,p}, \tag{8}$$

where the weight $\alpha_{t,p}$ is computed by

$$\alpha_{t,p} = \frac{\exp\left(\beta \left\|\mathbf{x}_{t,p} - \mathbf{c}_p\right\|_2\right)}{\sum_{j \in G} \exp\left(\beta \left\|\mathbf{x}_{j,p} - \mathbf{c}_p\right\|_2\right)}, \tag{9}$$

and

$$\mathbf{c}_p = \frac{1}{|G|} \sum_{t \in G} \mathbf{x}_{t,p}. \tag{10}$$

is the initial group center. This weighting assigns higher importance to tokens that deviate more from the stable-group average, helping preserve discriminative details during fusion. The compressed event representation is then formed by concatenating retained change-frame tokens and merged stable-frame tokens:

$$\tilde{\mathcal{E}} = \mathrm{Concat}\left(\{\mathbf{X}_t \mid m_t = 1\}, \{\bar{\mathbf{X}}_G\}\right). \tag{11}$$

In this way, the proposed temporal stable-frame token compression reduces redundant visual tokens within each event while keeping temporally changing frames intact. Compared with uniform token pruning, it applies compression only to stable regions and better preserves short-term visual changes needed for fine-grained reasoning.

3.2 Complementary Event-Residual Streaming Memory

After event partitioning and token compression, we maintain a complementary memory state $M_t = \{M_t^L, M_t^S\}$. M_t^L stores compact historical events and M_t^S preserves recent fine-grained visual evidence. For long-term memory, each event is represented as a multimodal memory node $m_i =$

$(\tilde{\mathcal{E}}, \mathbf{K}_i, c_i)$. $\tilde{\mathcal{E}}$ denotes the compressed event tokens, $\mathbf{K}_i$ is the set of selected key-frame tokens, and c_i is the event caption generated by the frozen VLM. The long-term memory is updated by appending new event nodes:

$$\mathcal{M}_t^L = \mathcal{M}_{t-1}^L \cup \{m_i\}_{i=1}^{N_t}. \tag{12}$$

To avoid fragmented events across adjacent buffers, we merge neighboring memory nodes when their visual similarity is high:

$$\mathrm{sim}(m_i, m_j) = \cos\left(\mathrm{Pool}(\mathbf{K}_i), \mathrm{Pool}(\mathbf{K}_j)\right), \tag{13}$$

$$m_i \leftarrow m_i \cup m_j, \qquad \text{if } \mathrm{sim}(m_i, m_j) > \tau_m. \tag{14}$$

While long-term memory provides compact historical context, it may lose local details during compression and captioning. Therefore, we introduce an anti-summary short-term memory that keeps raw visual tokens complementary to long-term memory. For each candidate frame f_i, we compute a selection score

$$s_i = \lambda_r R_i + \lambda_d D_i + \lambda_b B_i + \lambda_u U_i, \tag{15}$$

where R_i measures recency, D_i measures frame-level visual detail, B_i measures temporal boundary change, and U_i measures how much the frame is not covered by long-term memory. Specifically, given the pooled frame feature $\mathbf{z}_i$, the boundary score is

$$B_i = \max\left(1 - \cos(\mathbf{z}_i, \mathbf{z}_{i-1}),\ 1 - \cos(\mathbf{z}_i, \mathbf{z}_{i+1})\right). \tag{16}$$

and the long-memory residual score is

$$U_i = 1 - \max_{\mathbf{v} \in \mathcal{V}^L} \cos\left(\mathbf{z}_i, \mathbf{v}\right), \tag{17}$$

where $\mathcal{V}^L$ contains pooled visual representations from long-term memory nodes. We first retain the most recent K_r frames, and then greedily select additional frames by penalizing redundancy with already selected frames:

$$i^* = \arg\max_i \left(s_i - \gamma \max_{j \in \mathcal{S}} \cos\left(\mathbf{z}_i, \mathbf{z}_j\right) \right). \tag{18}$$

The selected raw frame tokens form the short-term memory $\mathcal{M}_t^S$. In this design, long-term memory summarizes historical events for efficient retrieval, while short-term memory preserves recent, detailed, transitional, and under-summarized visual evidence. The two memories are therefore complementary rather than redundant.

