EventMemAgent: Hierarchical Event-Centric Memory for Online Video Understanding with Adaptive Tool Use

Siwei Wen[1], Zhangcheng Wang[3], Xingjian Zhang[1], Lei Huang[1,2] ✉, and Wenjun Wu[1,2]

[1] Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, School of Artificial Intelligence, Beihang University, Beijing, China
[2] Hangzhou International Innovation Institute, Beihang University, Hangzhou, China
[3] 4Paradigm Inc., Beijing, China
huangleiai@buaa.edu.cn

Abstract. Online video understanding requires models to perform continuous perception and long-range reasoning within potentially infinite visual streams. Its fundamental challenge lies in the conflict between the unbounded nature of streaming media input and the limited context window of Multimodal Large Language Models (MLLMs). Current methods primarily rely on passive processing, which often face a trade-off between maintaining long-range context and capturing the fine-grained details necessary for complex tasks. To address this, we introduce **EventMemAgent**, an active online video agent framework based on a **hierarchical memory module**. Our framework employs a dual-layer strategy for online videos: short-term memory detects event boundaries and utilizes event-granular reservoir sampling to process streaming video frames within a fixed-length buffer dynamically; long-term memory structuredly archives past observations on an event-by-event basis. Furthermore, we integrate a **multi-granular perception toolkit** for active, iterative evidence capture and employ **Agentic Reinforcement Learning** (Agentic RL) to end-to-end internalize reasoning and tool-use strategies into the agent's intrinsic capabilities. Experiments show that **EventMemAgent** achieves competitive results on online video benchmarks. The code will be released here: `https://github.com/lingcco/EventMemAgent`.

Keywords: Online video understanding · Video agent

1 Introduction

Online video understanding has become a pivotal task in numerous real-world applications, such as autonomous driving, intelligent surveillance, and robotic assistants. Unlike offline video analysis, these scenarios require models to perform continuous perception and long-range reasoning within potentially infinite visual streams under constrained computational and memory budgets. The fundamental challenge lies in the conflict between the unbounded nature of streaming inputs and the finite context window of MLLMs.

Current studies have primarily sought to alleviate this challenge by managing historical information through passive processing. One line of work focuses on visual token management within the context window, employing techniques such as token pruning or sliding windows [27, 35, 42, 44, 50]. While these methods can mitigate the issue, they inevitably suffer from the decay and loss of historical information as the video stream progresses, and they remain structurally limited by the maximum context window. Another research direction utilizes external memory to preserve visual tokens from past streaming video [5, 8, 10, 36, 41, 45]. During inference, visual information with high similarity to the question is manually added to the model context as a reference. However, such storage imposes substantial overhead and lacks the semantic interpretability required for complex reasoning. More importantly, passive perceiving methods often struggle to achieve precise comprehension of fine-grained details when processing a vast number of visual tokens in a single pass.

We argue that overcoming these limitations requires a paradigm shift from a passive information processing paradigm to an active agentic framework. Instead of merely storing and recalling information passively, an agent can decompose complex video understanding tasks into multiple sub-tasks and actively, iteratively retrieve or perceive task-relevant information [15, 24, 32, 37, 52]. This approach allows the model to maintain long-range video understanding while executing fine-grained tasks with precision. However, the potential of such agentic intelligence in online video scenarios remains underexplored. Current online video agents are often restricted to limited toolsets and simplistically designed memory modules [23], which constrains their effectiveness in online video understanding tasks. Furthermore, these systems rely heavily on manual prompt engineering [36], failing to internalize reasoning and tool-invocation strategies into the agent's intrinsic capabilities through end-to-end training.

To bridge the gap between the active agentic paradigm and online video practice, we propose **EventMemAgent**, an online video agent framework based on a **hierarchical memory module**. Our framework introduces a dual-layer memory management strategy tailored for online video scenarios. Specifically, the short-term memory utilizes event-granular reservoir sampling to dynamically process streaming frames and detect event boundaries, while the long-term memory structuredly archives past observations into event-centric tuples, including captions, visual anchors, and change logs. Furthermore, we integrate a **multi-granular toolkit**, such as memory search and object detection, enabling the agent to actively and iteratively capture evidence across different levels of information granularity. Crucially, we employ **Agentic RL** to end-to-end optimize reasoning paths and tool-invocation policies, effectively internalizing the agentic framework into its intrinsic decision-making process.

The main contributions of this work are summarized as follows:

- We propose **EventMemAgent**, an active online video agent framework that enables precise comprehension by decomposing tasks and iteratively perceiving relevant information. By utilizing Agentic RL, our framework internalizes reasoning and tool-use strategies into its intrinsic capabilities.

- We design a **hierarchical memory module** that coordinates a short-term memory for online event segmentation and event-granular reservoir sampling with a long-term memory for structured event-centric tuples. This dual-layer architecture enables efficient event-centric management of infinite video streams.
- Experiments show that **EventMemAgent** achieves strong results on online video benchmarks using lightweight visual input. This demonstrates that our framework can handle infinite video streams efficiently while staying within fixed hardware constraints.

2 Related Work

2.1 Online Video Understanding

Online video understanding is designed to process continuous and potentially infinite visual inputs under fixed hardware constraints. However, most existing MLLMs [2,3,14,16,19,47,48] are primarily designed for predefined offline videos and struggle to handle online videos.

Therefore, recent research has focused primarily on processing online video through passive information processing strategies. One major research direction manages visual tokens through passive processing strategies within the context window, such as token pruning, sliding windows, and KV cache compression [27,35,42,44,50]. For instance, *StreamingVLM* [35] employs a sliding window to discard past information, while *VideoStreaming* [27] utilizes a trained encoder to compress historical information into fixed-length tokens. While efficient, these passive methods inevitably suffer from information decay as the video stream progresses and remain structurally limited by the maximum context window of the underlying model.

Another research direction explores utilizing external memory modules to archive historical visual tokens [5,8,36,41,45], manually retrieving relevant tokens based on similarity during inference. For example, *StreamKV* [5] compresses and offloads visual KV caches to external storage, while *Venus* [41] selects keyframes and stores their embeddings in memory. Although these approaches circumvent context length constraints, their passive recall mechanisms hinder the model's ability to achieve precise comprehension among vast tokens. Furthermore, such feature-level storage lacks the semantic interpretability required for complex reasoning.

