LOOK LESS, REMEMBER MORE: TRAINING-FREE DYNAMIC ENCODING WITH HIERARCHICAL VISUAL MEMORY FOR STREAMING VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming video understanding requires models to continuously process evolving visual content and answer questions under limited computational budgets. However, visual tokens grow linearly with video length, despite substantial temporal redundancy. Existing methods mainly reduce token redundancy within the language model through query-aware pruning or merging, leaving redundancy at the visual encoding stage largely unaddressed and risking the removal of information that may be useful for future queries. We propose **StreamSieve**, a training-free and query-agnostic framework for efficient streaming video understanding. StreamSieve first employs refined dynamic encoding to adapt visual computation to inter-frame temporal novelty and further suppress redundant tokens through saliency-guided spatiotemporal refinement. It then maintains a diversity-preserving hierarchical memory that compactly retains informative historical visual content without query guidance. Extensive experiments on online and offline long-video benchmarks show that **StreamSieve** delivers competitive video understanding performance with lower memory consumption and latency.

1 INTRODUCTION

Video understanding has long been dominated by an offline paradigm (Bai et al., 2025a; Wang et al., 2025b; Hurst et al., 2024; Li et al., 2024), where models analyze recorded videos with full access to all frames. However, real-world applications such as intelligent surveillance (Feng et al., 2026; Yang et al., 2025c), embodied robotics (Zhang et al., 2024; Fan et al., 2025), autonomous driving (Tian et al., 2024; Wang et al., 2025a; Xie et al., 2025), and real-time assistants (Huang et al., 2025; Yang et al., 2026a) require streaming video understanding capability: observations arrive as a continuous stream, and the model must sequentially process frame in a causal manner to support real-time interaction, maintaining a continuously updated understanding so that questions can be answered at any moment.

Advances in Multimodal Large Language Models (MLLMs) have demonstrated impressive performance on tasks requiring holistic reasoning over video content (Bai et al., 2025a; Wang et al., 2025b; Hurst et al., 2024; Li et al., 2024). Approaches ranging from segment-level retrieval and memory-based summarization (Yin et al., 2026; Lian et al., 2026) to KV-cache compression (Xiao et al., 2026) have been proposed in the offline paradigm, failing in many real-world applications where video arrives as an endless stream. Recently, an increasing number of works have extended offline video LLMs into online paradigm (Chen et al., 2024a; Wang et al., 2025c; Yang et al., 2026a). The dominant recipe maintains a persistent KV cache that is appended as each frame is encoded, giving the language model access to the history without reprocessing (Qian et al., 2025; Zhang et al., 2025); retrieval over the growing cache keeps the context tractable (Di et al., 2025; Ning et al., 2025); and streaming-oriented instruction tuning adapts models to continuous, multi-turn interaction (Ding et al., 2025; Xia et al., 2026).

Despite the above progress, existing streaming Video LLMs inherit a fundamental inefficiency from their offline counterparts: they typically process every incoming frame through a deep, computationally heavy vison encoder and feed all tokens into the language model (Chen et al., 2024a; Wang et al., 2025c). This practice raises two dilemmas in the streaming setting: **(1)** A video stream exhibits

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

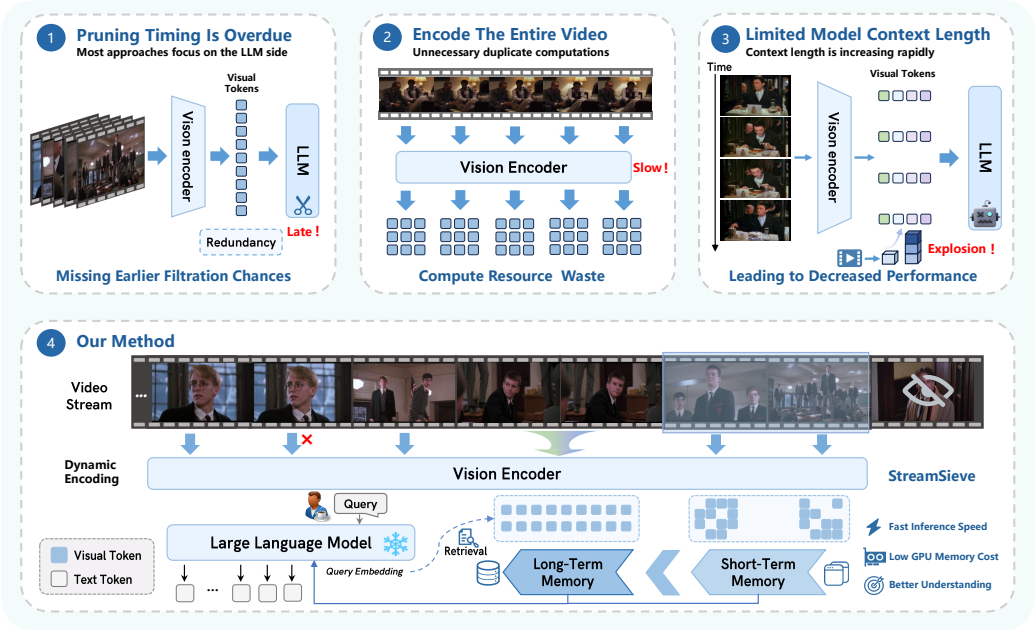


Figure 1: The comparison between StreamSieve and previous methods.

high temporal redundancy, such that adjacent frames are usually similar. Recent analysis shows that roughly 80% of visual tokens are naturally redundant (Yao et al., 2025). Therefore, fully encoding every frame creates a huge efficiency bottleneck. **(2)** Visual tokens of the sampled video frames accumulate rapidly as the video length increases, leading to a rapid expansion in context length. Most existing efficiency methods focus on language model-based context optimization, including token pruning (Chen et al., 2026c) and token compression (Wang et al., 2026b) while neglecting to filter out redundancy directly at the earlier visual encoding stage. Furthermore, in open-ended video stream or multi-turn interaction scenarios, the system does not know in advance which questions will be asked, and the user's intent, including the aspects and objects they focus on, also frequently changes. Query-guided pruning, which discards context information considered irrelevant to the current question, may be critical to the later question, deteriorating performance (Guo et al., 2026).

