TRACE: NOVELTY-AWARE SPATIOTEMPORAL EVIDENCE RETENTION FOR STREAMING VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming video understanding requires vision-language models to process evolving streams causally and preserve visual evidence for queries arriving at arbitrary times. However, models must make irreversible retention decisions under a bounded memory budget, without access to future observations or queries. Existing methods often treat temporal and spatial redundancy separately or rely on attention-based importance scores, potentially discarding useful evidence before complementary cues are jointly considered. To address this challenge, we propose TRACE, a training-free framework for novelty-aware spatiotemporal evidence retention. Following an observe–score–retain procedure, TRACE maintains a causal hierarchical visual memory that preserves recent observations at full granularity while characterizing visual tokens through two complementary query-agnostic signals: temporal novelty and spatial distinctiveness. Their joint evaluation enables adaptive evidence retention without relying on attention weights, while temporal gap-aware positional re-encoding supports long-horizon streaming inference. Extensive experiments across multiple VLM backbones demonstrate that TRACE achieves state-of-the-art performance among the compared open-source methods on streaming video understanding benchmarks, while preserving 91.1% of the corresponding backbone performance with only 8.2% of the visual tokens on long-video benchmarks. TRACE further maintains bounded visual KV memory and stable query latency over hour-scale streams.

1 INTRODUCTION

The rapid development of large language models (LLMs) (Achiam et al., 2023; Grattafiori et al., 2024) has driven substantial progress in vision-language models (VLMs) (Wang et al., 2025; Bai et al., 2025a), enabling increasingly powerful multimodal perception and reasoning. However, many latency-sensitive applications, such as autonomous driving (Cao et al., 2026), robotic interaction (Liu et al., 2026), and video assistants (Yan et al., 2026), require models to continuously process incoming visual observations and respond before the complete video becomes available. These requirements motivate streaming video understanding, where a VLM operates causally over an evolving video stream that may continue for an unknown duration, while user queries may arrive at arbitrary times. A practical streaming model must therefore preserve useful visual evidence from the observed history while keeping visual memory and inference latency bounded over time.

This setting introduces a distinctive challenge: *visual memory must be selected before its future utility is known*. At time t, the model has access only to the observed stream $\mathcal{F}_t = \{f_1, \ldots, f_t\}$, while neither future observations nor future queries are available. Once a visual state is removed from memory, the corresponding evidence cannot be recovered if it later becomes relevant to a query. Consequently, streaming memory management is not merely a problem of reducing redundant tokens, but requires *causal and query-agnostic evidence retention*: under a bounded memory budget, the model must preserve visual states that may support unknown future queries.

Existing streaming approaches address this challenge at different levels of granularity. Frame-level selection (Yang et al., 2026) can efficiently control memory growth, but discarding an entire frame may remove localized cues that become important later. Token-level methods provide finer control over visual history, yet require a reliable criterion for determining which visual states deserve long-term retention. Some feature-based approaches reduce temporal and spatial redundancy through

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.

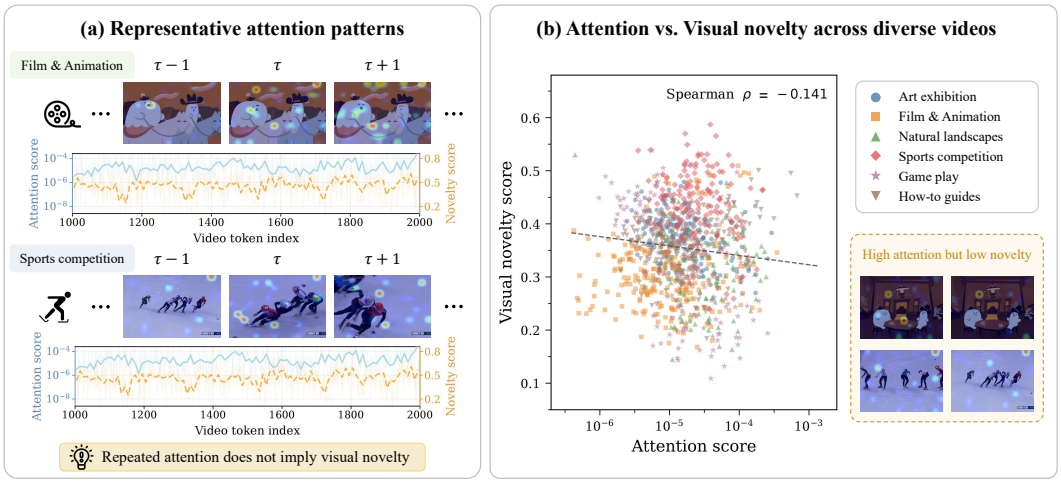


Figure 1: Relationship between attention and visual novelty in streaming video. (a) Two examples show that attention repeatedly focuses on similar regions across neighboring frames, while visual novelty does not consistently follow the attention pattern. (b) Across six video categories, attention and visual novelty are only weakly correlated (Spearman's $\rho = -0.141$), including highly attended tokens with low novelty. These observations motivate visual novelty as a query-agnostic signal for streaming visual-memory retention.

separate or sequential operations (Xie et al., 2026). Such sequential compression risks prematurely discarding complementary evidence: a region that is temporally stable may still contain a spatially distinctive detail, whereas a spatially subtle region may signal an important temporal transition. Reliable retention should therefore assess temporal and spatial evidence jointly.