In contrast, **EventMemAgent** transitions from passive processing to an active agentic framework. Supported by an event-centric hierarchical memory, it ensures long-term coherence through dynamic segmentation and structured archiving. Crucially, our agent actively and iteratively perceives task-relevant evidence via a multi-granular toolkit, overcoming the limitations of passive similarity-based recall.

2.2 Video Understanding Agents

Recent advancements in long video understanding have explored agent-based frameworks as a viable paradigm for processing large-scale visual inputs [15, 24, 32, 37, 51, 52]. Unlike conventional video MLLMs that struggle with lengthy sequences, these systems leverage the high-level reasoning, planning, and memory management capabilities of Large Language Models (LLMs). By decomposing complex video tasks into manageable sub-tasks and iteratively retrieving and synthesizing task-relevant information, these agents effectively enhance the performance in long video understanding.

In the context of online video, although some recent works [23,36] have begun to incorporate agentic frameworks, significant limitations persist. For instance, *M3-Agent* [23] is restricted to rudimentary memory retrieval tools during inference, failing to leverage a diverse toolset to unlock the potential of the agentic paradigm; moreover, its simple text-only memory module, which relies on fixed-length segmentation, remains suboptimal for capturing long-term temporal dynamics. Meanwhile, *StreamAgent* [36] relies heavily on manual prompt engineering to guide its decision-making, lacking exploration into utilizing Agentic RL to internalize reasoning and tool-invocation strategies as intrinsic capabilities. Our work addresses these limitations by introducing a hierarchical memory module and a multi-granular toolkit, leveraging reinforcement learning to optimize the agent's planning and execution within dynamic streaming environments.

3 Methodology

3.1 Overview

We propose an active online video agent framework designed to shift the passive information processing paradigm toward proactive perception for online video. As illustrated in Figure 1, our architecture consists of three integrated components: (1) a **Hierarchical Memory Module** designed to overcome information decay and semantic fragmentation by dynamically archiving visual streams into structured, event-centric representations; (2) a **Multi-granular Perception Toolkit** that enables the agent to actively and iteratively capture fine-grained evidence across diverse levels of granularity; and (3) an **Agentic Reinforcement Learning** framework that internalizes reasoning and tool-invocation strategies into the agent's intrinsic capabilities through end-to-end optimization.

3.2 Hierarchical Memory Module

To address the challenge of processing infinite video streams within a finite context window while overcoming information decay and semantic fragmentation, we design an event-centric hierarchical memory system. The architecture consists of a Short-Term Memory (STM) for immediate visual buffering and a Long-Term Memory (LTM) for structured history archiving. We consistently employ a sampling rate of 1 FPS on the incoming video stream $V = \{f_1, f_2, \ldots, f_t \ldots\}$, where

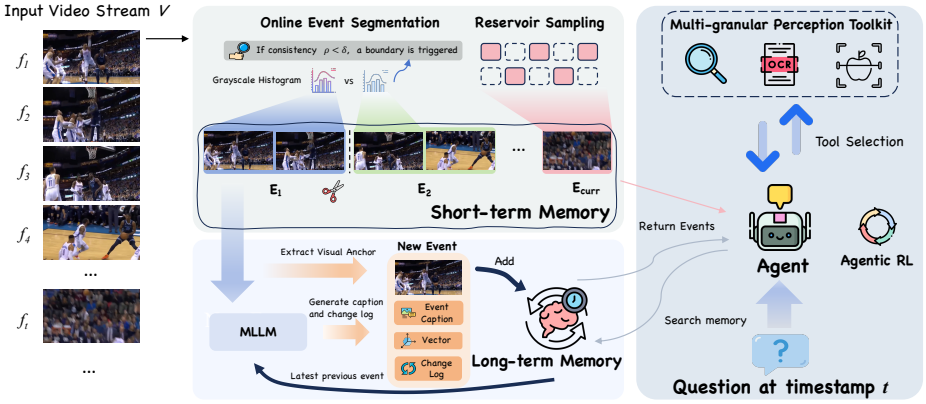

Fig. 1: Overview of EventMemAgent. It consists of three core components: a hierarchical memory module that archives video streams into structured event-centric representations, a multi-granular perception toolkit for active, iterative evidence capture, and an agentic reinforcement learning framework that optimizes tool-use strategies. The agent dynamically retrieves memory and utilizes perception tools to answer questions at specific timestamps.

f_t denotes the frame captured at time step t, to extract discrete frames before proceeding with further processing.

Event-Centric Short-Term Memory The Short-Term Memory operates as a fixed-size buffer with a maximum capacity of K frames, directly accessible during inference. Instead of a flat frame sequence, we structure the STM as a collection of m semantically coherent events. Formally, let $\mathcal{E}_{st}$ denote the set of events in the short-term memory:

$$\mathcal{E}_{st} = \{E_1, E_2, \ldots, E_m\}, \tag{1}$$

where $\{E_1, \ldots, E_{m-1}\}$ represents completed historical segments, and E_m denotes the active event accumulating incoming frames. Each event $E_i = \{f_{i,j}\}_{j=1}^{n_i}$ consists of n_i frames belonging to the same semantic segment. The total frame count is strictly constrained by $\sum_{i=1}^{m} n_i \leq K$.

Online Event Segmentation. To maintain semantic coherence, for each incoming frame f_t, the system determines whether it belongs to the current active event E_m or signifies a new segment. We evaluate content consistency by comparing the normalized grayscale histogram $\mathbf{h}_t$ of the new frame with the average histogram $\bar{\mathbf{h}}_m$ of all frames currently in E_m. Specifically, we employ the Pearson correlation coefficient ρ as a quantitative metric to measure the distributional similarity between the new frame and the event history. The coefficient is computed as:

$$\rho(\bar{\mathbf{h}}_m, \mathbf{h}_t) = \frac{\mathrm{cov}(\bar{\mathbf{h}}_m, \mathbf{h}_t)}{\sigma_{\bar{\mathbf{h}}_m} \sigma_{\mathbf{h}_t}}, \tag{2}$$