To address these limitations, we propose **StreamSieve**, a novel training-free framework that can dynamically and more compactly encode video frames, retain informative video content without query guidance, and requires no additional training and annotation. (1) We introduce **Refined Dynamic Encoding for Spacetime Redundancy (RDESR)**, which adaptively assigns incoming frames to different encoding levels according to their temporal novelty, thereby reducing redundant visual computation. The resulting frame tokens are further compressed through Spatiotemporal Saliency-guided Token Refinement, which removes redundant visual tokens while preserving informative content. (2) We further propose **Query-agnostic Diversity-preserving Hierarchical Memory (QDHPM)** to maintain a compact yet informative representation of the streaming history. QDHPM combines a short-term memory that preserves recent, directly accessible visual information with a Diversity-preserving Long-term Memory (DPLM) that compactly stores representative historical content. The long-term memory is updated based on information novelty, enabling the model to retain diverse historical evidence without relying on query-guided retrieval or compression. Figure 1 illustrates the the comparasion between **StreamSieve** and previous mainstream methods.

In summary, our key contributions are as follows:

- We propose **StreamSieve**, a novel training-free and query-agnostic framework that relies solely on the intrinsic semantics of a pretrained Video-LLM, achieving superior video understanding performance with limited memory usage and lower latency compared with SOTA methods over online and offline long-video benchmarks.

- To improve the efficiency of video representation and eliminate spatiotemporal redundancy at early stage, the Refined Dynamic Encoding for Spacetime Redundancy (RDESR) module is designed to adaptively encode video frames according to temporal novelty and further refine visual tokens by spatiotemporal saliency.
- To prevent excessive historical information from interfering with model reasoning, while avoiding information loss caused by query-guided compression, we propose the Query-agnostic Diversity-preserving Hierarchical Memory (QDPHM) to effectively manage contextual representation.

2 Related Work

Streaming Video Understanding. Streaming video understanding requires models to process continuous video streams in real-time, posing unique challenges in temporal reasoning and latency control (Tang et al., 2025; Yang et al., 2026b). To manage the ever-growing context, one line of work maintains a persistent KV Cache to avoid re-encoding the entire history (Zhang et al., 2025; Yang et al., 2025b), Beyond passive response, recent works attempt to equip the model with the ability to proactive interaction, enabling it to initiate responses whenever appropriate, without being explicitly asked. Some works introduce the EOS (End-of-Sequence) token prediction mechanism that decides when to respond (Chen et al., 2024a; Yu et al., 2026). Meanwhile, some propose to leverage auxiliary signals, such as lightweight prediction heads, to anticipate user intent and guide timely response (Wang et al., 2025c; 2026a; Fu et al., 2025b), alleviating the computational burden of token generation. More recently, LiveStar (Yang et al., 2026a) proposes a verification-based paradigm to monitor the semantic changes in the video stream.

Visual Token Compression and Pruning Existing methods can be regarded as optimizations in two dimensions: operational location and decision signal. Regarding operational location, outer-LLM pruning typically selects informative tokens or merges redundant tokens at the encoder output (Tao et al., 2025). Whereas inner-LLM pruning operates within the Transformer layers, such as attention-guided token pruning or merging to alleviate redundancy (Chen et al., 2024b; Guo et al., 2026). In terms of the decision signal, query-agnostic methods operate without question signals, relying on token similarity or density (Chen et al., 2026c; Shen et al., 2026; Shao et al., 2026). Query-aware methods leverage question tokens to guide pruning (Zhang et al., 2026b), achieving high compression for fixed VQA tasks but risking irreversible information loss when future queries shift. However, a shared limitation across both categories is that they operate on dense tokens already generated by the vision encoder, unable to suppress redundancy at the encoding stage itself.

Memory and Context Management As video streams progressively extend, the model context grows unbounded. Some studies (Yang et al., 2025b) explore KV cache compression to reduce the context stored within the LLM. StreamMem (Yang et al., 2025a) maintains a fixed budget cache through pruning low-scoring tokens for pruning and merging KV cache and InfiniPot-V (Kim et al., 2026) maintains a fixed memory budget by periodically compressing the KV cache. Memory strategies mitigate this issue by maintaining separate memory banks outside the LLM to mitigate. VideoLLaMB (Wang et al., 2025d) compresses clips into memory tokens and StreamForest (Zeng et al., 2026) organizes frames into hierarchical event trees based on temporal distance and content similarity. Complementing these, retrieval-augmented methods build external indices that enable on-demand access to historical content, retrieving relevant information when a query arrives. (Chen et al., 2026a;b), exhibiting solid performance on tasks requiring long-term historical information.

3 Method

3.1 Architectural Overview

As illustrated in Figure 2, the proposed StreamSieve consists of two main modules: the **Refined Dynamic Encoding for Spatiotemporal Redundancy** (RDESR) module and the **Query-agnostic Diversity-preserving Hierarchical Memory** (QDPHM) module. Streaming video frames are partitioned into different levels according to temporal novelty and conduct different encoding strategies. Video frame tokens are further compressed using Spatiotemporal Saliency-guided Token Refine-