3.3 MEMORY-AUGMENTED INFERENCE

Given a timestamped question q_t, we combine short-term visual evidence with retrieved long-term events for answer generation. The short-term memory is directly used as recent fine-grained context:

$$\mathbf{V}_t^S = \mathrm{Concat}\left(\mathcal{M}_t^S\right). \tag{19}$$

We retrieve the top-k long-term memory nodes according to their relevance to q_t, measured by a weighted combination of caption-text similarity and key-frame visual similarity:

$$\mathcal{R}_t = \mathrm{TopK}_{m_i \in \mathcal{M}_t^L} \mathrm{Rel}(q_t, m_i). \tag{20}$$

The retrieved event tokens are concatenated with short-term tokens:

$$\mathbf{V}_t = \mathrm{Concat}\left(\mathbf{V}_t^S, \left\{\tilde{\mathcal{E}}_i \mid m_i \in \mathcal{R}_t\right\}\right). \tag{21}$$

Meanwhile, the retrieved event captions are appended to the textual prompt as memory context. The frozen VLM then generates the answer by

$$a_t = \mathrm{VLM}\left(q_t, \{c_i \mid m_i \in \mathcal{R}_t\}, \mathbf{V}_t\right). \tag{22}$$

This training-free inference procedure enables the model to use both recent visual details and compact historical memories.

Table 1: Comparison resluts on the StreamingBench Real-Time benchmark, which are listed in accuracy (%). Among training-free online methods, **bold** and underlined values represent the best and second-best results, respectively.

Methods	OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	Avg.
Performance by Human	89.47	92.00	93.60	91.47	95.65	92.52	88.00	88.75	89.74	91.30	91.46
Offline MLLMs											
LongVA	70.03	63.28	61.20	70.92	62.73	59.50	61.11	53.66	54.67	34.72	59.96
GPT-4o	77.11	80.47	83.91	76.47	70.19	83.80	66.67	62.19	69.12	49.22	73.28
LLaVA-OneVision	80.38	74.22	76.03	80.72	72.67	71.65	67.59	65.45	65.72	45.08	71.12
InternVL2	68.12	60.94	69.40	77.12	67.70	62.93	59.26	53.25	54.96	56.48	63.72
Qwen2.5-VL	78.32	80.47	78.86	80.45	76.73	78.50	79.63	63.41	66.19	53.19	73.68
Online MLLMs (Training-Based)											
VideoLLM-online	39.07	40.06	34.49	31.05	45.96	32.40	31.48	34.16	42.49	27.89	35.99
Flash-VStream	25.89	43.57	24.91	23.87	27.33	13.08	18.52	25.20	23.87	48.70	23.23
ViSpeak	79.84	88.28	83.28	81.05	76.40	75.08	70.37	65.85	77.34	34.20	74.36
StreamForest	83.11	82.81	82.65	84.26	77.50	78.19	76.85	69.11	75.64	54.40	77.26
StreamBridge	84.74	82.68	88.92	89.77	77.36	85.36	84.26	69.92	71.67	35.75	77.04
Online MLLMs (Training-Free)											
LiveVLM	81.50	78.10	83.30	79.10	69.60	74.10	75.00	69.10	67.70	40.40	72.90
ReKV	74.40	78.90	78.60	77.10	68.30	67.90	67.60	62.60	64.30	44.60	69.10
StreamChat	–	–	–	–	–	–	–	–	–	–	60.14
FreshMem	**84.47**	**83.59**	77.60	83.28	**78.12**	**80.37**	70.37	**74.39**	66.86	30.05	74.20
OASIS	–	–	–	–	–	-	–	–	–	–	70.59
LongVA + **EventMem**	68.02	65.63	70.03	75.00	66.67	64.80	64.81	47.97	52.84	46.81	62.68
Qwen2.5-VL + **EventMem**	81.03	73.43	**86.43**	**86.85**	77.34	76.32	**80.56**	69.51	**69.87**	**47.34**	**76.00**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks. To thoroughly assess EventMem, we adopt three benchmarks for streaming video: Streaming-Bench real-time (Lin et al. (2026)), StreamBench (Xiong et al. (2025)), and OVO-Bench real-time (Niu et al. (2025)). StreamingBench real-time measures video understanding across real-time vision, multi-source input, and context comprehension using five-question multi-turn interactions per video. StreamBench is a streaming video understanding benchmark centered on open-ended question answering, using multi-turn dialogues to evaluate real-time perception, contextual understanding, and reasoning over continuous video streams. OVO-Bench Real-Time measures the ability to respond to events occurring at the current timestamp in a streaming video.