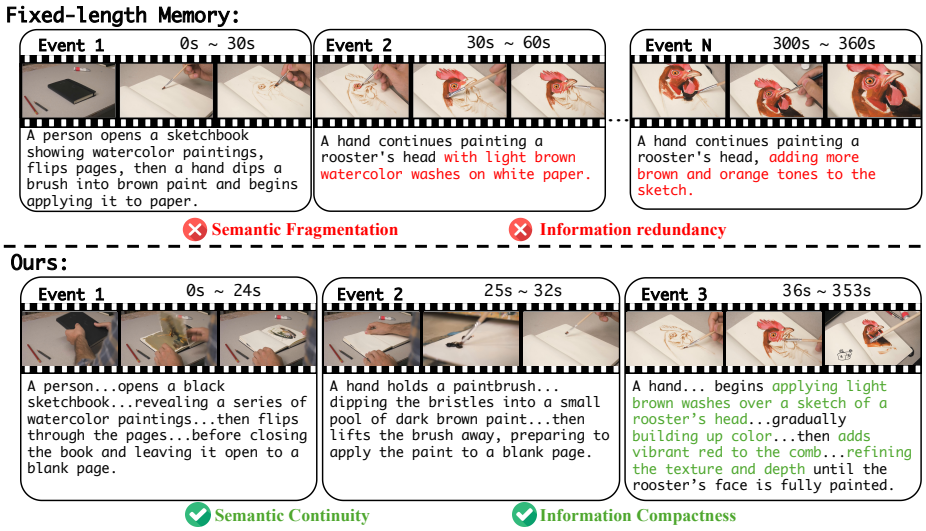

Fig. 2: Comparison of memory management strategies. Fixed-length Memory (top) frequently suffers from semantic fragmentation and information redundancy due to rigid temporal boundaries that bisect continuous actions. In contrast, our event-centric hierarchical memory (bottom) preserves semantic continuity and information compactness.

where $\mathrm{cov}(\cdot)$ and σ denote the covariance and standard deviation calculated over the histogram bins, respectively. A boundary is triggered if the correlation $\rho < \delta$. Upon a boundary, the finalized E_m is archived into the historical memory, and a new event E_{m+1} is initialized with f_t.

Intra-event Reservoir Sampling. Under our online event segmentation mechanism, semantically coherent events—such as those where the subject remains stationary—typically manifest substantial informational redundancy in the visual stream (As shown in Figure 2). To circumvent the inefficient accumulation of redundant data and preclude a single protracted event from depleting the K-frame context budget, we adopt a reservoir sampling strategy within E_m when event segmentation is not triggered.

Specifically, incoming frames are directly appended to the active event E_m as long as the total frame count across all STM events remains below K. Once the STM reaches its K-frame capacity, we resolve the overflow via two mutually exclusive branches: (1) Inter-event FIFO: If the STM contains multiple events, we execute a First-In-First-Out strategy, transitioning the earliest event E_1 to the LTM to accommodate the new frame. (2) Intra-event Reservoir Sampling: If E_m is the sole event in the STM, indicating it has entirely saturated the global budget, we apply reservoir sampling. The n-th processed frame of this event ($n > K$) is accepted with probability $P = K/n$, replacing a randomly selected frame in E_m.

This mechanism guarantees that the inclusion probability for each frame remains uniform, ensuring that E_m consistently serves as an unbiased representative summary of an event stream with an indeterminate duration. Consequently, this approach effectively maximizes the temporal horizon covered within the fixed context window while mitigating short-term memory overflow. See Appendix A for detailed short-term memory update rules.

Structured Long-Term Memory When an event E_i is evicted from the short-term buffer to accommodate new observations, it is archived in the Long-Term Memory as a structured tuple $E_i' = \{I_i^{first}, c_i, \mathbf{e}_i, \Delta_i\}$. Each entry is anchored by the first frame I_i^{first}, providing a visual reference, while the semantic content is captured by a natural language caption c_i generated by MLLMs. To facilitate efficient retrieval, c_i is projected into a high-dimensional space as a semantic embedding $\mathbf{e}_i$. Furthermore, we address the potential for narrative fragmentation by incorporating Change Logs Δ_i, which record the state transitions and logical shifts between consecutive events. This hierarchical transition ensures that while the STM focus remains on immediate dynamics, the historical context is preserved indefinitely in a structured and searchable format.

In summary, the hierarchical memory module reconciles the inherent conflict between unbounded video input and limited context length. Furthermore, as illustrated in Figure 2, our event-centric design maximizes the utility of the limited short-term memory, enhancing narrative continuity and informational compactness while avoiding semantic fragmentation and redundancy common in fixed-length partitioning. Additionally, our long-term memory archives both captions and visual anchors to preserve more fine-grained details. Detailed visualizations of the memory contents are provided in Appendix E.

3.3 Multi-granular Perception Toolkit

To extend the agent's perceptual depth beyond raw visual signals, we integrate a toolkit $\mathcal{T}$ that enables active information acquisition. A key component is the *Memory Search* tool, which empowers the agent to retrieve historical contexts from the LTM through two modalities: (i) Temporal Retrieval, where the agent filters events E_i' whose timestamps $[t_i^{start}, t_i^{end}]$ intersect with a specific query range; and (ii) Semantic Retrieval, which identifies relevant episodes by maximizing the cosine similarity between the query embedding $\mathbf{e}_q$ and stored event embeddings $\mathbf{e}_i$. The retrieval returns the complete event content from the LTM, such as the caption and change log.

Beyond memory retrieval, the agent invokes specialized perception tools to capture fine-grained evidence. Specifically, it utilizes *OCR* to extract textual cues and *Object Detection* to localize specific entities. The agent can autonomously choose to apply these operations either to visual anchors associated with events in the LTM or to specific frames within the STM. The observations are integrated into the context to facilitate further reasoning and response generation.

3.4 Agentic Reinforcement Learning

The agent follows a multi-turn reasoning trajectory adhering to the ReAct paradigm [39]:

$$\tau = \{(t_1, a_1, o_1), \ldots, (t_n, a_n, o_n)\}, \tag{3}$$

In this structure, *Thought* (t) represents the agent's internal reasoning for task decomposition, *Action* (a) specifies either a tool invocation or the terminal response, and *Observation* (o) captures the sensory feedback returned by the perception toolkit.

