CurveStream: Boosting Streaming Video Understanding in MLLMs via Curvature-Aware Hierarchical Visual Memory Management

Chao Wang[1*], Xudong Tan[1*], Jianjian Cao[1], Kangcong Li[1], and Tao Chen[1,2†]

[1] College of Future Information Technology, Fudan University, Shanghai, China
chaowang25@m.fudan.edu.cn, eetchen@fudan.edu.cn
[2] Shanghai Innovation Institute, Shanghai, China

Abstract. Multimodal Large Language Models have achieved significant success in offline video understanding, yet their application to streaming videos is severely limited by the linear explosion of visual tokens, which often leads to Out-of-Memory (OOM) errors or catastrophic forgetting. Existing visual retention and memory management methods typically rely on uniform sampling, low-level physical metrics, or passive cache eviction. However, these strategies often lack intrinsic semantic awareness, potentially disrupting contextual coherence and blurring transient yet critical semantic transitions. To address these limitations, we propose CurveStream, a training-free, curvature-aware hierarchical visual memory management framework. Our approach is motivated by the key observation that high-curvature regions along continuous feature trajectories closely align with critical global semantic transitions. Based on this geometric insight, CurveStream evaluates real-time semantic intensity via a Curvature Score and integrates an online K-Sigma dynamic threshold to adaptively route frames into clear and blurred memory states under a strict token budget. Evaluations across diverse temporal scales confirm that this lightweight framework, CurveStream, consistently yields absolute performance gains of over 10% (e.g., 10.69% on StreamingBench and 13.58% on OVOBench) over respective baselines, establishing new state-of-the-art results for streaming video perception.

Keywords: Streaming Video Understanding · Visual Memory Management · Training-free · Feature Manifold Curvature

1 Introduction

While Multimodal Large Language Models (MLLMs) have achieved remarkable success in offline video understanding [1, 2, 20, 22, 34, 48], their application to streaming video scenarios is still hindered by fundamental bottlenecks. Streaming videos are theoretically infinite in length, inevitably leading to a linear explosion

[*] Equal contribution. [†] Corresponding author.

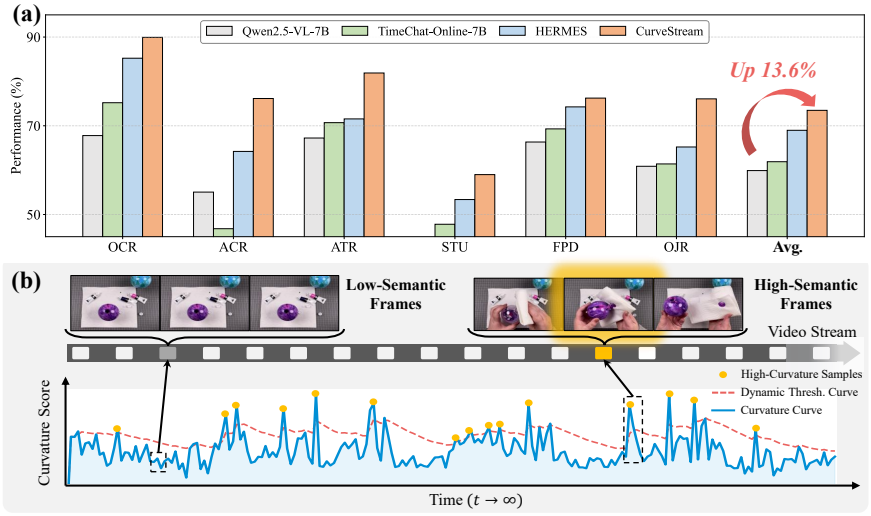


Fig. 1: Performance and mechanism of CurveStream. (a) CurveStream achieves state-of-the-art on OVOBench among training-free paradigms, boosting performance by 13.6% over the Qwen2.5-VL-7B baseline. **(b)** Curvature-aware memory management over infinite streams ($t \to \infty$). By evaluating real-time semantic intensity (blue curve) against a K-Sigma dynamic threshold (pink dashed line), it adaptively filters redundant Low-Semantic Frames. Critical High-Semantic Frames (yellow dots) at curvature peaks are preserved, ensuring optimal visual context retention under strict token limits.

of visual tokens. Under stringent GPU memory constraints, models are highly susceptible to Out-of-Memory (OOM) errors or suffer from catastrophic forgetting caused by naive truncation strategies [37]. Consequently, continuously and dynamically managing visual memory within a fixed memory budget emerges as the core challenge in achieving long-term streaming video understanding.

To address the challenge of linear token explosion, existing methods primarily focus on two aspects: visual information retention and long-term memory management. Visual information retention strategies typically utilize uniform sampling [13, 29, 41] or low-level difference metrics (including inter-frame similarity [19,35] or optical flow [36]). However, these approaches are often sensitive to local noise and prioritize low-level physical motion, making it difficult to robustly capture the high-level global semantic transitions required for multimodal reasoning. Building upon these retained visual features, long-term memory management mechanisms further process the context. Mainstream solutions predominantly include rule-based cache eviction [16, 37, 39, 46], feature clustering and merging, and retrieval paradigms utilizing external storage [9].

Despite their progress, these visual retention and memory management methods share common limitations that hinder efficient streaming video understanding: 1) Semantic Fragmentation: They mostly employ passive eviction or smooth-

ing compression strategies lacking intrinsic semantic awareness, which disrupts contextual coherence. 2) Information Blurring: During indiscriminate feature compression, they irreversibly blur transient yet critical semantic transition points. 3) Delayed Perception: Retrieval mechanisms conditioned on post-hoc queries restrict the model's capability for real-time, proactive perception in unbounded streaming scenarios.