Baselines. To provide a comprehensive comparison, we select three kinds of models. Firstly, We benchmark against offline MLLMs to define an upper bound on performance for general visual reasoning, which includes LongVA Zhang et al. (2024b), GPT-4o OpenAI (2024), LLaVA-OneVision Li et al. (2025a), InternVL2 Chen et al. (2024c), Qwen2.5-VL Bai et al. (2025b), Video-LLaVA Lin et al. (2024a), VILA1.5 Lin et al. (2024b), LLaMA-VID Li et al. (2025b). Secondly, we conduct comparison with training -Based online MLLMs, including VideoLLM-online Chen et al. (2024a), Flash-VStream Zhang et al. (2024a), ViSpeakFu et al. (2025a) , StreamForest Zeng et al. (2026), StreamBridge Wang et al. (2025). Finally, we also compare with training-free online MLLMs, such as LiveVLM Ning et al. (2026), ReKV Di et al. (2025), FreeVA Wu (2024), StreamChat Xiong et al. (2025), FreshMem Li et al. (2026), OASIS Liang et al. (2026).

Implementation Details. We conduct our experiments using Qwen2.5-VL-7B Bai et al. (2025b), LongVA-7BZhang et al. (2024b). For EVENTMEM's complementary event-residual memory, we set the maximum adaptive event number to $K_{max} = 5$, the change threshold to $\tau_c = 0.6$, the most recent frames to $K_r = 20, 40, 40$, and the max length of residual frames to 40, 60,60 on StreamBench, StreamingBench real-time and OVO-Bench real-time, respectively. Two NVIDIA A800 80GB GPUs were needed for all experiments .

4.2 MAIN RESULTS

StreamingBench real-time's results. As reported in Table 1, EventMem obtains the average accuracy of **76.00%**, establishing the best performance among training-free online methods. It surpasses the strongest training-free competitor, FreshMem, by **1.80** points, while remaining com-

Table 2: Comparison resluts on StreamBench. We report both the evaluation score (*Sco.*) and accuracy (*Acc.*(%)) for six capability dimensions and their average. Among training-free online methods, **bold** and underlined values represent the best and second-best results, respectively.