To effectively internalize the agentic framework's reasoning and tool-invocation strategies into its intrinsic decision-making capabilities, we employ Group Relative Policy Optimization (GRPO) [11]. For each query q, we sample a group of trajectories $\{\tau_1, \ldots, \tau_G\}$ from the old policy $\pi_{\theta_{old}}$. GRPO optimizes the following objective by utilizing group-based relative advantages instead of a separate critic:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(\frac{\pi_\theta(\tau_i|q)}{\pi_{\theta_{old}}(\tau_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(\tau_i|q)}{\pi_{\theta_{old}}(\tau_i|q)}, 1-\epsilon, 1+\epsilon\right)A_i\right)\right] \tag{4}$$

where the advantage A_i is computed by normalizing the rewards within each group: $A_i = (r_i - \text{mean}(\mathbf{r}))/\text{std}(\mathbf{r})$.

In our framework, the reward is assigned solely based on the terminal correctness of the final answer:

$$r_i = \begin{cases} 1, & \text{if the predicted answer is correct} \\ 0, & \text{otherwise} \end{cases}. \tag{5}$$

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate **EventMemAgent** on two popular benchmarks for online video question answering: OVO-Bench [25] and StreamingBench [18]. In these benchmarks, models must follow the online video format, which means they can only use the video content before the question's timestamp to answer. OVO-Bench divides online video tasks into three scenarios: (1) Real-Time Visual Perception, (2) Backward Tracing, and (3) Forward Active Responding, to fully test model performance in different online video settings. StreamingBench includes 12 tasks to evaluate real-time understanding in various situations.

Implementation Details. We employ Qwen3-VL-8B-Instruct [2] as the backbone MLLM for **EventMemAgent**, using Grounding DINO [21] as the object detection model and Deepseek-OCR [33] as the OCR model. By default, input videos are sampled at 1 FPS, and the maximum capacity of short-term memory is set to 32. The content consistency threshold δ is set to 0.2. In short-term memory,

we added frame numbers and timestamps to the images to improve the model's spatiotemporal perception [34]. Since Qwen3-VL already has basic tool-use capabilities, we use 10K MovieChat [29] samples labeled by VideoMarathon [17] for direct end-to-end agentic RL training. All training and evaluation are conducted on eight 80G A100 GPUs. For more deployment details, please refer to Appendix B.

Compared Baselines. Our compared baselines mainly include: (1) Proprietary MLLMs, such as GPT-4o [12] and Gemini-1.5-Pro [30]. (2) Advanced open-source offline MLLMs, including InternVL-V2 [6], Qwen3-VL [2], and LLaVA-Video [16]. To adapt these offline models to the online video setting, fixed FPS or uniform sampling is applied to the video segment before the query timestamp. (3) Recent open-source online MLLMs, such as Flash-VStream [43] and Dispider [26]. (4) Online video agents, StreamAgent [36].

Table 1: The experimental results on the OVO-Bench benchmark. The benchmark contains many tasks, including Optical Character Recognition (OCR), Action Recognition (ACR), Attribute Recognition (ATR), Spatial Understanding (STU), Future Prediction (FPD), Object Recognition (OJR), Episodic Memory (EPM), Action Sequence Identification (ASI), Hallucination Detection (HLD), Repetition Event Count (REC), Sequential Steps Recognition (SSR), and Clues Reveal Responding (CRR).

Model	Params	Frames	Real-Time Visual Perception							Backward Tracing				Forward Active Responding				**Overall**
			OCR	ACR	ATR	STU	FPD	OJR	**Avg.**	EPM	ASI	HLD	**Avg.**	REC	SSR	CRR	**Avg.**	
Human																		
Human Agents	-	-	93.96	92.57	94.83	92.70	91.09	94.02	93.20	92.59	93.02	91.37	92.33	95.48	89.67	93.56	92.90	92.81
Proprietary MLLMs																		
Gemini 1.5 Pro [30]	-	1fps	85.91	66.97	79.31	58.43	63.37	61.96	69.32	58.59	76.35	52.64	62.54	35.53	74.24	61.67	57.15	63.00
GPT-4o [12]	-	64	69.8	64.22	71.55	51.12	70.3	59.78	64.46	57.91	75.68	48.66	60.75	27.58	73.21	59.4	53.40	59.54
Open-Source Offline MLLMs																		
Qwen2-VL-72B [31]	72B	64	65.77	60.55	69.83	51.69	69.31	54.35	61.92	52.53	60.81	57.53	56.95	38.83	64.07	45.00	49.30	56.27
LLaVA-Video-7B [16]	7B	64	69.13	58.72	68.83	49.44	74.26	59.78	63.52	56.23	57.43	7.53	40.4	34.10	69.95	60.42	54.82	52.91
LLaVA-OneVision-7B [14]	7B	64	66.44	57.80	73.28	53.37	71.29	61.96	64.02	54.21	55.41	21.51	43.71	25.64	67.09	58.75	50.50	52.74
Qwen2-VL-7B [31]	7B	64	60.40	50.46	56.03	47.19	66.34	55.43	55.98	47.81	35.48	56.08	46.46	31.66	65.82	48.75	48.74	50.39
InternVL-V2-8B [6]	8B	64	67.11	60.55	63.79	46.07	68.32	56.52	60.39	48.15	57.43	24.73	43.44	26.5	59.14	54.14	46.60	50.15
LongVU-7B [28]	7B	1fps	53.69	53.21	62.93	47.75	68.32	59.78	57.61	40.74	59.46	4.84	35.01	12.18	69.48	60.83	47.50	46.71
Qwen3-VL-8B [2]	8B	64	75.17	58.72	72.41	57.30	70.30	59.24	65.52	56.57	69.59	12.37	46.18	47.13	67.57	52.50	55.73	55.81
Open-Source Online MLLMs																		
Flash-VStream-7B [44]	7B	1fps	24.16	29.36	28.45	33.71	25.74	28.80	28.37	39.06	37.16	5.91	27.38	8.02	67.25	60.00	45.09	33.61
VideoLLM-online-8B [4]	8B	2fps	8.05	23.85	12.07	14.04	45.54	21.20	20.79	22.22	18.80	12.18	17.73	-	-	-	-	-
Dispider [26]	7B	1fps	57.72	49.54	62.07	44.94	61.39	51.63	54.55	48.48	55.41	4.3	36.06	18.05	37.36	48.75	34.72	41.78
StreamForest-7B [42]	7B	1fps	68.46	53.21	71.55	47.75	65.35	60.87	61.20	58.92	64.86	32.26	52.02	32.81	70.59	57.08	53.49	55.57
Online Video Agents																		
StreamAgent [36]	7B	1fps	71.20	53.20	63.60	53.90	67.30	58.70	61.30	54.80	58.10	25.80	41.70	35.90	48.40	52.00	45.40	49.40
Ours	8B	$\leq$32	75.84	69.72	73.28	55.62	67.33	67.93	68.29	59.60	70.95	43.55	58.03	33.67	72.02	62.08	55.92	60.75