To overcome these limitations, we re-examine the evolutionary dynamics of video streams within the feature space. We observe a critical phenomenon: when mapping a continuous video stream into a trajectory within the feature space, the high-curvature regions along this trajectory precisely correspond to high-quality visual semantic transitions. Unlike uniform sampling or physical motion metrics that treat frames equally or focus on local noise, curvature geometrically measures the intensity of semantic shifts. A sharp turn (high curvature) in the feature trajectory signifies the emergence of a new event, a sudden viewpoint change, or a critical action boundary. This implies that utilizing "curvature" as an evaluation metric enables the precise extraction of the most valuable contextual information for reasoning, thereby offering a novel perspective for constructing highly efficient, adaptive streaming video memory management systems. As illustrated in Fig. 1 (b), this geometric approach effectively identifies critical semantic transitions by monitoring the trajectory's curvature peaks.

Building upon this curvature observation, we propose CurveStream, a training-free, curvature-aware hierarchical visual memory management framework. Diverging from uniform sampling strategies that periodically drop frames, we formulate streaming video processing as a dynamic, semantic-aware memory update process under a fixed token capacity limit (N). Specifically, CurveStream first calculates a Curvature Score in real time to represent the intensity of semantic transitions, integrating motion variation of consecutive frames with the geometric angle between feature displacement vectors. To achieve adaptive memory management in non-stationary video streams, we introduce an online-updating K-Sigma rule ($g = \mu + k\sigma$). This mechanism dynamically generates an admission threshold based on the running mean and variance of the historical curvature, adaptively categorizing high-value visual tokens into distinct hierarchical states (Clear Memory and Blurred Memory). When the memory bank reaches its capacity limit, the system systematically evicts the oldest tokens following strict queue rules. This design ensures that models maintain an acute perception of core visual semantic trajectories under a constant memory footprint.

To comprehensively evaluate CurveStream, we conduct extensive experiments across diverse temporal scales, encompassing 10 Real-Time Visual Understanding tasks in StreamingBench [23], 6 Real-Time Visual Perception tasks in OVOBench [27], and 3 offline video datasets (15–1200s) [10, 21, 26]. As a lightweight, model-agnostic module, CurveStream demonstrates broad architectural compatibility across the LLaVA-OneVision and Qwen-VL (2/2.5/3) series at 4B, 7B, 8B, and 32B parameter scales. As shown in Fig. 1a, integrating our framework into the Qwen2.5-VL-7B baseline yields accuracies of 84.00% and 73.48% on StreamingBench and OVOBench, respectively, delivering absolute

performance gains of 10.69% and 13.58%. Furthermore, CurveStream enables 7B-parameter open-source models to consistently surpass closed-source commercial systems, including GPT-4o and Gemini 1.5 Pro, validating its robust generalizability and practical efficacy.

In summary, the main contributions of this paper are as follows:

1. Revealing the "curvature" effect in streaming videos. We discover that high-curvature regions in the latent feature space align with critical global semantic transitions, providing a geometric metric for evaluating visual information that overcomes local noise.
2. Proposing CurveStream, a training-free hierarchical memory management framework. By integrating real-time curvature scoring with a dynamic K-Sigma threshold, it adaptively routes frames into clear and blurred memory states to handle non-stationary streams under fixed token budgets.
3. Achieving state-of-the-art performance on streaming benchmarks. CurveStream effectively mitigates OOM issues and consistently improves diverse MLLMs by approximately 10% in streaming scenarios, showing broad applicability on benchmarks like StreamingBench and OVOBench.

2 Related Work

2.1 Existing Visual Information Retention Strategies

Existing strategies for visual information retention in long videos encompass various directions, with prominent approaches focusing on rule-based token compression and query-driven feature retrieval [17, 20, 31, 49]. Rule-based methods mitigate redundancy by evaluating local feature similarities. AKS [29] and M-LLM [13] employ adaptive keyframe selection algorithms to maximize video coverage. FLoC [8], FlexSelect [25], and METok [33] dynamically prune redundant tokens during inference utilizing attention weights or facility location functions. Query-driven approaches perform goal-oriented extraction by fetching relevant frames conditioned on user instructions. DIG [18], APVR [11], BOLT [24], and MemVid [43] compute semantic similarities between post-hoc text queries and visual frames. These paradigms generally rely on delayed user queries or low-level physical metrics (including inter-frame cosine similarity). This makes them susceptible to local motion noise in dynamic scenes and limits their capacity for proactive perception. To address this, our method diverges from traditional metrics by leveraging the "curvature" of feature trajectories in the feature space. This perspective intrinsically captures global semantic transitions, ensuring robust retention that is resilient to local physical disturbances.

2.2 Existing Streaming Video Memory Management Mechanisms

Processing theoretically infinite streaming videos inherently causes a linear explosion in memory footprint. Existing methods mainly address this issue through