Methods	OS		LM		SM		CI		KG		SF		Avg.	
	Sco.	Acc.	Sco.	Acc.	Sco.	Acc.	Sco.	Acc.	Sco.	Acc.	Sco.	Acc.	Sco.	Acc.
Performance by Human	3.95	71.8	3.81	69.3	4.07	81.5	4.14	82.6	4.06	80.7	4.30	80.7	4.06	77.8
Offline MLLMs														
LongVA	2.61	41.8	2.81	47.4	3.20	57.6	3.29	59.8	4.01	80.7	3.48	66.1	3.23	58.9
Video-LLaVA	2.25	31.2	2.31	35.9	2.50	41.8	3.18	56.1	3.81	74.6	2.93	54.8	2.83	49.1
GPT-4o-50	3.27	60.5	3.35	61.2	3.41	64.4	3.81	72.3	4.58	93.9	3.83	74.7	3.71	71.2
GPT-4o-35	3.22	59.6	3.28	58.6	3.45	65.3	3.76	71.7	4.45	93.3	3.50	66.1	3.61	69.1
GPT-4o-mini-35	2.52	46.8	2.70	45.8	2.80	51.0	3.50	64.0	4.67	95.2	2.90	53.3	3.18	59.4
InternVL2	2.49	38.5	2.70	46.6	2.89	50.9	3.61	67.6	4.02	81.0	3.29	62.2	3.17	57.8
VILA1.5	2.33	36.1	2.54	44.3	2.87	50.8	3.59	68.3	3.97	78.6	3.38	65.5	3.11	57.3
LLaMA-VID	2.32	33.9	2.43	38.2	2.63	44.1	3.31	58.4	3.93	76.9	3.06	57.1	2.95	51.4
Qwen2.5-VL	–	36.1	–	41.3	–	41.0	–	50.5	–	76.5	–	73.8	–	53.2
Online MLLMs (Training-Based)														
VideoLLM-online	2.61	41.4	2.87	48.8	3.01	52.9	3.31	62.7	3.58	69.2	3.39	64.1	3.13	56.5
Flash-VStream	2.38	37.1	2.64	44.5	2.78	48.6	3.13	58.1	3.34	66.4	3.17	59.2	2.91	52.3
Online MLLMs (Training-Free)														
MovieChat	1.45	18.6	1.42	20.4	1.76	26.5	2.28	42.3	3.39	67.2	2.05	35.8	2.06	35.1
FreeVA	2.39	35.6	2.33	37.5	2.62	43.7	3.16	58.8	4.24	84.0	2.87	53.7	2.94	52.2
StreamChat	**3.01**	51.7	2.93	53.9	3.21	57.8	**3.86**	68.5	4.38	**88.1**	3.57	69.3	3.49	64.9
OASIS	–	47.0	–	53.6	–	54.3	–	**71.7**	–	86.1	–	67.6	–	63.4
LongVA + **EventMem**	2.98	**53.5**	**3.04**	**54.1**	**3.22**	**60.8**	3.52	67.68	**4.40**	87.3	**3.90**	**76.2**	**3.51**	**66.6**
Qwen2.5-VL + **EventMem**	**3.19**	**54.9**	**3.36**	**55.8**	**3.40**	**59.3**	**3.76**	**68.7**	**3.44**	**68.8**	**3.76**	**74.2**	**3.49**	63.6

Table 3: Comparison results on OVO-Bench real-time, which are listed in accuracy (%). Among training-free online methods, **bold** and underlined values represent the best and second-best results, respectively.

Methods	OCR	ACR	ATR	STU	FPD	OJR	Avg.
Performance by Human	93.96	92.57	94.83	92.70	91.09	94.02	93.20
Offline MLLMs							
LLaVA-Next-Video	69.80	59.63	66.38	50.56	72.28	61.41	65.81
GPT-4o	69.80	64.22	71.55	51.12	70.30	59.78	64.46
LLaVA-OneVision	66.44	57.80	73.28	53.37	71.29	61.96	62.04
Qwen2-VL-7B	40.40	50.46	63.76	47.19	66.34	53.55	50.98
Qwen2.5-VL	76.51	57.80	68.10	46.63	66.34	56.52	60.93
Online MLLMs (Training-Based)							
VideoLLM-online	8.05	23.85	12.07	14.04	45.54	21.20	20.79
Flash-VStream	25.50	32.11	29.31	33.71	29.70	28.80	29.86
ViSpeak	75.17	58.72	71.55	51.12	74.26	66.85	66.28
StreamForest	68.46	53.21	71.55	47.75	61.39	51.63	61.20
StreamBridge	84.56	71.56	74.14	49.44	75.25	72.83	71.30
Online MLLMs (Training-Free)							
StreamChat	51.68	47.71	58.62	41.01	61.06	47.83	50.12
FreshMem	–	–	–	–	–	–	66.67
OASIS	**85.23**	**72.48**	66.38	52.25	67.33	64.67	67.26
Qwen2.5-VL + EventMem	81.88	60.55	**68.10**	**56.18**	**72.28**	**67.39**	**67.38**

petitive with training-based methods such as StreamForest (77.30%) and StreamBridge (77.04%). Notably, EventMem improves LongVA and Qwen2.5-VL by **2.68** and **2.10** average accuracy points, respectively, without any additional training. The gains are especially pronounced on CS, ATP, PR, and SU for Qwen2.5-VL, suggesting that event-aware compression effectively removes redundant streaming observations while the complementary event-residual memory preserves evidence required for temporally detailed reasoning. These results exhibit that EventMem achieves a model-agnostic and training-free mechanism for understanding streaming video better.