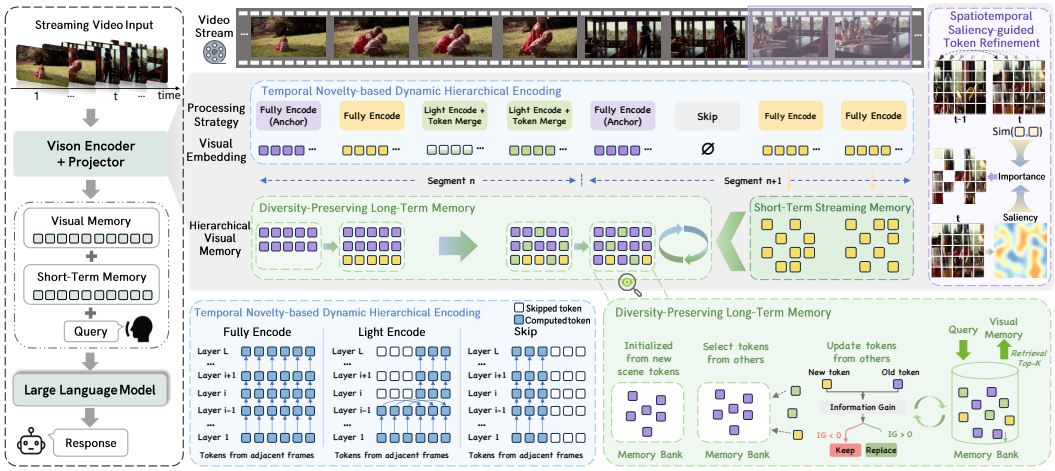


Figure 2: Overview of the StreamSieve framework. Overview of StreamSieve. Incoming video frames will be dynamically encoded by TNDHE and then further prune to preserve informative contents. The short-term streaming memory retains recent visual perception and the diversity-preserving long-term memory is used to long reasoning.

ment (SSTR), yielding the final visual tokens. The Query-agnostic Diversity-preserving Hierarchical Memory (QDPHM) utilizes the Short-term Stream Memory for preserving recent visual content at high fidelity, as well as the Diversity-preserving Long-term Memory (DPLM) for retaining sufficient historical information to support accurate understanding.

3.2 Refined Dynamic Encoding for Spatiotemporal Redundancy

Video redundancy is rarely an all-or-nothing property, as a frame may still contain subtle motion or newly emerging objects. Therefore, we introduce the Refined Dynamic Encoding for Spatiotemporal Redundancy module, which implements Temporal Novelty-based Dynamic Hierarchical Encoding (TNDHE) instead of applying a uniform full encoding or a hand-crafted sampling schedule. To avoid excessive computation of redundant information during encoding, similarity cues derived directly from the raw pixel space are leveraged to determine the encoding strategy.

Pixel-level Similarity. To quantify the dynamic variation of the video frames over time, the cosine similarity between consecutive video frames are exploited. Formally, the neighborhood similarity score of consecutive frames is calculated in a causal manner:

$$\mathcal{S}(f_i, f_{i-1}) = \frac{1}{2}\left(1 + \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} \mathrm{sim}\left(\boldsymbol{p}_{i-1}^{h,w}, \boldsymbol{p}_{i}^{h,w}\right)\right), \tag{1}$$

The cosine similarity score $\mathrm{sim}\left(\boldsymbol{p}_{i-1}^{h,w}, \boldsymbol{p}_{i}^{h,w}\right)$ is rescaled from its native range [-1,1] to [0,1].

Temporal Novelty-based Dynamic Hierarchical Encoding (TNDHE). The encoding strategy is dynamically determined according to the novelty of each video frame, which is estimated using neighborhood similarity score $s_i = \mathcal{S}(f_{i-1}, f_i)$ between adjacent frames. To implement this adaptive scheme, we partition the similarity range using three thresholds, $\tau_{\mathrm{vh}} > \tau_{\mathrm{h}} > \tau_{\mathrm{m}}$, and assign each frame f_i to one of four encoding levels:

$$c_i = \begin{cases} \mathrm{Duplicate}, & s_i \geq \tau_{\mathrm{vh}} \\ \mathrm{Similar}, & \tau_{\mathrm{h}} \leq s_i < \tau_{\mathrm{vh}} \\ \mathrm{New_info}, & \tau_{\mathrm{m}} \leq s_i < \tau_{\mathrm{h}} \\ \mathrm{New_scene}, & s_i < \tau_{\mathrm{m}} \end{cases} \tag{2}$$

These four levels correspond to progressively more intensive encoding strategies with increasing computational costs. (1) For Duplicate frames, which are nearly identical to their predecessors and

contain negligible new information, visual encoding is skipped entirely, thereby introducing no additional encoder FLOPs. (2) For $\mathrm{Similar}$ frames, only minor variations are observed relative to their preceding frames. Instead of passing each frame through the entire visual encoder, we process it only through the first L_s layers to obtain a shallow representation $\mathbf{h}_i^{L_s}$. Consecutive $\mathrm{Similar}$ frames are then progressively aggregated into an accumulated representation $\mathbf{h}_{\mathrm{accu},i}$, which is subsequently propagated through the remaining deeper layers of the encoder. To adapt the aggregation strength to the degree of inter-frame redundancy, we employ a similarity-weighted Exponential Moving Average (EMA): (3) Frames categorized as $\mathrm{New_info}$ exhibit noticeable changes while remaining within the same scene. Such frames are therefore processed by the full visual encoder to preserve newly introduced visual content. (4) Frames categorized as $\mathrm{New_scene}$ correspond to substantial visual changes and are likewise fully encoded. Low similarity scores typically indicate potential scene transitions, we therefore allocate the largest token budget to them in the subsequent stage.

Spatiotemporal Saliency-guided Token Refinement (SSTR). To further fine-grained redundancy with the frame, the Spatiotemporal Saliency-Guided Token Refinement are introduced. Based on the encoding level assigned to each frame, corresponding token retention budget are determined. Let n_i represents the number of tokens in frame f_i and c_i represent the encoding level calculated by Eq. (2). The number of tokens retained for frame i is then computed as: $b_i = \mathrm{round}(\varphi(c_i) \cdot n_i)$, where $\varphi(c_i)$ denotes the token retention ratio associated with encoding level c_i.