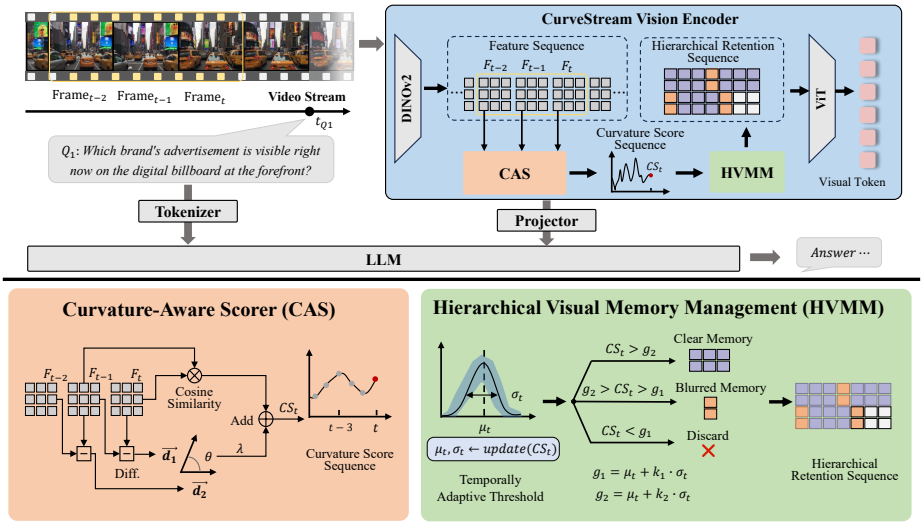


Fig. 2: Overview of the CurveStream framework. This training-free vision encoder enables infinite streaming video understanding by replacing traditional sampling with a dynamic-retention perception layer designed to prevent Out-of-Memory (OOM) errors in long-term sequences. The Curvature-Aware Scorer (CAS) evaluates semantic transition intensity by fusing first-order motion variation and second-order trajectory curvature within the latent feature manifold, while the Hierarchical Visual Memory Management (HVMM) module dynamically routes incoming tokens into a fixed-capacity ($N_{\max}$) queue. By utilizing temporally adaptive K-Sigma thresholds, the encoder adaptively categorizes visual information into Clear, Blurred, or Discard states based on the intensity of semantic shifts, thereby ensuring a constant memory footprint while preserving critical visual anchors for long-term multimodal reasoning.

KV cache eviction and external structured memory [5, 9, 45]. KV cache eviction methods reduce memory cost by discarding or compressing historical tokens according to sliding-window constraints, spatio-temporal redundancy, or token-importance criteria [4,16,37,46]. External memory methods instead offload long-term visual context into hierarchical trees, feature banks, or retrieval-based storage, and reactivate relevant information when needed [9, 42, 44, 50]. More recent approaches further introduce trained reasoning or prediction signals for memory update, such as textual reasoning, prediction-error-based surprise, or latent future-event modeling [12, 15, 32, 38]. However, these mechanisms either rely on additional trained reasoning or predictive modules, or treat memory management as queue-based smoothing or isolated retrieval. Consequently, they may blur transient semantic shifts and disrupt natural in-context coherence. In contrast, CurveStream formulates memory management as a dynamic, semantic-aware in-context update process: its online K-Sigma rule continuously evaluates historical curvature and adaptively updates clear and blurred memory within a strict token limit.

3 Methods

To achieve precise understanding of infinitely long streaming videos under strict memory constraints, we propose CurveStream, a training-free vision encoder architecture (illustrated in Fig. 2). The framework operates as an online selective-retention pipeline: it first utilizes a Curvature-Aware Scorer (CAS) to extract semantic transition intensity from the latent feature manifold trajectory, which is then processed by a Hierarchical Visual Memory Management (HVMM) module. Guided by temporally adaptive thresholds derived from online manifold statistics, this mechanism dynamically routes incoming frames into a fixed-capacity memory bank, categorizing them as Clear, Blurred, or Discarded.

3.1 Problem Formulation

Let $\mathcal{V} = \{I_t\}_{t=1}^{\infty}$ be an infinitely long, continuous video stream, where I_t denotes the visual observation at time step t. Suppose the system receives a natural language query Q regarding the current or historical states at timestamp t_q. Due to the large parameter size Θ of Multimodal Large Language Models (MLLMs) and the quadratic complexity of self-attention mechanisms, it is computationally intractable to directly feed the entire historical sequence $\mathcal{V}_{\leq t_q}$ into the model. Therefore, the system must maintain a dynamic visual memory queue $\mathcal{M}_t$ restricted by a maximum capacity limit $N_{\max}$.

We frame the streaming video understanding task as an online information extraction problem within a constrained space. At each time step t, the system needs to derive an efficient memory scheduling policy π. This policy evaluates the informative value of the current frame I_t and outputs a state tuple (s_t, r_t) containing the retention and resolution decisions to update the memory bank:

$$\mathcal{M}_t = \text{Update}(\mathcal{M}_{t-1}, I_t, s_t, r_t) \tag{1}$$

where $s_t \in \{\text{Clear}, \text{Blurred}, \text{Discard}\}$ represents the hierarchical routing state, and r_t denotes the corresponding spatial resolution.

The primary optimization objective of CurveStream is to maximize the conditional probability of the MLLM generating the correct answer A under a strict queue length constraint ($|\mathcal{M}_t| \leq N_{\max}$):

$$\max_{\pi} P(A \mid Q, \mathcal{M}_{t_q}; \Theta) \tag{2}$$

To solve this online decision-making problem lacking direct supervisory signals, we leverage the intrinsic geometric properties of the visual feature manifold to construct a lightweight scheduling policy π, realized through the CAS and HVMM modules described below.

3.2 Curvature-Aware Scorer (CAS)

In continuous visual streams, adjacent frames often exhibit high temporal redundancy. Especially in embodied AI or first-person perspectives, traditional sampling strategies based on simple feature differences are highly prone to overfitting

to large translational motions. To accurately localize high-value information, we design the Curvature-Aware Scorer (CAS).

CAS utilizes a frozen visual encoder to extract the global feature representation $\mathbf{F}_t \in \mathbb{R}^D$ of the input frame I_t, followed by L_2 normalization. To characterize the evolutionary trajectory of features within the latent space manifold, we integrate both the first-order motion intensity and the second-order geometric curvature. Based on the cosine similarity between consecutive frames, the first-order Motion Variation is defined as:

$$M_t = 1 - \frac{\langle \mathbf{F}_t, \mathbf{F}_{t-1} \rangle}{\|\mathbf{F}_t\| \|\mathbf{F}_{t-1}\|} \tag{3}$$

To filter out constant-velocity background changes caused by smooth camera movements, we compute an approximation of the second-order partial derivative of the feature trajectory. Let the feature displacement vectors of adjacent time steps be $\mathbf{d}_1 = \mathbf{F}_{t-1} - \mathbf{F}_{t-2}$ and $\mathbf{d}_2 = \mathbf{F}_t - \mathbf{F}_{t-1}$. The local Geometric Curvature of the feature manifold is approximately represented by the angular deviation between these displacement vectors:

$$C_t = 1 - \frac{\langle \mathbf{d}_1, \mathbf{d}_2 \rangle}{\|\mathbf{d}_1\| \|\mathbf{d}_2\|} \tag{4}$$

When $\mathbf{d}_1$ and $\mathbf{d}_2$ are aligned in direction, C_t approaches 0, indicating a smooth transition period. Conversely, when the direction of feature evolution changes abruptly (e.g., a new entity intrudes or a sharp viewpoint shift occurs), C_t increases significantly. The final Curvature Score CS_t is formulated as a linear combination of the two:

$$CS_t = M_t + \lambda C_t \tag{5}$$

where λ serves as the balancing coefficient for the geometric penalty term.