StreamBench's results. As reported in Table 2, EventMem substantially improves the memory and retrieval capabilities of the LongVA backbone without additional training. In particular, accuracy on **OS** (Object Search) increases from 41.8% to 53.5%, while **LM** (Long-term Memory) improves from 47.4% to 54.1% and **SM** (Short-term Memory) from 57.6% to 60.8%. These gains directly validate the effectiveness of EventMem in retrieving target-specific visual evidence, preserving long-

term event context, and retaining recent fine-grained information. Overall, EventMem improves the average accuracy of LongVA from 58.9% to **66.5%**, achieving the best overall performance among training-free online MLLMs.

OVO-Bench real-time's results. Table 3 shows that EventMem achieves the best average performance among training-free online MLLMs, reaching 67.73%, which surpasses OASIS by 0.47 points and StreamChat by 17.62 points. The gains are particularly evident on ATR, STU, and FPD (68.10%/56.18%/72.28%), showing that event-aware compression and complementary long–short memory better preserve temporal changes and fine-grained streaming evidence. Notably, EventMem remains competitive with training-based online methods without requiring additional model training.

Overall, across StreamingBench real-time, StreamBench and OVO-Bench real-time, EventMem consistently improves streaming video understanding by jointly addressing visual redundancy and long-term information retention without additional training.

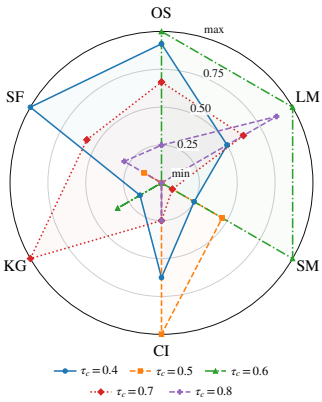


Figure 3: τ_c's effect

4.3 Ablation Study

Table 4 verifies the complementary effects of the proposed components, where the fully ablated baseline (first row) means building long-term memory nodes by directly splitting the video into fixed-length segments based on LongVa, while short-term memory simply keeps the most recent frames in a fixed-length queue. The complete EventMem achieves the best average accuracy of **66.58**, outperforming the baseline by **2.73** points. Removing RSM causes the largest overall degradation from 66.58 to 65.12, with a particularly substantial drop on SM from 60.77 to 54.28, demonstrating the importance of retaining fine-grained recent visual evidence. Meanwhile, removing TAKC or TSAFM also consistently reduces the average performance, confirming that adaptive event discovery and redundancy-aware compression are complementary to residual short-term memory.

Table 4: Ablation results on StreamBench.

ELM		RSM	OS	LM	SM	CI	KG	SF	Avg.
TAKC	TSAFM								
×	×	×	44.82	47.85	59.00	67.00	88.26	76.16	63.85
×	✓	✓	50.84	50.83	58.11	66.33	89.26	**77.15**	65.42
✓	×	✓	49.16	53.80	58.41	**69.02**	89.93	75.17	65.92
✓	✓	×	51.84	50.50	54.28	68.01	**90.27**	75.83	65.12
✓	✓	✓	**53.51**	**54.13**	**60.77**	67.68	87.25	76.16	**66.58**

4.4 Case Study

As shown in Fig. 3, performance remains relatively stable across different values of τ_c, with $\tau_c = 0.6$ providing the best overall trade-off. It achieves the best results on OS, LM, and SM (53.51, 54.13, and 60.77), indicating that a moderate threshold effectively balances temporal change preservation and redundancy suppression. We therefore adopt $\tau_c = 0.6$ by default. As shown in Fig. 4, A larger $K_{\max}$ produces finer event segmentation, which may fragment temporal context and weaken the global coherence of event memory, while a smaller $K_{\max}$ yields overly sparse segmentation with ambiguous event boundaries. Overall, $K_{\max} = 5$ achieves the best performance, suggesting that a moderate event granularity better balances global event semantics and fine-grained temporal discrimination.

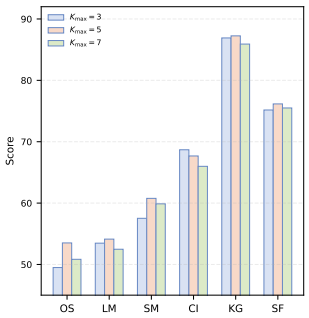


Figure 4: K_{max}'s effect

Conclusion

We introduced EventMem, an event-centric memory framework withot additional training for streaming video understanding. By combining adaptive event segmentation, change-preserving token compression, and complementary long–short-term memory, EventMem efficiently preserves both historical context and recent fine-grained evidence. Extensive experiments on streaming video benchmarks demonstrate consistent improvements in accuracy and response efficiency, highlighting the effectiveness and generalizability of EventMem for long-term streaming video reasoning.