We combined two complementary criteria, temporal novelty and spatial saliency, to eliminate token level redundancy. Formally, let $\boldsymbol{v}_i^{h,w}$ indicate the token of frame f_i at spatial location (h, w). The temporal novelty of $\boldsymbol{v}_i^{h,w}$ is measured by its cosine distance from the token at the corresponding spatial location in the preceding frame: $d(\boldsymbol{v}_i^{h,w}) = 1 - \mathrm{sim}(\boldsymbol{v}_{i-1}^{h,w}, \boldsymbol{v}_i^{h,w})$, where $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity. In parallel, we quantify the spatial saliency $s(\boldsymbol{v}_i^{h,w})$ of each token using the L2-norm. Finally, the overall token importance score is calculated as:

$$g(\boldsymbol{v}_i^{h,w}) \;=\; \lambda \cdot d(\boldsymbol{v}_i^{h,w}) \;+\; (1-\lambda) \cdot s(\boldsymbol{v}_i^{h,w}), \tag{3}$$

where $\lambda \in [0, 1]$ is a weighting coefficient that controls the trade-off between novelty and saliency. The tokens of frame i are then ranked by their corresponding overall importance score $g(\cdot)$, and only the top-k_i of them are retained:

$$\mathcal{R}_i \;=\; \mathrm{TopK}_{(h,w)} \left(\{\boldsymbol{v}_i^{h,w}\}; g(\boldsymbol{v}_i^{h,w}), k_i \right), \tag{4}$$

where $\mathcal{R}_i$ denotes the resulting set of retained tokens. In this way, the refinement process preferentially preserves tokens that exhibit strong temporal novelty and high spatial saliency.

3.3 Query-agnostic Diversity-preserving Hierarchical Memory

As video frames continuously arrive and the video content evolves, the user's intent including the aspects and objects they focus on may also change over the course of interaction. To enable long-duration video understanding under a fixed context budget, we introduce a hierarchical memory architecture consisting of a short-term streaming memory which preserves recent visual evidence at relatively high fidelity, and a bounded Long-term memory which compactly maintain extensive historical content.

Short-term Streaming Memory. The short-term memory is implemented as a first-in-first-out (FIFO) queue containing the visual tokens of the W most recent frames. This memory explicitly preserves directly accessible recent visual context. Once a frame falls outside the W-frame window, its retained tokens are transferred to the long-term memory rather than being immediately discarded.

Diversity-Preserving Long-term Memory. To preserve informative evidence over arbitrarily long video streams, tokens evicted from the short-term memory are consolidated into a bounded long-term memory bank $\mathcal{M}$, which is initialized using tokens associated with a $\mathrm{New_scene}$ frame. When the memory bank has not yet reached its capacity, candidate tokens are selected according to their novelty with respect to the current memory. Given a collection of candidate tokens $\mathcal{E}$, we greedily select the token that is most dissimilar to the existing memory:

$$\boldsymbol{k}^* = \arg \max_{\boldsymbol{k} \in \mathcal{E} \setminus \mathcal{M}} d(\boldsymbol{k}, \mathcal{M}), \tag{5}$$

Instead of evicting tokens solely according to their residence time, we perform content-aware replacement to better preserve the diversity and coverage of the memory bank. Specifically, once $\mathcal{M}$ reaches its capacity, for an incoming candidate $\boldsymbol{k}$, we first identify its most similar existing memory item, The incoming token is admitted only when its novelty is sufficiently greater than that of the existing memory item:

$$\boldsymbol{m}_r = \arg\max_{\boldsymbol{m}\in\mathcal{M}} \mathrm{sim}(\boldsymbol{k}, \boldsymbol{m}), \qquad d(\boldsymbol{k}, \mathcal{M}) > \gamma \cdot d(\boldsymbol{m}_r, \mathcal{M}), \tag{6}$$

where γ is a replacement threshold controlling the aggressiveness of memory updates. A larger γ imposes a stricter admission criterion and therefore reduces the replacement frequency. When the condition in Eq. (6) is satisfied, $\boldsymbol{m}_r$ is replaced by $\boldsymbol{k}$. In our implementation, the novelty of a token can be measured by its cosine distance from the memory centroid:

$$d(\boldsymbol{k}, \mathcal{M}) = 1 - \mathrm{sim}(\boldsymbol{k}, \overline{\boldsymbol{m}}),$$

When a user query arrives, we retrieve the long-term memory entries that are most relevant to the query. Let $\boldsymbol{Q}$ denote the representation of the user query. The retrieved token set is defined as:

$$\mathcal{R}_{\boldsymbol{Q}} = \mathrm{TopK}_{\boldsymbol{m}\in\mathcal{M}} \left(\mathrm{sim}(\boldsymbol{Q}, \boldsymbol{m}), K_L\right), \tag{7}$$

where K_L denotes the retrieval budget. The retrieved long-term memory tokens $\mathcal{R}_{\boldsymbol{Q}}$ are concatenated with the tokens retained in the short-term memory and the query tokens, and the resulting sequence is provided to the LLM for response generation.

4 Experiments

4.1 Implementation Setup

Datasets and metrics. We evaluate the effectiveness of the StreamSieve framework on both online and offline benchmarks. For streaming video understanding, We ultize OVO-Bench (Niu et al., 2025) and StreamingBench (Lin et al., 2026). OVO-Bench evaluates online video understanding under three scenarios: backward tracing, realtime visual perception, and forward active responding. StreamingBench targets real-time comprehension of live streams. For offline video evaluation, we adopt Video-MME (Fu et al., 2025a) and LongVideoBench (Wu et al., 2024). Video-MME provides videos of varying lengths, up to one hour in length, paired with questions that assess both short-term perception and long-range reasoning abilities. LongVideoBench stresses very long videos whose query-relevant evidence is sparsely scattered throughout the video. For all question-answering tasks, we adopt accuracy as the primary performance metric.

Models and Baselines. To verify the broad applicability of our method, we select two popular open-source MLLM series, Qwen2.5-VL (Bai et al., 2025b) and InternVL2.5 (Chen et al., 2024c). We compare against a diverse set of representative streaming models to ensure a comprehensive analysis. For training-free methods, we compare with ReKV (Di et al., 2025), LiveVLM (Ning et al., 2025), StreamingToM (Chen et al., 2026a) and QueryStream (Zhang et al., 2026a). Additionally, we compare with training-based methods, including Flash-VStream(Zhang et al., 2025), VideoLLM-online (Chen et al., 2024a), Dispider (Qian et al., 2025) and TimeChat-Online (Yao et al., 2025).