3.3 Hierarchical Visual Memory Management (HVMM)

After obtaining the CS_t sequence, the Hierarchical Visual Memory Management (HVMM) module utilizes temporally adaptive dynamic thresholds to route high-value frames into a fixed-capacity memory bank at differentiated resolution levels, effectively suppressing KV Cache bloat.

Online Manifold Distribution Estimation In untrimmed embodied or first-person streaming videos, the temporal pacing typically exhibits significant dynamics. For instance, a subject might suddenly break into a vigorous run after a prolonged period of stationary observation. Under such complex scenarios, employing any a priori static threshold is highly likely to lead to memory bank collapse or severe loss of critical information.

Therefore, HVMM models the filtering of high-value information as an online distribution-aware process. To capture the dynamic pacing of the video stream

in real time, we update the transient expectation μ_t and variance σ_t^2 of the curvature scores using an Exponential Moving Average (EMA) formulation:

$$\mu_t = \gamma\mu_{t-1} + (1-\gamma)CS_t, \quad \sigma_t^2 = \gamma\sigma_{t-1}^2 + (1-\gamma)(CS_t - \mu_t)^2 \tag{6}$$

where $\gamma \in (0,1)$ is the momentum factor controlling the size of the historical observation window. As the time step t advances, the newly observed curvature score CS_t smoothly calibrates the transient distribution parameters in a recursive manner. Based on this online evolutionary mechanism, we construct Gaussian distribution-aware dynamic dual thresholds: $g_1 = \mu_t + k_1\sigma_t$ and $g_2 = \mu_t + k_2\sigma_t$ $(k_1 < k_2)$. This design enables CurveStream to adaptively scale its sensitivity to visual shifts according to the current intensity of the scene.

Hierarchical State Transition Guided by the adaptive dual thresholds, HVMM executes a resolution-aware hierarchical state transition strategy. Specifically, the retention state s_t for an incoming frame I_t is dynamically determined as follows:

$$s_t = \begin{cases} \text{Clear Memory}, & \text{if } CS_t \geq g_2 \\ \text{Blurred Memory}, & \text{if } g_1 \leq CS_t < g_2 \\ \text{Discard}, & \text{if } CS_t < g_1 \end{cases} \tag{7}$$

Clear Memory. Frames satisfying $CS_t \geq g_2$ break through the current local dynamic distribution and capture significant semantic shifts. The system retains their original high-resolution features ($r_t = \text{High}$) and stores them in the memory bank to support subsequent fine-grained visual reasoning. Notably, the current frame I_{t_q} that triggers the query is deterministically assigned this state to ensure immediate context awareness.

Blurred Memory. Frames falling within $g_1 \leq CS_t < g_2$ are identified as intermediate transitional observations consistent with the current dynamic pacing. To preserve necessary temporal causal associations and action coherence while significantly compressing token overhead, these frames are downsampled to a minimal resolution ($r_t = \text{Low}$) before storage.

Discard. Frames with $CS_t < g_1$ represent low-information redundant observations below the local expected mean. The system directly discards these features to protect the scarce memory space.

Finally, to ensure a constant memory footprint without OOM risks, whenever the memory bank exceeds its capacity ($|\mathcal{M}_t| > N_{max}$), the system executes a strict First-In-First-Out (FIFO) eviction, removing the oldest tokens from the queue regardless of their retention states.

4 Experiments

4.1 Experimental Setup

Datasets. To comprehensively evaluate the effectiveness of the proposed adaptive visual memory framework under various temporal dynamics, we conducted

extensive experiments across five mainstream multimodal benchmarks encompassing three video paradigms. As the core of our evaluation for streaming video understanding, we selected StreamingBench [23] and OVOBench [27]. These two benchmarks rigorously test the model's capability for long-range event association and instantaneous dynamic response within continuous data streams. To address complex dynamic scenes, we utilized EgoSchema [26], a highly challenging egocentric benchmark that rigorously tests the model's ability to accurately capture micro-actions and perform causal reasoning amidst drastic viewpoint changes and redundant backgrounds. Furthermore, to explore the extreme limits of memory capacity, we introduced VideoMME [10], comprehensively examining the model's feature retention and generalizability across short, medium, and extremely long (up to several hours) contexts. Finally, we incorporated the MVBench [21] short video benchmark to verify that the system's dynamic frame filtering and resolution reduction strategies do not compromise the model's spatio-temporal perception of fine-grained local actions.

Baselines. Our comparative analysis involves two major categories of baseline methods. The first category comprises state-of-the-art open-source Multimodal Large Language Models (Base MLLMs), specifically including LLaVA-OneVision [17] and multiple iterations of the Qwen-VL series (i.e., Qwen2-VL [34], Qwen2.5-VL [2], Qwen3-VL [1]). The second category encompasses recent advanced frameworks specifically optimized for streaming video understanding or long-context visual processing (SOTA Streaming Methods), including Flash-VStream [45], FreshMem [19], HERMES [46], and ReKV [9]. By integrating our proposed training-free memory module into the base MLLMs, we conduct a direct performance comparison with these specialized SOTA methods under strictly equivalent visual token constraints.