Table 2: The experimental results on the real-time visual understanding part of the StreamingBench benchmark. The benchmark contains many tasks, including 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	Params	Frames	OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	**ALL**
Human	-	-	89.47	92.00	93.60	91.47	95.65	92.52	88.00	88.75	89.74	91.30	91.46
Proprietary MLLMs													
Gemini 1.5 pro [30]	-	1 fps	79.02	80.47	83.54	79.67	80.00	84.74	77.78	64.23	71.95	48.70	75.69
GPT-4o [12]	-	64	77.11	80.47	83.91	76.47	70.19	83.80	66.67	62.19	69.12	49.22	73.28
Claude 3.5 Sonnet [1]	-	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 [7]	7B	32	55.86	55.47	57.41	58.17	52.80	43.61	39.81	42.68	45.61	35.23	49.52
VILA-1.5 [22]	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 [9]	14B	96	56.40	57.81	65.30	62.75	64.60	51.40	42.59	47.97	49.58	31.61	53.96
LongVA [46]	7B	128	70.03	63.28	61.20	70.92	62.73	59.50	61.11	53.66	54.67	34.72	59.96
InternVL-V2 [6]	8B	16	68.12	60.94	69.40	77.12	67.70	62.93	59.26	53.25	54.96	56.48	63.72
Qwen2-VL [31]	7B	0.2-1fps	75.20	77.45	73.19	82.81	68.32	71.03	72.22	61.19	61.47	46.11	69.04
Kangaroo [20]	7B	64	71.12	84.38	70.66	73.20	67.08	61.68	56.48	55.69	62.04	38.86	64.60
LLaVA-NeXT-Video [49]	32B	64	78.20	70.31	73.82	76.80	63.35	69.78	57.41	56.10	64.31	38.86	66.96
MiniCPM-V-2.6 [40]	8B	32	71.93	71.09	77.92	75.82	64.60	65.73	70.37	56.10	62.32	53.37	67.44
Qwen3-VL [2]	8B	0.2-1 fps	73.57	71.09	76.34	80.39	65.22	74.77	78.70	67.89	61.47	47.67	70.20
Open-source Online MLLMs													
Flash-VStream [43]	7B	-	25.89	43.57	24.91	23.87	27.33	13.08	18.52	25.20	23.87	48.70	23.23
VideoLLM-online [4]	8B	2 fps	39.07	40.06	34.49	31.05	45.96	32.40	31.48	34.16	42.49	27.89	35.99
Dispider [26]	7B	1 fps	74.92	75.53	74.10	73.08	74.44	59.92	76.14	62.91	62.16	45.80	67.63
TimeChat-Online [38]	7B	1 fps	80.22	82.03	79.50	83.33	76.10	78.50	78.70	64.63	69.60	57.98	75.36
StreamForest [42]	7B	1 fps	83.11	82.81	82.65	84.26	77.50	78.19	76.85	69.11	75.64	54.40	77.26
Online Video Agents													
StreamAgent [36]	7B	1fps	79.63	78.31	79.28	75.87	74.74	76.92	82.94	66.31	73.69	55.40	74.28
Ours	8B	$\leq$32	83.92	76.56	87.70	85.29	77.02	79.44	77.78	68.29	72.24	48.70	77.00

4.2 Main Results

Results on OVO-Bench. We report the detailed results of **EventMemAgent** on OVO-Bench in Table 1. With input frames limited to 32, **EventMemAgent** achieves an average accuracy of 60.75%, surpassing all existing open-source models and even the proprietary model **GPT-4o** (59.54%). Specifically, our model outperforms current open-source methods by 4.27% in Real-Time Visual Perception, 1.08% in Backward Tracing, and 1.1% in Forward Active Responding. These gains are driven by our hierarchical event-centric memory, which organizes streaming video into structured event-level representations. Furthermore, the multi-granular perception tools and agentic RL training allow the model to actively capture fine-grained details and internalize efficient tool-use strategies. This combination ensures high-fidelity perception and precise reasoning within a limited context, leading to superior stability and accuracy across diverse online video scenarios.

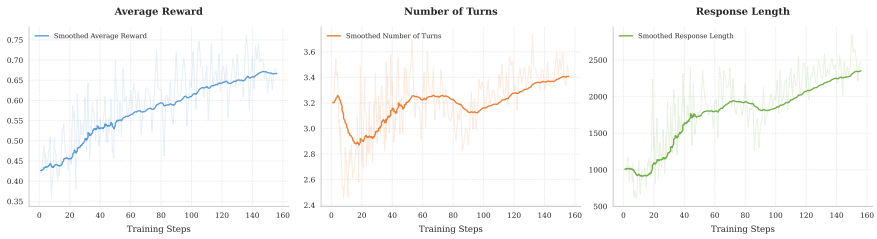

Fig. 3: Training Statistics of Agentic RL. The evolution of Average Reward (left), Number of Turns (middle), and Response Length (right) throughout the training process. The steady increase in both reward and response length demonstrates the model's improving capability in complex reasoning and tool usage.

Results on StreamingBench. We report the detailed performance of **EventMemAgent** on StreamingBench in Table 2, which covers 12 diverse real-time understanding tasks. Our model delivers highly competitive performance, demonstrating its strength in streaming video scenarios. A key observation is that success in these tasks depends heavily on the precise perception of current content rather than the complex integration of long-term history, which suggests that the potential of LTM in our framework has not yet been fully realized. Even with no more than 32 input frames, **EventMemAgent** achieves impressive accuracy, proving its efficiency in processing immediate visual streams. This capability is supported by our multi-granular perception tools and event-centric short-term memory, which enable the agent to capture essential details within the current event window.