Implementation details. For Qwen2.5-VL-7B, we sample at 1 fps and maintain its native dynamic resolution on video inputs, ensuring a fair comparison with the base model. For InternVL2.5-8B, the frames are sampled at 1 fps and resized to the resolution of 448×448, reflecting its official ingestion pipeline. In the dynamically encoding stage, we set the similarity thresholds to $\tau_{\mathrm{vh}} = 0.925$, $\tau_{\mathrm{h}} = 0.85$ and $\tau_{\mathrm{m}} = 0.55$, the shallow encoding depth to $L_s = 8$. and the retention ratios are set to 1.0, 0.75, 0.55. The short-term memory length to set to $W = 4$ frames in streaming benchmarks. All experiments are conducted on a single NVIDIA A800 GPU(80G).

4.2 Main Results on streaming video benchmarks

Results on OVO-Bench. As shown in Table 1 , our method achieves strong performance with existing streaming models on OVO-Bench. Without any training, the StreamSieve framework improves overall accuracy on both backbones. It improves Qwen2.5-VL from 52.3 to 56.5 with fewer visual tokens, surpassing fine-tuned online model despite the backbone not being trained for the streaming

Table 1: **Evaluation results on OVO-Bench.** Real-Time Visual Perception tasks consist of Optical Character Recognition (OCR), Action Recognition (ACR), Attribute Recognition (ATR), Spatial Understanding (STU), Future Prediction (FPD), Object Recognition (OJR). Backward Tracing tasks consists of Episodic Memory (EPM), Action Sequence Identification (ASI), Hallucination Detection (HLD). Best results among online models are in **bold** and the second-best results are underlined.

Model	#Frames	Real-Time Visual Perception							Backward Tracing				Overall
		OCR	ACR	ATR	STU	FPD	OJR	Avg.	EPM	ASI	HLD	Avg.	Avg.
Human Agents	-	94.0	92.6	94.8	92.7	91.1	94.0	93.2	92.6	93.0	91.4	92.3	92.8
Proprietary Multimodal Models													
Gemini 1.5 Pro	1 fps	87.3	67.0	80.2	54.5	68.3	67.4	70.8	68.6	75.7	52.7	62.3	66.6
GPT-4o	64	69.1	65.1	65.5	50.0	68.3	63.7	63.6	49.8	71.0	55.4	58.7	61.2
Open-source Offline VideoLLMs													
LLaVA-NeXT-Video-7B	64	69.8	59.6	66.4	50.6	72.3	61.4	63.3	51.2	64.2	9.7	41.7	52.5
LLaVA-OneVision-7B	64	67.1	58.7	69.8	49.4	71.3	60.3	62.8	52.5	58.8	23.7	45.0	53.9
Qwen2-VL-7B	64	69.1	53.2	63.8	50.6	66.3	60.9	60.7	44.4	66.9	34.4	48.6	54.7
InternVL-V2-8B	64	68.5	58.7	69.0	44.9	67.3	56.0	60.7	43.1	61.5	27.4	44.0	52.4
InternVL-2.5-8B	64	71.1	60.6	69.8	47.8	70.3	58.7	63.0	49.8	67.6	37.6	51.7	57.4
LongVU-7B	1 fps	55.7	49.5	59.5	48.3	68.3	63.0	57.4	43.1	66.2	9.1	39.5	48.5
Qwen2.5-VL-7B	1 fps	67.8	55.1	67.2	42.1	66.3	60.9	59.9	51.5	58.8	23.7	44.7	52.3
Open-source Online VideoLLMs													
Flash-VStream-7B	1 fps	25.5	32.1	29.3	33.7	29.7	28.8	29.9	36.4	33.8	5.9	25.4	27.7
VideoLLM-online-8B	2fps	8.1	23.9	12.1	14.0	45.5	21.2	20.8	22.2	18.8	12.2	17.7	19.3
Dispider-7B	1 fps	57.7	49.5	62.1	44.9	61.4	51.6	54.6	48.5	55.4	4.3	36.1	45.4
TimeChat-Online-7B	1 fps	75.2	46.8	70.7	47.8	69.3	61.4	61.9	55.9	59.5	9.7	41.7	51.8
StreamAgent-7B	1 fps	71.2	53.2	63.6	53.9	67.3	58.7	61.3	54.8	58.1	25.8	46.2	53.8
QueryStream-7B	1 fps	75.2	49.5	69.8	50.0	71.3	62.5	63.1	56.9	65.5	12.4	44.9	54.0
Qwen2.5-VL-7B + **StreamSieve**	1 fps	77.9	60.6	69.8	52.8	70.3	60.9	65.4	46.8	58.1	37.6	47.5	56.5
InternVL-2.5-8B + **StreamSieve**	1 fps	75.2	67.0	74.1	56.7	71.3	65.2	**68.3**	46.8	59.5	45.7	**50.7**	**59.5**

setting. The consistent improvements on both backbones demonstrate its superiority in real-time reasoning over a video stream. Likewise, significant gains could be observed on HLD sub-task of backward tracing, which requires the model to answer questions irrelevant to existing video inputs, indicating the long-short memory enables the model to retrieve more useful information from the historical context and thereby answer questions about past events more accurately.

Results on StreamingBench. For StreamingBench, which emphasizes the gap between offline processing and real-time human-like interaction, we evaluate on the official real-time visual understanding category. Table 2 demonstrates that our method achieves competitive performance on the Real-time Visual Understanding subtask of StreamingBench. Built on Qwen2.5-VL-7B, significant gains concentrate on Clips Summarization (+5.1) and Spatial Understanding (+4.9), and the resulting 70.33 on Spatial Understanding is the highest compared to the other model. For InternVL2.5-8B, the improvement can also be found in tasks that demand fine-grained perception of the current frames, such as Object Perception(81.84) and Event Understanding(79.87). This is consistent with what the framework eliminates: by removing redundant information and representing visual content in a compact form, it prevents contextual noise from affecting model inference.