Implementation Details. In all comparative experiments, to ensure evaluation fairness and strictly simulate the physical GPU memory constraints inherent in streaming video processing, we establish a uniform memory bank capacity upper limit (i.e., a maximum token budget N) across all methods. At the feature extraction frontend of our framework, we employ the lightweight DINOv2-small model to acquire local geometric representations of temporal features. During the adaptive memory allocation phase, for high-curvature core transition frames that trigger clear memory, the system retains the native dynamic high-resolution input of the base model. Conversely, for blurred memory frames representing smooth transition states, the resolution is uniformly downsampled to a fixed 224×224 to conserve memory space. All benchmark evaluations are independently executed on a single inference GPU to fully validate the robustness of our framework under severely limited memory conditions.

4.2 Online Benchmark Results

Tab. 1 presents the quantitative evaluation results of various methods on the streaming video benchmarks. Under strict visual token capacity constraints, our

method achieves stable and significant performance leaps across different base models. Specifically, when utilizing Qwen2-VL-7B as the base model, our method achieves accuracies of 81.04% and 70.73% on StreamingBench and OVOBench, respectively, yielding absolute performance gains of 12.0% and 10.08% compared to the uniform sampling baseline.

More importantly, among training-free streaming video understanding frameworks, our method establishes a new state-of-the-art (SOTA). Compared to recent advanced specialized streaming video methods (*e.g.*, FreshMem and HERMES), our framework further achieves absolute accuracy improvements of 6.84% and 4.06% on StreamingBench and OVOBench, respectively.

This comprehensively leading performance is directly attributed to our adaptive visual memory mechanism. By introducing manifold curvature as a dynamic prior, the framework not only effectively strips away redundant static backgrounds in long videos but also precisely allocates limited memory resources to high-frequency visual transition points. This strategy, which highly aligns the memory queue with the underlying dynamic evolution of the video, fundamentally overcomes catastrophic forgetting in long-range reasoning, thereby preserving the highest-quality temporal context for the model.

4.3 Offline Benchmark Results

Tab. 2 presents the evaluation results of our framework on the short-video benchmark (MVBench) and long-video benchmark (VideoMME). Although our adaptive memory framework is specifically designed for streaming video scenarios, it also exhibits strong generalization ability in conventional offline short- and long-video understanding tasks.

As can be observed, our method consistently brings stable performance improvements across different base models. For instance, when built upon Qwen2.5-VL-7B, our method achieves a 1.03% absolute gain (up to 66.03%) on MVBench, a fine-grained action-oriented short-video benchmark, compared with the uniform sampling baseline. Meanwhile, when integrated into LLaVA-OneVision-7B, our method also yields a 1.77% absolute improvement (up to 59.44%) on VideoMME, a comprehensive long-video benchmark.

It is worth noting that for Qwen2.5-VL-7B on VideoMME, there is a slight performance drop (from 64.52% to 62.97%). This is because, to maintain a strictly constant memory footprint without OOM risks over hours-long videos, the system inevitably trades off some fine-grained global details to preserve the most critical semantic transitions. These quantitative results sufficiently verify the universality and effectiveness of the proposed framework in offline settings.

4.4 Scalability Across Model Parameters

Fig. 3a presents the evaluation results of our framework across the Qwen3-VL series with different parameter scales. Taking StreamingBench and OVOBench as examples, after being integrated into the 4B, 8B, and 32B versions of Qwen3-VL, our method yields absolute performance improvements of 8.7%, 12.4%, and

Table 1: Quantitative comparison of the average accuracy across 10 Real-Time Visual Understanding sub-tasks in StreamingBench and 6 Real-Time Visual Perception sub-tasks in OVOBench. The best results are highlighted in bold, with absolute performance gains over the respective baselines indicated in red. Note that our frame count (10-20) reflects the dynamically changing size of the adaptive memory queue.

Method	Frame	StreamingBench [23]	OVOBench [27]
Human	-	91.46	93.20
Proprietary MLLMs			
Gemini 1.5 Pro [30]	1fps	75.69	69.32
GPT-4o [14]	64	73.28	64.46
Open-source Offline MLLMs			
Qwen2-VL-7B [34]	64	69.04	60.65
InternVL-V2-8B [6]	16	63.72	60.73
Open-Source Online MLLMs			
Flash-VStream-7B [45]	-	23.23	29.86
VideoLLM-online-8B [3]	2fps	35.99	20.79
Dispider-7B [28]	1fps	67.63	54.55
TimeChat-Online-7B [40]	1fps	75.36	61.90
StreamForest-7B [44]	1fps	77.26	61.20
Training-free Offline-to-Online Methods			
LLaVA-OneVision-7B [17]	64	71.34	63.06
+ ReKV [9]	0.5fps	69.22	57.33
+ HERMES [46]	1fps	73.23	66.34
+ **Ours**	10-20	**75.12** (↑ 3.78)	**70.57** (↑ 7.51)
Qwen2-VL-7B [34]	1fps	69.04	60.65
+ HERMES [46]	1fps	-	-
+ FreshMem [19]	1fps	74.20	66.67
+ **Ours**	10-20	**81.04** (↑ 12.00)	**70.73** (↑ 10.08)
Qwen2.5-VL-7B [2]	1fps	73.31	59.90
+ HERMES [46]	1fps	79.44	68.98
+ **Ours**	10-20	**84.00** (↑ 10.69)	**73.48** (↑ 13.58)
Qwen3-VL-8B [1]	1fps	73.20	70.10
+ **Ours**	10-20	**85.56** (↑ 12.36)	**80.76** (↑ 10.66)

11.5% on StreamingBench, respectively, compared to their corresponding uniform sampling baselines. Similarly, it achieves robust gains of 11.2%, 10.7%, and 10.6% on OVOBench.