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing and grammar checking of the manuscript. All technical content, research ideas, methodology, experiments, analyses, and conclusions were developed and verified by the authors. We reviewed all AI-assisted edits and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve the recruitment of human subjects or the collection of new personally identifiable or private data.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, experimental settings, datasets, and evaluation protocols in the main paper and appendix. Additional implementation details and hyperparameters are included in the supplementary materials. Source code and configuration files will be released to support reproducibility.

REFERENCES

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. *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 preprint arXiv:2502.13923*, 2025b.

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. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024b.

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. Streamingtom: Streaming token compression for efficient video understanding. In *CVPR*, 2026.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, 2024c. doi: 10.1109/CVPR52733.2024.02283.

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. In *International Conference on Learning Representations*, 2025.

Junhao Du, Jialong Xue, Anqi Li, Jincheng Dai, and Guo Lu. Unified spatiotemporal token compression for video-llms at ultra-low retention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17661–17671, 2026.

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. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21778–21788, 2025a.

Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. Framefusion: Combining similarity and importance for video token reduction on large vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22654–22663, 2025b.

Zhenghui Guo, Yuanbin Man, Junyuan Sheng, Bowen Lin, Ahmed Ahmed, Bo Jiang, Boyuan Zhang, Miao Yin, Sian Jin, Omprakash Gnawali, et al. Event-vstream: Event-driven real-time understanding for long video streams. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3060–3069, 2026.

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.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13700–13710, 2024.

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

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.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 323–340, Cham, 2025b. Springer Nature Switzerland.

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*, pp. 2821–2831, 2026.

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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984. Association for Computational Linguistics, 2024a.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26689–26699, 2024b.

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

Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. Multi-stage vision token dropping: Towards efficient multimodal large language model. *arXiv preprint arXiv:2411.10803*, 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. In *Proceedings of the 63rd ACM/IEEE Design Automation Conference (DAC)*, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18902–18913, 2025.

OpenAI. Gpt-4o system card. `https://openai.com/index/gpt-4o-system-card/`, 2024.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360, 2024.

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*, 2025.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Boyuan Sun, Jiaxing Zhao, Xihan Wei, and Qibin Hou. Llava-scissor: Token compression with semantic connected components for video llms. *arXiv preprint arXiv:2506.21862*, 2025.

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*, 2025.

Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via hybrid architecture. *arXiv preprint arXiv:2409.02889*, 2(5):6, 2024.

Yifei Wang, Zhenkai Li, Tianwen Qian, Huanran Zheng, Zheng Wang, Yuqian Fu, and Xiaoling Wang. Streameqa: Towards streaming video understanding for embodied scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 9422–9432, 2026.

Siwei Wen, Zhangcheng Wang, Xingjian Zhang, Lei Huang, and Wenjun Wu. Eventmemagent: Hierarchical event-centric memory for online video understanding with adaptive tool use. *arXiv preprint arXiv:2602.15329*, 2026.

Wenhao Wu. Freeva: Offline mllm as training-free video assistant. 2024.

Yiweng Xie, Bo He, Junke Wang, Xiangyu Zheng, Ziyi Ye, and Zuxuan Wu. Fluxmem: Adaptive hierarchical memory for streaming video understanding. In *CVPR*, 2026.

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. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 69332–69351, 2025.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams. *arXiv preprint arXiv:2510.09608*, 2025.

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, 2025.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, Lingpeng Kong, Qi Liu, Yuanxing Zhang, and Xu Sun. TimeChat-Online: 80% visual tokens are naturally redundant in streaming videos. *arXiv preprint arXiv:2504.17343*, 2025.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. In *International Conference on Learning Representations*, volume 2025, pp. 38057–38081, 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. *Advances in Neural Information Processing Systems*, 38:75804–75835, 2026.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024a.

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*, 2024b.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024c.