4.3 Performance on Offline Long-video Tasks.

To evaluate the general video understanding capability of StreamSieve, we further evaluate its efficacy on two offline long video understanding benchmarks(VideoMME (Fu et al., 2025a) and LongVideoBench (Wu et al., 2024)). The results in Table 3 show that our framework achives compelling performance across both scenarios.

Results on LongVideoBench. As illustrated in Table 3, our framework exhibits superior performance on LongVideoBench compared to recent online Video MLLMs. Specifically, it achieves the leading performance with 59.3 accuracy, surpassing the strongest training-free baseline at 58.0 and outperforming other training-based methods listed. On InternVL2.5-8B it reaches 56.2, which indicates that the gain does not depend on a particular backbone. We attribute this phenomenon to the structure of long videos: they are dominated by redundant content, and that redundancy dilutes the evidence on which a question depends. This is precisely what our framework addresses: it can effectively filter out redundant noise from the videos, allowing the model to concentrate on the key information.

Table 2: **Performance comparison on StreamingBench.** Accuracy comparison on Streaming-Bench focusing on Real-Time Visual Understanding tasks. Real-Time Visual Understanding tasks consists of Object Perception (OP), Causal Reasoning (CR), Clips Summarization (CS), Attribute Perception (ATP), Event Understanding (EU), Text-Rich Understanding (TR), Prospective Reasoning (PR), Spatial Understanding (SU), Action Perception (ACP), and Counting (CT).

Model	#Frames	OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	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	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	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	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 VideoLLMs												
Video-LLaMA2-7B	32	55.86	55.47	57.41	58.17	52.80	43.61	39.81	42.68	45.61	35.23	49.52
VILA-1.5-8B	14	53.68	49.22	70.98	56.86	53.42	53.89	54.63	48.78	50.14	17.62	52.32
Video-CCAM-14B	96	56.40	57.81	65.30	62.75	64.60	51.40	42.59	47.97	49.58	31.61	53.96
LongVA-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-V2-8B	16	68.12	60.94	69.40	77.12	67.70	62.93	59.26	53.25	54.96	56.48	63.72
Kangaroo-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-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-2.6-8B	32	71.93	71.09	77.92	75.82	64.60	65.73	70.37	56.10	62.32	53.37	67.44
LLaVA-OneVision-7B	32	80.38	74.22	76.03	80.72	72.67	71.65	67.59	65.45	65.72	45.08	71.12
InternVL2.5-8B	16	71.00	73.44	73.82	80.71	69.62	68.85	58.33	59.76	60.80	57.45	68.21
Qwen2.5-VL-7B	1 fps	77.93	76.56	78.55	80.86	76.73	76.95	80.56	65.45	65.45	52.85	73.31
Open-source Online VideoLLMs												
Flash-VStream-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-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-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
ReKV-7B	0.5 fps	76.02	81.25	77.92	76.90	66.04	66.04	69.44	60.98	64.31	49.22	69.22
LiveVLM-7B	0.5 fps	81.47	78.13	83.28	79.08	69.57	74.14	75.00	69.11	67.71	40.41	72.92
TimeChat-Online-7B	1 fps	81.03	83.59	78.55	81.09	76.73	80.37	75.93	63.82	68.47	47.87	**74.32**
QueryStream-7B	1 fps	82.11	83.59	78.23	82.69	75.47	80.06	79.63	63.01	67.90	42.55	74.04
Qwen2.5-VL-7B + **StreamSieve**	1 fps	79.13	78.12	83.60	83.33	76.10	76.32	70.37	70.33	65.62	48.40	<u>74.16</u>
InternVL2.5-8B + **StreamSieve**	1 fps	81.84	71.09	82.65	85.26	79.87	67.91	59.26	67.48	66.19	62.77	73.88

Results on Video-MME. On VideoMME, our framework reaches 52.9 on Qwen2.5-VL-7B and 52.8 on InternVL2.5-8B for long-duration videos, the best long-video score reported in the table and above every training-based method, together with 62.5 and 62.1 overall, within 0.7 points of the strongest training-based result. StreamSieve excels particularly on Medium and Long subsets while showing comparable performance on Short videos, demonstrating robust capability across different temporal scales. The consistent improvements observed across different video durations demonstrate that our method can preserve both recent visual details and long-range information. Taken together, these results demonstrate the effectiveness and generality of our framework for long video understanding.

4.4 Ablation Studies

Contribution of Different Components. Table 4, our full StreamSieve achieves the best performance and also delivered consistent performance improvements when using individual modules. Notably, although TNDHE retains fewer tokens, it still achieves comparable performance to that of the individually optimized methods, indicating that these modules play complementary roles in enhancing video understanding.

Effect of short-term memory length. To investigate the impact of short-term memory length on understanding performance, we conduct ablations by varying the short-term memory Length W from 2 frames to 16 frames. As shown in Table 5, expanding the length from 2 frames to 4 frames improves performance. However, further increasing the length to 8 and 16 frames results in a decline in performance. This demonstrates the effectiveness of our hierarchical memory.

Efficiency Analysis. To evaluate the efficiency of our framework for long streaming video, we employ three metrics: peak GPU memory usage, Vison Encoder FLOPs and Time to First Token (TTFT), defined as the latency measured from the moment a user inputs a query to the decoding of the first output token. As shown in Figure3 (a), the baseline method quickly ran out of memory (OOM) as the number of video frames increased, while our method shows a slight increase. Without TNDHE, every frame would be fully encoded, resulting in higher GPU memory usage. In Figure 3

Table 3: **Results on offline long-video benchmarks.** We report accuracy on LongVideoBench and VideoMME (w/o subtitles). Here, "w/o sub." denotes evaluation on VideoMME without subtitles. **StreamSieve-7B** denotes the model based on Qwen2.5-VL-7B and **StreamSieve-8B** denotes the model based on InternVL2.5-8B.