These consistent quantitative improvements fully demonstrate that our curvature-aware adaptive memory mechanism does not overfit to models of a specific parameter volume. Instead, as a plug-and-play module, it maintains stable positive gains across multimodal base models ranging from small to large parameters, exhibiting exceptionally strong architectural universality and scalability.

Table 2: Quantitative comparison of the average accuracy on MVBench (20 subtasks), EgoSchema, and VideoMME benchmarks. The best results are highlighted in bold, with absolute performance gains over the respective baselines indicated in red.

Method	Frame	MVBench [21]	EgoSchema [26]	VideoMME [10]
Proprietary MLLMs				
GPT-4V [47]	1fps	43.7	55.6	60.7
GPT-4o [14]	64	64.6	72.2	77.2
Open-source Offline MLLMs				
LLaVA-NeXT-Video [48]	32	33.7	43.9	46.5
Qwen2-VL-7B [34]	64	67.0	66.70	69.0
VideoChat2 [21]	16	60.4	54.4	54.6
VideoLLaMA2 [7]	32	54.6	51.7	46.6
Open-Source Online MLLMs				
Dispider-7B [28]	1fps	-	55.60	57.20
TimeChat-Online-7B [40]	1fps	75.36	61.90	53.22
StreamForest-7B [44]	1fps	70.20	-	61.40
Training-free Offline-to-Online Methods				
Qwen2.5-VL-7B [2]	1fps	65.00	58.47	**64.52**
+ HERMES [46]	1fps	65.53	59.47	60.63
+ **Ours**	1fps	**66.03**(↑ 1.03)	**64.29**(↑ 5.82)	62.97

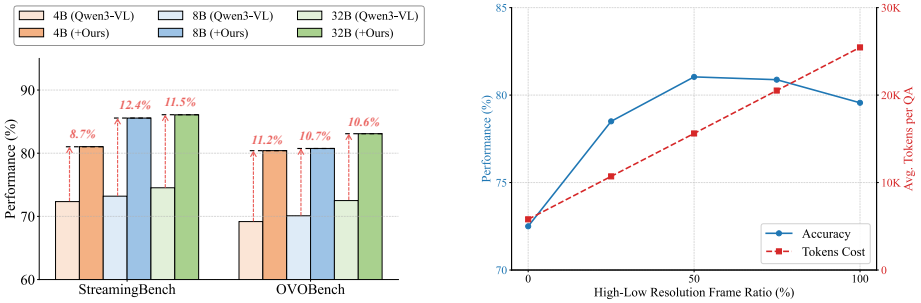
4.5 Ablation Studies

To validate the independent contributions and synergistic effects of the core components in our adaptive memory framework, we conduct systematic ablation analyses on the Qwen-based model.

Effectiveness of Curvature Metric. To evaluate the superiority of manifold curvature in capturing temporal information increments, we compare different frame sampling strategies under identical visual token constraints (see Table 3). The results demonstrate that our curvature metric significantly outperforms both uniform sampling and motion sampling based on cosine similarity. This confirms that pure motion similarity struggles to distinguish redundant smooth-panning shots from sudden semantic shifts. Furthermore, compared to dense optical flow, which is computationally expensive and highly susceptible to pixel noise, temporal manifold curvature serves as a lightweight second-order geometric prior,enabling more precise and robust localization of core turning points.

Adaptive Hierarchical Visual Memory Management. We further ablate the allocation ratio between clear memory (native high-resolution keyframes) and blurred memory (down-projected low-resolution transition frames) in the memory queue. As illustrated in Fig. 3b, forcing a 100% clear memory strategy accelerates context window depletion, triggering catastrophic forgetting of early memory. Conversely, adopting a 0% clear memory ("all-blur") strategy discards critical spatial details, leading to a drastic performance drop.

In contrast, the content-aware hybrid mechanism of our framework dynamically balances the clear memory ratio at approximately 50% based on temporal



(a) Scaling Performance Comparison **(b)** Impact of High-Res Ratio

Fig. 3: Scalability and memory allocation analysis. (a) CurveStream consistently delivers significant performance gains across varying model capacities (4B, 8B, 32B) of the Qwen3-VL series. (b) Impact of the clear memory (High-Res) retention ratio on overall accuracy and token cost. An adaptive 50% ratio achieves the optimal trade-off between semantic integrity and computational overhead.

Table 3: Comparison of frame sampling strategies on StreamingBench with the same selected-frame budget ($N = 10$). Training-free methods use Qwen2-VL-7B; StreamForest is reported with Qwen2-7B as a trained tree-memory reference.

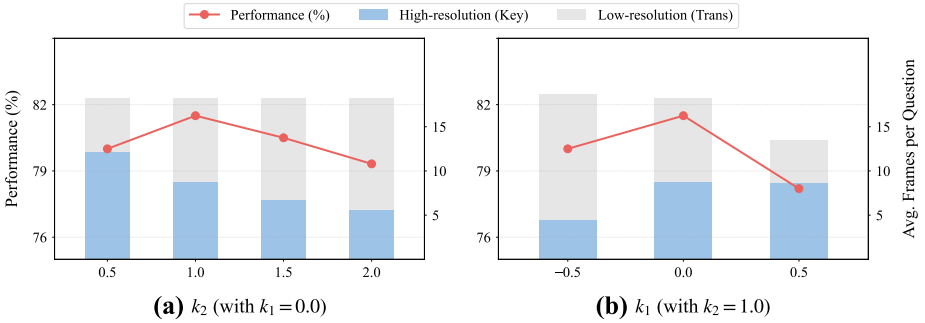
Sampling Strategy	Accuracy (%)
Uniform Sampling	69.04
Cosine Similarity	73.28
Optical Flow	46.54
Pyramid Optical Flow	75.69
StreamForest (train)	77.26
Ours (Curvature)	**77.31**

Table 4: Ablation on curvature score weights (λ) using Qwen2-VL-7B under identical token constraints on OVOBench. CurveStream maintains stable high performance across diverse weights, validating its plug-and-play reliability.