Training Statistics We visualize the training dynamics of the Agentic RL process in Figure 3. The **Average Reward** (left) exhibits a steady upward trend, confirming that the GRPO algorithm effectively optimizes the agent's policy. Simultaneously, we observe a significant increase in the **Response Length** (right), which indicates that the model is learning to generate more detailed reasoning chains (CoT) and tool invocations to ensure answer correctness. The **Number of Turns** (middle) initially fluctuates before stabilizing, reflecting the agent's adaptation to the multi-turn interaction environment. These metrics collectively demonstrate the stability and effectiveness of our training framework.

4.3 Efficiency Analysis

To evaluate the practical deployment capabilities of EventMemAgent, we conduct a comprehensive efficiency analysis on the OVO-Bench benchmark, with the model served by vLLM [13] on a single 40GB A100 GPU. As illustrated in Figure 4(a), the median response latency remains remarkably stable at approximately 1.78 seconds, even as the input video stream extends up to 30 minutes. This consistent performance demonstrates that our hierarchical memory module successfully bounds the computational overhead. By structurally archiving past

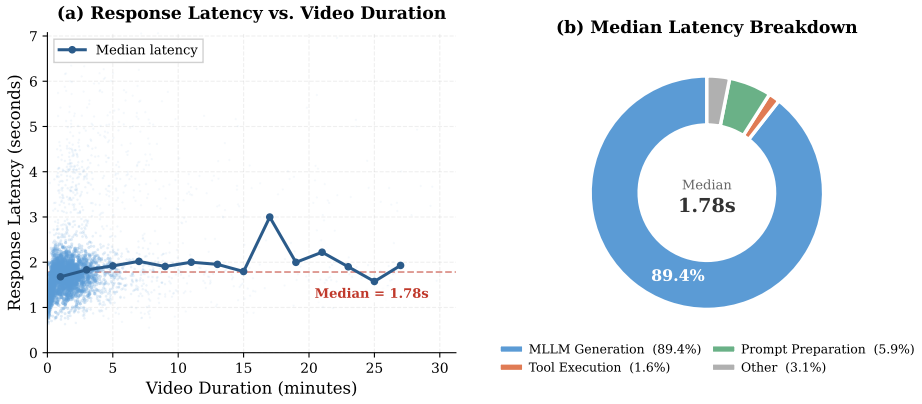

Fig. 4: Efficiency analysis of EventMemAgent on OVO-Bench. (a) The relationship between response latency and video duration. (b) The detailed breakdown of the median response latency.

observations and dynamically managing the active context, the framework effectively prevents the severe latency degradation that typically occurs as historical information accumulates in continuous visual streams.

Furthermore, Figure 4(b) provides a detailed breakdown of the median latency components. The vast majority of the processing time (89.4%) is allocated to MLLM generation, while prompt preparation accounts for 5.9%. Notably, the active invocation of the multi-granular perception toolkit introduces negligible overhead, accounting for only 1.6% of the total latency. This indicates that the efficiency of our framework is primarily determined by the intrinsic generation speed of the underlying MLLM, rather than the memory management mechanisms or tool execution. Consequently, EventMemAgent is highly scalable and well-suited for real-time online video understanding under constrained computational budgets.

4.4 Ablation Study

Effectiveness of hierarchical memory module. To verify the effectiveness of our hierarchical memory system, we compare it with a fixed-length memory while keeping other agent components constant. Specifically, for the fixed-length approach, the input video stream is divided into segments every 30 seconds (such as Figure 2). As shown in Table 3, the fixed-length memory shows a performance decline on both OVO-Bench and StreamingBench compared to our proposed module. This indicates that the combination of online event segmentation and event-granular reservoir sampling improves semantic integrity, helping the model better understand information in the video stream. These results confirm the effectiveness of our hierarchical memory system.

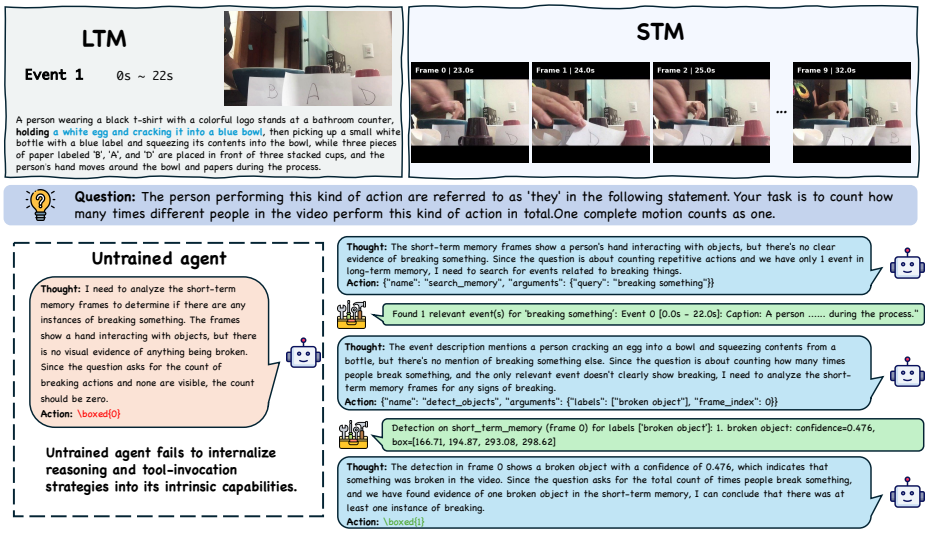

Fig. 5: Qualitative comparison of reasoning trajectories on OVO-Bench. (Left) The untrained agent fails to use tools flexibly. (Right) EventMemAgent successfully internalizes complex reasoning and tool-use strategies, allowing it to adaptively retrieve information from long-term memory and use tools for more precise observations to provide accurate answers.

Effectiveness of Perception Tools. Similarly, we evaluate the impact of the multi-granular perception toolset by comparing the full **EventMemAgent** with versions that exclude specific tools. As shown in Table 4, removing tools such as *OCR* or *Object Detection* results in a performance decrease on OVO-Bench. This shows that while the base MLLM possesses general visual reasoning capabilities, specialized tools are essential to capture fine-grained details like text and small objects, helping the model to understand and reason more comprehensively. These results confirm that multi-granular perception is highly effective for achieving a comprehensive understanding of diverse and complex video scenarios.

Case Study. We provide a qualitative comparison between **EventMemAgent** and an untrained version in Figure 5. While sharing the same architecture, the

Table 3: Comparison between our hierarchical memory and fixed-length segmentation memory on OVO-Bench and StreamingBench (Average Score).