Method	Frames	Offline Video				
		VideoMME				LongVideoBench
		Short	Medium	Long	All	Val
Open-source Offline MLLMs						
Videochat2-7B	16	48.3	37.0	33.2	39.5	39.3
LongVA-7B	128	61.1	50.4	46.2	52.6	–
LongVU-7B	1 fps	–	–	–	60.6	–
LLaVA-OneVision-7B	32	70.1	56.4	48.8	58.4	56.5
Qwen2.5-VL-7B	1 fps	–	–	50.4	63.2	61.5
Open-source Online MLLMs (Training-Based)						
Dispider-7B	1 fps	–	53.7	49.7	57.2	–
VideoChat-Online-4B	2 fps	–	–	44.9	52.8	54.1
Flash-VStream-7B	1 fps	**72.0**	61.1	50.3	61.2	–
TimeChat-Online-7B	1 fps	–	–	49.2	62.5	57.7
Open-source Online MLLMs (Training-Free)						
ReKV-7B	0.5 fps	70.4	55.4	47.3	57.7	–
LiveVLM-7B	0.5/0.2 fps	66.7	56.4	48.8	57.3	–
QueryStream-7B	1 fps	–	–	49.8	**63.2**	58.0
StreamingTOM-7B	1 fps	71.3	57.8	50.6	59.9	–
StreamSieve-7B	1 fps	70.9	**63.7**	**52.9**	62.5	**59.3**
StreamSieve-8B	64	71.7	61.8	52.8	62.1	56.2

Table 4: **Ablation study of different components on OVO-Bench.** We analyze the contribution of the *temporal router*, *token pruning*, and *memory*.

Model	Variants			All.
	Temporal	Spatial	Memory	
Qwen2.5-VL-7B	–	–	–	52.3
+ Dynamic Encoding	✓	–	–	53.4
+ Token Pruning	–	✓	–	53.7
+ TNDHE	✓	✓	–	53.5
Full StreamSieve	✓	✓	✓	**56.5**

Table 5: **Effect of the short-term memory length.** Effect of the recent-window length W on OVO-Bench. Real-Time and Bwd. report the average accuracy over the *real-time visual perception* and *Backward Tracing* categories, and Avg. is their mean.

Length	Real-Time	Bwd.	Avg.
2 Frames	64.2	46.9	55.6
4 Frames	**65.4**	47.5	**56.5**
8 Frames	64.5	47.3	55.9
16 Frames	63.9	**47.6**	55.8

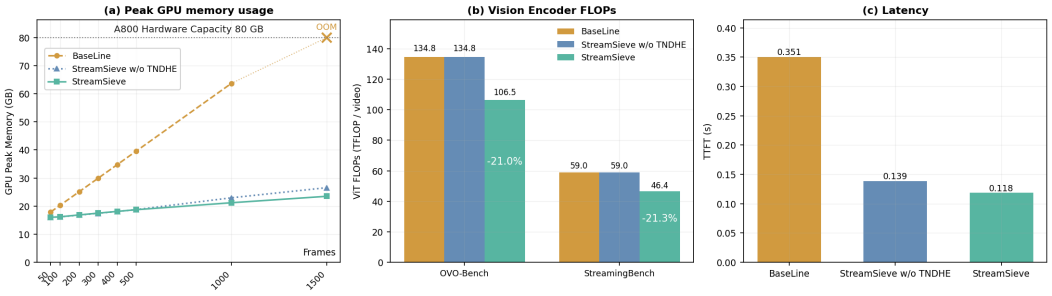


Figure 3: FLOPs, peak GPU memory usage and latency comparison.

(b), compared to the baseline model, TNDHE performs optimization before the video frames enter the vision encoder, thereby effectively reducing computational FLOPs (-21%). In terms of latency, StreamSieve improves the response speed (2.98x), with the result shown in Figure3 (c). These results collectively indicate our method's real time responsiveness in practical streaming scenarios.

5 Conclusion

In this work, we propose StreamSieve, a training-free and query-agnostic framework for streaming video understanding. By introducing the Refined Dynamic Encoding for Spatiotemporal Redundancy, our method effectively mitigate the redundancy in the video. We further develop a Query-agnostic Diversity-preserving Hierarchical Memory, which enables the model to perceive recent content and access relevant historical visual information when needed. Extensive experiments across different Video LLMs and benchmarks demonstrate the superior performance of StreamSieve.

AI USE STATEMENT

In line with the conference policy, we disclose that LLMs were used for literature discovery, drafting preliminary text for select narrative sections (e.g., portions of the Introduction and Related Work), and language polishing. All AI-assisted text was critically reviewed, verified, and heavily revised by the authors. All scientific contributions—including core ideas, methodology, experiments, and conclusions—were made exclusively by the authors, who take full responsibility for this work.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of our work. The source code for StreamSieve will be made publicly available upon publication. Our framework is built upon publicly available checkpoints, and our experiments are conducted using two frozen open-weight video language models: the released TimeChat-Online-7B checkpoint for Qwen2.5-VL and InternVL2.5-8B. All relevant citations are provided in the main text. We evaluate our method on four standard and publicly available benchmarks, OVO-Bench, StreamingBench, Video-MME and LongVideoBench, following their official evaluation protocols; for StreamingBench we report the Real-Time Visual Understanding category, and accuracy is computed by each benchmark's own script. StreamSieve requires no additional training and is applied entirely at inference time.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

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

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024b.

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. Streamingtom: Streaming token compression for efficient video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24675–24685, 2026a.

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

Yuhao Chen, Bin Shan, Xin Ye, and Cheng Chen. Evoprune: Early-stage visual token pruning for efficient mllms. *arXiv preprint arXiv:2603.03681*, 2026c.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024c.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 42115–42127, 2025.