Method	λ	Accuracy (%)
Qwen2-VL-7B	-	60.65
Ours	0.2	**65.83**
Ours	0.4	62.50
Ours	0.6	63.33
Ours	0.8	62.50
Ours	1.0	65.00

dynamics. This approach achieves the best accuracy while substantially reducing computational overhead by about 40%. This indicates that, compared to static constraints, dynamically allocating clear and blurred memory more effectively strikes a balance between the integrity of long-term context and the capture of fine-grained actions. Specifically, due to the temporal non-stationarity of streaming videos, forcing a fixed high-resolution retention often overfits to local translational motion noise. In contrast, our adaptive 50% dynamic hybrid strategy essentially leverages localized blurred memory to serve as smooth transition states for action continuity, thereby freeing up the most critical clear memory space for high-curvature semantic transitions under the same token budget.

Hyperparameter Robustness. To verify the generalization stability of our framework, we evaluate the model's sensitivity to core hyperparameters. As



(a) k_2 (with $k_1 = 0.0$) **(b)** k_1 (with $k_2 = 1.0$)

Fig. 4: Ablation on K-Sigma dual thresholds. CurveStream exhibits strong hyperparameter robustness across various k_1 and k_2 configurations on OVOBench. The dynamic mechanism effectively balances memory allocation between High-Res and Low-Res frames, ensuring an optimal accuracy-efficiency trade-off without tedious tuning.

Table 5: Efficiency comparison under 1 FPS video ingestion.

Model	Video Length	E2E (ms/tok)	TTFT (s)
Base	5 / 10 / 15 min	45.92 / 54.19 / 63.28	1.503 / 3.550 / 5.606
Ours	5 / 10 / 15 min	43.70 / 42.31 / 44.69	1.011 / 0.791 / 1.234
Ratio (Base / Ours)		**1.05×** / **1.28×** / **1.42×**	**1.49×** / **4.49×** / **4.54×**

shown in Table 4, when the curvature comprehensive score weight λ varies across a broad range of $[0.2, 1.0]$, the model accuracy remains steadily above 62.5%, peaking at 65.83%. The maximum absolute fluctuation is merely 3.33%, and it consistently outperforms the baseline method. Similarly, the dual-threshold parameters, `K_SIGMA_TRANS` (k_1) and `K_SIGMA_KEY` (k_2), maintain highly stable performance and robust frame sampling ratios across different settings (see Fig. 4). Such exceptionally low hyperparameter sensitivity strongly corroborates the intrinsic robustness of our framework as a plug-and-play module, capable of adapting to diverse underlying data streams without tedious heuristic tuning for real-world streaming tasks.

4.6 Efficiency Evaluation

We further evaluate the practical efficiency of CurveStream under the same 1 FPS video ingestion setting. As shown in Fig. 5, the base model (e.g., Qwen-VL) exhibits rapid token and memory growth as video length increases, while CurveStream remains stable at around 10K–12K tokens and about 20GB memory. Table 5 further shows that CurveStream consistently reduces latency, achieving up to 1.42× E2E speedup and 4.54× TTFT speedup on 15-minute videos. These results verify its practical efficiency for long streaming inputs.

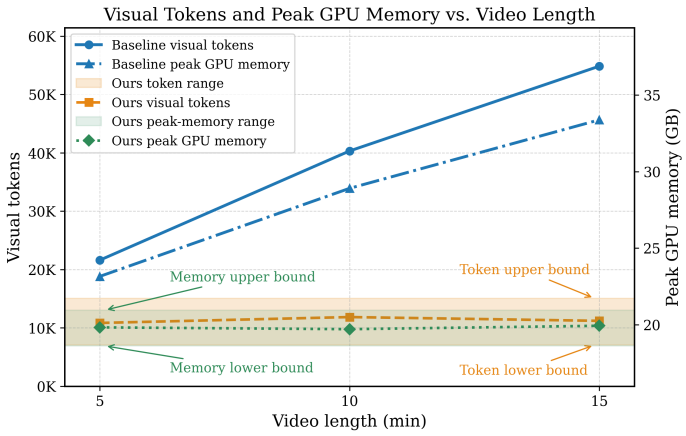


Fig. 5: Visual tokens and peak GPU memory versus video length. CurveStream keeps both metrics nearly stable as video length increases.

5 Conclusion

We present CurveStream, a training-free hierarchical memory management framework to boost streaming video understanding in MLLMs by tackling the inherent token explosion and Out-of-Memory (OOM) bottlenecks. Driven by the geometric insight that high-curvature regions in feature trajectories align with critical semantic transitions, CurveStream integrates a real-time Curvature Score with an online K-Sigma threshold. This dynamic mechanism adaptively routes incoming frames into clear or blurred memory states, ensuring MLLMs retain essential long-term visual context under strict token budgets.

Extensive experiments demonstrate that this lightweight, model-agnostic module exhibits broad architectural compatibility and consistently yields substantial performance gains over respective baselines. By establishing new state-of-the-art results on challenging benchmarks like StreamingBench and OVOBench, CurveStream offers a robust solution for continuous video perception. Future work will extend this geometric memory paradigm to broader embodied AI applications, such as autonomous navigation and prolonged robotic manipulation, where real-time adaptive reasoning and decision-making are paramount.