Memory Mechanism	OVO-Bench	StreamingBench
Ours (Hierarchical Memory)	60.75	77.00
Fixed-length Memory	58.27 (↓ 2.48)	75.09 (↓ 1.91)

Table 4: Ablation Study on OVO-Bench. The values in parentheses indicate the performance change compared to the Full model, reported only for average and overall metrics.

Configuration	Real-Time Visual Perception							Backward Tracing				Forward Active Responding				Overall
	OCR	ACR	ATR	STU	FPD	OJR	**Avg.**	EPM	ASI	HLD	**Avg.**	REC	SSR	CRR	**Avg.**	
Ours (Full)	75.84	69.72	73.28	55.62	67.33	67.93	68.29	59.60	70.95	43.55	58.03	33.67	72.02	62.08	55.92	60.75
w/o Detect	75.84	69.72	66.50	55.62	61.20	63.10	65.33 (↓ 2.96)	57.50	68.20	40.50	55.40 (↓ 2.63)	31.50	69.00	57.50	52.67 (↓ 3.25)	57.80 (↓ 2.95)
w/o OCR	64.50	69.00	72.00	55.00	66.80	66.50	65.63 (↓ 2.66)	58.20	69.50	42.50	56.73 (↓ 1.30)	32.50	71.00	60.20	54.57 (↓ 1.35)	58.98 (↓ 1.77)

untrained agent fails to use tools flexibly. In contrast, through Agentic RL, our agent internalizes tool-use strategies into its reasoning process. This bridges the gap between the infinite video stream and the limited context window.

The impact of Agentic RL on the agent's tool-use strategy. As shown in Figure 6, we analyze the tool-use strategies on OVO-Bench before and after Agentic RL training. It is evident that the model before training fails to utilize tools flexibly to solve problems. In Backward Tracing, which primarily requires historical information, the untrained model rarely invokes Search Memory to retrieve necessary data. Furthermore, its strategy is highly polarized, with over 96% of cases either making no tool calls or repeatedly calling tools until the maximum turn limit is reached. In contrast, the trained model actively employs Search Memory for questions regarding past events. In Real-Time Visual Perception, it more frequently utilizes OCR and Detect Objects to enhance immediate perception. For a detailed quantitative validation of the performance improvements brought by Agentic RL, please refer to Appendix C.

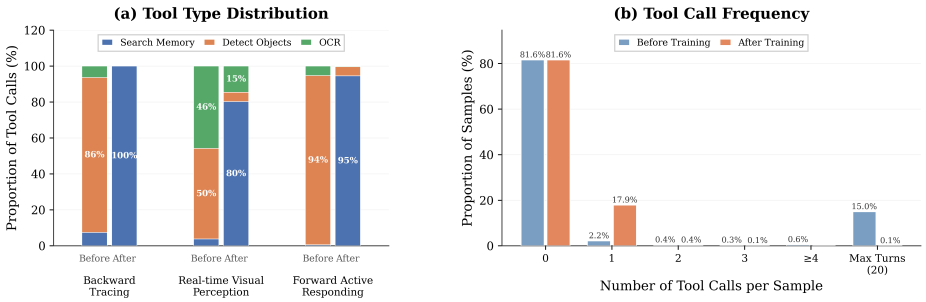

Fig. 6: Analysis of tool usage patterns on OVO-Bench before and after training. (a) Distribution of tool types invoked across the three task categories. (b) Distribution of the number of tool calls per sample.

5 Conclusion

In this paper, we propose **EventMemAgent**, an active agent framework that solves the conflict between infinite video streams and the limited context window of MLLMs. By shifting from passive information processing to proactive perception, our approach ensures long-term semantic coherence while maintaining precise understanding. This is achieved through a hierarchical memory module—integrating online event segmentation and reservoir sampling—and a multi-granular perception toolkit. Furthermore, we use Agentic Reinforcement Learning to internalize reasoning and tool-use strategies into the agent's intrinsic capabilities. Experiments on OVO-Bench and StreamingBench show that **EventMemAgent** achieves competitive performance with low hardware costs. We hope our work inspires further research on autonomous agents for continuous perception in dynamic visual environments.

Acknowledgments

This work was partially supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0122603), National Natural Science Foundation of China (Grant No. 62476016 and 62441617), the Fundamental Research Funds for the Central Universities, Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing.

References

1. Anthropic: Claude 3.5 sonnet model card addendum (Jun 2024), https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Chen, J., Lv, Z., Wu, S., Lin, K.Q., Song, C., Gao, D., Liu, J.W., Gao, Z., Mao, D., Shou, M.Z.: 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 (2024)
5. Chen, Y., Bai, X., Wang, Z., Bai, C., Dai, Y., Lu, M., Zhang, S.: Streamkv: Streaming video question-answering with segment-based kv cache retrieval and compression. arXiv preprint arXiv:2511.07278 (2025)

6. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences **67**(12), 220101 (2024)
7. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
8. Di, S., Yu, Z., Zhang, G., Li, H., Zhong, T., Cheng, H., Li, B., He, W., Shu, F., Jiang, H.: Streaming video question-answering with in-context video kv-cache retrieval. arXiv preprint arXiv:2503.00540 (2025)
9. Fei, J., Li, D., Deng, Z., Wang, Z., Liu, G., Wang, H.: Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. arXiv preprint arXiv:2408.14023 (2024)
10. Guan, Y., Yin, L., Liang, D., Ju, J., Luo, Z., Luan, J., Liu, Y., Bai, X.: Video streaming thinking: Videollms can watch and think simultaneously. arXiv preprint arXiv:2603.12262 (2026)
11. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
12. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
13. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the 29th symposium on operating systems principles. pp. 611–626 (2023)
14. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
15. Li, J., Wei, K., Xu, Z., Su, Z., Yang, X., Deng, C.: Perceive, reflect and understand long video: Progressive multi-granular clue exploration with interactive agents. arXiv preprint arXiv:2509.24943 (2025)
16. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: 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 (2024)
17. Lin, J., Wu, J., Sun, X., Wang, Z., Liu, J., Su, Y., Yu, X., Chen, H., Luo, J., Liu, Z., et al.: Unleashing hour-scale video training for long video-language understanding. arXiv preprint arXiv:2506.05332 (2025)
18. Lin, J., Fang, Z., Chen, C., Wan, Z., Luo, F., Li, P., Liu, Y., Sun, M.: Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. arXiv preprint arXiv:2411.03628 (2024)
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
20. Liu, J., Wang, Y., Ma, H., Wu, X., Ma, X., Wei, X., Jiao, J., Wu, E., Hu, J.: Kangaroo: A powerful video-language model supporting long-context video input. arXiv preprint arXiv:2408.15542 (2024)
21. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)

22. Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., et al.: Nvila: Efficient frontier visual language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4122–4134 (2025)
23. Long, L., He, Y., Ye, W., Pan, Y., Lin, Y., Li, H., Zhao, J., Li, W.: Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. arXiv preprint arXiv:2508.09736 (2025)
24. Ma, Z., Gou, C., Shi, H., Sun, B., Li, S., Rezatofighi, H., Cai, J.: Drvideo: Document retrieval based long video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18936–18946 (2025)
25. Niu, J., Li, Y., Miao, Z., Ge, C., Zhou, Y., He, Q., Dong, X., Duan, H., Ding, S., Qian, R., et al.: Ovo-bench: How far is your video-llms from real-world online video understanding? In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18902–18913 (2025)
26. Qian, R., Ding, S., Dong, X., Zhang, P., Zang, Y., Cao, Y., Lin, D., Wang, J.: 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)
27. Qian, R., Dong, X., Zhang, P., Zang, Y., Ding, S., Lin, D., Wang, J.: Streaming long video understanding with large language models. Advances in Neural Information Processing Systems **37**, 119336–119360 (2024)
28. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., et al.: Longvu: Spatiotemporal adaptive compression for long video-language understanding. arXiv preprint arXiv:2410.17434 (2024)
29. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., 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)
30. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
31. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
32. Wang, X., Zhang, Y., Zohar, O., Yeung-Levy, S.: Videoagent: Long-form video understanding with large language model as agent. In: European Conference on Computer Vision. pp. 58–76. Springer (2024)
33. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr: Contexts optical compression. arXiv preprint arXiv:2510.18234 (2025)
34. Wu, Y., Hu, X., Sun, Y., Zhou, Y., Zhu, W., Rao, F., Schiele, B., Yang, X.: Number it: Temporal grounding videos like flipping manga. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13754–13765 (2025)
35. Xu, R., Xiao, G., Chen, Y., He, L., Peng, K., Lu, Y., Han, S.: Streamingvlm: Real-time understanding for infinite video streams. arXiv preprint arXiv:2510.09608 (2025)
36. Yang, H., Tang, F., Zhao, L., An, X., Hu, M., Li, H., Zhuang, X., Lu, Y., Zhang, X., Swikir, A., et al.: Streamagent: Towards anticipatory agents for streaming video understanding. arXiv preprint arXiv:2508.01875 (2025)
37. Yang, Z., Wang, S., Zhang, K., Wu, K., Leng, S., Zhang, Y., Li, B., Qin, C., Lu, S., Li, X., et al.: Longvt: Incentivizing" thinking with long videos" via native tool calling. arXiv preprint arXiv:2511.20785 (2025)

38. Yao, L., Li, Y., Wei, Y., Li, L., Ren, S., Liu, Y., Ouyang, K., Wang, L., Li, S., Li, S., 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)
39. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The eleventh international conference on learning representations (2022)
40. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
41. Ye, S., Ouyang, B., Qian, T., Zeng, L., Yuan, M., Chu, X., Hong, W., Chen, X.: Venus: An efficient edge memory-and-retrieval system for vlm-based online video understanding. arXiv preprint arXiv:2512.07344 (2025)
42. Zeng, X., Qiu, K., Zhang, Q., Li, X., Wang, J., Li, J., Yan, Z., Tian, K., Tian, M., Zhao, X., et al.: Streamforest: Efficient online video understanding with persistent event memory. arXiv preprint arXiv:2509.24871 (2025)
43. Zhang, H., Wang, Y., Tang, Y., Liu, Y., Feng, J., Dai, J., Jin, X.: Flash-vstream: Memory-based real-time understanding for long video streams. arXiv preprint arXiv:2406.08085 (2024)
44. Zhang, H., Wang, Y., Tang, Y., Liu, Y., Feng, J., Jin, X.: Flash-vstream: Efficient real-time understanding for long video streams. arXiv preprint arXiv:2506.23825 (2025)
45. Zhang, H., Yang, S., Fu, J., Ng, S.K., Qiu, X.: Hermes: Kv cache as hierarchical memory for efficient streaming video understanding. arXiv preprint arXiv:2601.14724 (2026)
46. Zhang, P., Zhang, K., Li, B., Zeng, G., Yang, J., Zhang, Y., Wang, Z., Tan, H., Li, C., Liu, Z.: Long context transfer from language to vision. arXiv preprint arXiv:2406.16852 (2024)
47. Zhang, X., Wen, S., Wu, W., Huang, L.: Tinyllava-video-r1: Towards smaller lmms for video reasoning. arXiv preprint arXiv:2504.09641 (2025)
48. Zhang, X., Weng, X., Yue, Y., Fan, Z., Wu, W., Huang, L.: Tinyllava-video: Towards smaller lmms for video understanding with group resampler. arXiv preprint arXiv:2501.15513 (2025)
49. Zhang, Y., Li, B., Liu, h., Lee, Y.j., Gui, L., Fu, D., Feng, J., Liu, Z., Li, C.: Llava-next: A strong zero-shot video understanding model (April 2024), `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`
50. Zheng, N., Huang, J., Guo, Q., Zhao, F.: Videoscaffold: Elastic-scale visual hierarchies for streaming video understanding in mllms. arXiv preprint arXiv:2512.22226 (2025)
51. Zhi, Z., Wu, Q., Li, W., Li, Y., Shao, K., Zhou, K., et al.: Videoagent2: Enhancing the llm-based agent system for long-form video understanding by uncertainty-aware cot. arXiv preprint arXiv:2504.04471 (2025)
52. Zuo, J., Deng, Y., Kong, L., Yang, J., Jin, R., Zhang, Y., Sang, N., Pan, L., Liu, Z., Gao, C.: Videolucy: Deep memory backtracking for long video understanding. arXiv preprint arXiv:2510.12422 (2025)