Xin Ding, Hao Wu, Yifan Yang, Shiqi Jiang, Qianxi Zhang, Donglin Bai, Zhibo Chen, and Ting Cao. Streammind: Unlocking full frame rate streaming video dialogue through event-gated cognition. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13448–13459. IEEE, 2025.

Yue Fan, Xiaojian Ma, Rongpeng Su, Jun Guo, Rujie Wu, Xi Chen, and Qing Li. Embodied videoagent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6342–6352. IEEE, 2025.

Yue Feng, Yingshi Liu, Jingrou Zhang, Jie Qin, et al. Monitor: Exploiting large language models with instruction for online video anomaly detection. *Advances in Neural Information Processing Systems*, 38:35851–35871, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, 2025a.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Yi-Xing Peng, Kun-Yu Lin, Xihan Wei, Jian-Fang Hu, Xiaohua Xie, and Wei-Shi Zheng. Vispeak: Visual instruction feedback in streaming videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21778–21788. IEEE, 2025b.

Yansong Guo, Chaoyang Zhu, Jiayi Ji, Jianghang Lin, and Liujuan Cao. Hieravid: Hierarchical token pruning for fast video large language models. *arXiv preprint arXiv:2604.01881*, 2026.

Yifei Huang, Jilan Xu, Baoqi Pei, Yuping He, Guo Chen, Mingfang Zhang, Lijin Yang, Zheng Nie, Jinyao Liu, Guoshun Fan, et al. An egocentric vision-language model based portable real-time smart assistant. *arXiv preprint arXiv:2503.04250*, 2025.

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

Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. Infinipot-v: Memory-constrained kv cache compression for streaming video understanding. *Advances in Neural Information Processing Systems*, 38:138983–139013, 2026.

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

Niu Lian, Yuting Wang, Hanshu Yao, Jinpeng Wang, Bin Chen, Yaowei Wang, Min Zhang, and Shu-Tao Xia. From verbatim to gist: Distilling pyramidal multimodal memory via semantic information bottleneck for long-horizon video agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11601–11617, 2026.

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

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

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

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24045–24055. IEEE, 2025.

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Holitom: Holistic token merging for fast video large language models. *Advances in Neural Information Processing Systems*, 38:135547–135570, 2026.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, et al. Fastvid: Dynamic density pruning for fast video large language models. *Advances in Neural Information Processing Systems*, 38:123553–123581, 2026.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):1355–1376, 2025.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18992–19001. IEEE, 2025.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. Streambridge: Turning your offline video large language model into a proactive streaming assistant. *Advances in Neural Information Processing Systems*, 38:132332–132359, 2026a.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22442–22452. IEEE, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Yiyu Wang, Xuyang Liu, Xiyan Gui, Xinying Lin, Boxue Yang, Chenfei Liao, Tailai Chen, and Linfeng Zhang. Accelerating streaming video large language models via hierarchical token compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18523–18533, 2026b.

Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Jiansheng Wei, Huishuai Zhang, and Dongyan Zhao. Videollm knows when to speak: Enhancing time-sensitive video comprehension with video-text duet interaction format. In *EMNLP (Findings)*, pp. 6338–6359, 2025c.

Yuxuan Wang, Yiqi Song, Cihang Xie, Yang Liu, and Zilong Zheng. Videollamb: Long streaming video understanding with recurrent memory bridges. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24170–24181. IEEE, 2025d.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Jiaer Xia, Peixian Chen, Mengdan Zhang, Xing Sun, and Kaiyang Zhou. Streaming video instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31219–31229, 2026.

Junbin Xiao, Jiajun Chen, Tianxiang Sun, Xun Yang, and Angela Yao. Mukv: Multi-grained kv cache compression for long streaming video question-answering. 2026.

Yichen Xie, Runsheng Xu, Tong He, Jyh-Jing Hwang, Katie Luo, Jingwei Ji, Hubert Lin, Letian Chen, Yiren Lu, Zhaoqi Leng, et al. S4-driver: Scalable self-supervised driving multimodal large language model with spatio-temporal visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1622–1632, 2025.

Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. Streammem: Query-agnostic kv cache memory for streaming video understanding. *arXiv preprint arXiv:2508.15717*, 2025a.

Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 42522–42556, 2025b.

Zhenyu Yang, Kairui Zhang, Yuhang Hu, Bing Wang, Shengsheng Qian, Bin Wen, Fan Yang, Tingting Gao, Weiming Dong, and Changsheng Xu. Livestar: Live streaming assistant for real-world online video understanding. *Advances in Neural Information Processing Systems*, 38:31266–31304, 2026a.

Zhenyu Yang, Kairui Zhang, Qi Liu, Tiancheng Liu, Long Ying, Dizhan Xue, Qibin Hou, Shengsheng Qian, and Changsheng Xu. Towards online interactors: A comprehensive survey on streaming video understanding. *Preprints*, June 2026b. doi: 10.20944/preprints202606.1674.v1. URL https://doi.org/10.20944/preprints202606.1674.v1.

Zhiwei Yang, Chen Gao, Jing Liu, Peng Wu, Guansong Pang, and Mike Zheng Shou. Assistpda: An online video surveillance assistant for video anomaly prediction, detection, and analysis. *arXiv preprint arXiv:2503.21904*, 2025c.

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

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. Videoarm: Agentic reasoning over hierarchical memory for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Xueyang Yu, Cheng Shi, Yang Wang, and Sibei Yang. Eyes wide open: Ego proactive video-llm for streaming video. *Advances in Neural Information Processing Systems*, 38:13420–13463, 2026.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, et al. Streamforest: Efficient online video understanding with persistent event memory. *Advances in Neural Information Processing Systems*, 38:75804–75835, 2026.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21059–21069. IEEE, 2025.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *arXiv preprint arXiv:2412.06224*, 2024.

Kairui Zhang, Zhenyu Yang, Bing Wang, Shengsheng Qian, and Changsheng Xu. Querystream: Advancing streaming video understanding with query-aware pruning and proactive response. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *Advances in Neural Information Processing Systems*, 38:25438–25468, 2026b.