Acknowledgements

This work was supported by the National Key R&D Program of China (No. 2026YFE0101200) and the Shanghai Natural Science Foundation (No. 23ZR1402900). The computations in this research were performed on the CFFF platform of Fudan University.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. 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)
4. Chen, X., Tao, K., Shao, K., Wang, H.: Streamingtom: Streaming token compression for efficient video understanding. arXiv preprint arXiv:2510.18269 (2025)
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., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
7. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., Bing, L.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024), https://arxiv.org/abs/2406.07476
8. Cho, J., Lee, J., Hayat, M., Hwang, K., Porikli, F., Choi, S.: Floc: Facility location-based efficient visual token compression for long video understanding. arXiv preprint arXiv:2511.00141 (2025)
9. Di, S., Yu, Z., Zhang, G., Li, H., Cheng, H., Li, B., He, W., Shu, F., Jiang, H., et al.: Streaming video question-answering with in-context video kv-cache retrieval. In: ICLR (2025)
10. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24108–24118 (2025)
11. Gao, H., Bao, Y., Tu, X., Zhong, B., Yue, L., Zhang, M.: Apvr: Hour-level long video understanding with adaptive pivot visual information retrieval. arXiv preprint arXiv:2506.04953 (2025)
12. 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)
13. Hu, K., Gao, F., Nie, X., Zhou, P., Tran, S., Neiman, T., Wang, L., Shah, M., Hamid, R., Yin, B., et al.: M-llm based video frame selection for efficient video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13702–13712 (2025)
14. 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)

15. Jiang, T., Wu, L., Xia, S., Li, S., Yan, Z., Yang, H., Qiao, Y., Wang, Y.: Imagine before you predict: Interleaved latent visual reasoning for video event prediction. arXiv preprint arXiv:2606.05769 (2026)
16. Kim, M., Shim, K., Choi, J., Chang, S.: Infinipot-v: Memory-constrained KV cache compression for streaming video understanding. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=hFxOZjHyTg
17. 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)
18. Li, J., Li, B., Li, J., Lu, Y.: Divide, then ground: Adapting frame selection to query types for long-form video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11369–11380 (2026)
19. Li, K., Ye, P., Zhang, L., Wang, C., Qin, H., Chen, T.: Freshmem: Brain-inspired frequency-space hybrid memory for streaming video understanding. arXiv preprint arXiv:2602.01683 (2026)
20. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. Science China Information Sciences **68**(10), 200102 (2025)
21. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024)
22. 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)
23. 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)
24. Liu, S., Zhao, C., Xu, T., Ghanem, B.: Bolt: Boost large vision-language model without training for long-form video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3318–3327 (2025)
25. Lu, Y., Wang, T., Rao, F., Yang, Y., Zhu, L., et al.: Flexselect: Flexible token selection for efficient long video understanding. Advances in Neural Information Processing Systems **38**, 102751–102777 (2026)
26. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems **36**, 46212–46244 (2023)
27. 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)
28. 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)
29. Tang, X., Qiu, J., Xie, L., Tian, Y., Jiao, J., Ye, Q.: Adaptive keyframe sampling for long video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29118–29128 (2025)

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 (2024), https://arxiv.org/abs/2403.05530
31. Tu, C., Zhang, L., Chen, P., Ye, P., Zeng, X., Cheng, W., Yu, G., Chen, T.: Favorbench: A comprehensive benchmark for fine-grained video motion understanding. arXiv preprint arXiv:2503.14935 (2025)
32. Wang, L., Jin, Z., Hao, Y., Chen, Y., Liu, K., Ao, Y., Zhao, J.: Think while watching: Online streaming segment-level memory for multi-turn video reasoning in multimodal large language models. arXiv preprint arXiv:2603.11896 (2026)
33. Wang, M., Chen, S., Kersting, K., Tresp, V., Ma, Y.: Metok: Multi-stage event-based token compression for efficient long video understanding. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 18881–18895 (2025)
34. 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)
35. Wang, Y., Song, Y., Xie, C., Liu, Y., Zheng, Z.: Videollamb: Long streaming video understanding with recurrent memory bridges. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24170–24181 (2025)
36. Xiong, H., Yang, Z., Yu, J., Zhuge, Y., Zhang, L., Zhu, J., Lu, H.: Streaming video understanding and multi-round interaction with memory-enhanced knowledge. arXiv preprint arXiv:2501.13468 (2025)
37. 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)
38. Yang, S., Yang, J., Huang, P., Brown II, E.L., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., et al.: Cambrian-s: Towards spatial supersensing in video. In: The Fourteenth International Conference on Learning Representations (2025)
39. Yang, Y., Zhao, Z., Shukla, S.N., Singh, A., Mishra, S.K., Zhang, L., Ren, M.: Streammem: Query-agnostic kv cache memory for streaming video understanding. arXiv preprint arXiv:2508.15717 (2025)
40. 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)
41. Ye, J., Wang, Z., Sun, H., Chandrasegaran, K., Durante, Z., Eyzaguirre, C., Bisk, Y., Niebles, J.C., Adeli, E., Fei-Fei, L., et al.: Re-thinking temporal search for long-form video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8579–8591 (2025)
42. 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)
43. Yuan, H., Liu, Z., Qin, M., Qian, H., Shu, Y., Dou, Z., Wen, J.R., Sebe, N.: Memory-enhanced retrieval augmentation for long video understanding. arXiv preprint arXiv:2503.09149 (2025)
44. 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)
45. Zhang, H., Wang, Y., Tang, Y., Liu, Y., Feng, J., Jin, X.: Flash-vstream: Efficient real-time understanding for long video streams. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21059–21069 (2025)

46. Zhang, H., Yang, S., Fu, J., Ng, S.K., Qiu, X.: Hermes: Kv cache as hierarchical memory for efficient streaming video understanding (2026), https://arxiv.org/abs/2601.14724
47. Zhang, X., Lu, Y., Wang, W., Yan, A., Yan, J., Qin, L., Wang, H., Yan, X., Wang, W.Y., Petzold, L.R.: Gpt-4v (ision) as a generalist evaluator for vision-language tasks. arXiv preprint arXiv:2311.01361 (2023)
48. 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/
49. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
50. 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)