Another common strategy estimates token importance from attention weights (Zhang et al., 2026). However, attention is not necessarily an appropriate proxy for long-term visual evidence, as it reflects the model's current computation whereas streaming retention must preserve information for *unknown future queries*. As illustrated in Fig. 1, highly attended visual states can remain redundant across neighboring frames, and attention shows only weak correspondence with visual novelty. Moreover, efficient attention implementations such as FlashAttention (Dao, 2024) avoid explicitly constructing the full attention matrix, making attention-based scoring inconvenient and potentially costly during streaming inference. These observations motivate an attention-free retention criterion that directly characterizes the novelty of visual evidence.

We therefore propose **TRACE**, a training-free framework for **novelty-aware spatiotemporal evidence retention** in streaming video understanding. The central premise of TRACE is simple: when future queries are unknown, visual states should be retained according to the new information they contribute to the observed stream rather than their importance to the current computation. We characterize this information through two complementary signals: *temporal novelty*, which captures visual changes across neighboring observations, and *spatial distinctiveness*, which measures how strongly a visual state differs from its local spatial context. Their joint evaluation provides a query-agnostic proxy for evidence retention without allowing either dimension to discard visual states before the other is considered.

TRACE realizes this principle through a causal hierarchical visual memory. Newly observed frames are preserved at full granularity in a short-term observation window before irreversible long-term retention decisions are made. As observations leave this window, TRACE evaluates their spatiotemporal novelty and retains the most informative visual KV states under a bounded visual-memory budget. For extended streams, TRACE further uses temporal gap-aware positional re-encoding to control the growth of temporal coordinates while preserving temporal order and coarse temporal separation between retained observations.

Extensive experiments demonstrate the effectiveness and scalability of TRACE. On OVO-Bench and StreamingBench, TRACE consistently improves different VLM backbones in both real-time visual understanding and historical information retrieval. On long-video benchmarks, TRACE retains only 8.2% of the original visual tokens while preserving 91.1% of the corresponding backbone per-

formance on MLVU and VideoMME. TRACE further maintains bounded visual KV memory and stable query latency over streams of up to 4,096 frames, corresponding to more than one hour of video at 1 fps. Our main contributions are summarized as follows:

- We formulate visual memory maintenance in streaming video understanding as a causal and query-agnostic evidence-retention problem under a bounded memory budget, and introduce spatiotemporal novelty as a retention signal for visual evidence that may support unknown future queries.
- We propose TRACE, a training-free visual KV retention framework that preserves recent observations at full granularity and jointly evaluates temporal novelty and spatial distinctiveness for historical evidence retention. A causal hierarchical memory and temporal gap-aware positional re-encoding further support bounded long-horizon streaming inference.
- Extensive experiments on online and long-video understanding benchmarks demonstrate consistent gains across multiple VLM backbones, strong performance under substantial visual-token reduction, and stable hour-scale streaming inference.

2 Related Work

2.1 Streaming Video Understanding

Conventional video understanding primarily follows an offline setting (Li et al., 2024a; Qin et al., 2026; Gao et al., 2026), where the complete video and user query are available before inference. Streaming video understanding instead processes continuously arriving observations under causal constraints, requiring models to update their visual context as events unfold and respond using only the video content observed so far (Wang et al., 2026a; Sun et al., 2026; Lu et al., 2026). Existing streaming methods can be broadly divided according to how interaction is triggered. Query-triggered approaches continuously maintain historical visual information and access the accumulated context when a user query arrives (Xiong et al., 2025; Di et al., 2025; Wu et al., 2026). Proactive approaches, by contrast, operate under an initial query or instruction and continuously monitor subsequent observations, autonomously deciding whether and when to respond (Qian et al., 2025; Ma et al., 2026). Many proactive methods couple continuous perception with response-timing decisions through additional training (Chen et al., 2025; Wang et al., 2026b). TRACE focuses on the query-triggered setting, where future queries are unknown while visual memory must be maintained causally under a bounded budget.

2.2 Visual Token Compression and Streaming Memory Management

VLMs encode video frames into dense visual-token sequences containing substantial temporal and spatial redundancy (Huang et al., 2025a; Fang et al., 2026). Existing compression methods mainly reduce this redundancy through token pruning or merging. Pruning retains a subset of visual tokens (Ji et al., 2026), whereas merging aggregates related tokens into compact representations (Bolya et al., 2023; Zeng et al., 2026). Compression decisions are commonly guided by attention scores (Xiao et al., 2026) or feature similarity (Luo et al., 2026). For streaming video, memory management must further preserve useful historical context as new observations continuously arrive. Hierarchical memory designs organize visual information at multiple temporal scales or granularities (Huang et al., 2025b; Xie et al., 2026; Wang et al., 2026a; Zhang et al., 2026), typically preserving fine-grained recent observations while selectively compressing or discarding older memory entries. TRACE instead jointly evaluates temporal novelty and spatial distinctiveness before any token removal, providing a query-agnostic criterion for historical evidence retention.

3 Method

3.1 Method Overview

We consider query-triggered streaming video understanding, where frames arrive sequentially and user queries may be issued at any time. At streaming step t, the model can access only the observed video $\mathcal{F}_t = \{f_1, \ldots, f_t\}$, while future observations and future queries remain unknown. Our goal is to preserve useful visual evidence under a bounded KV memory budget without using future information.

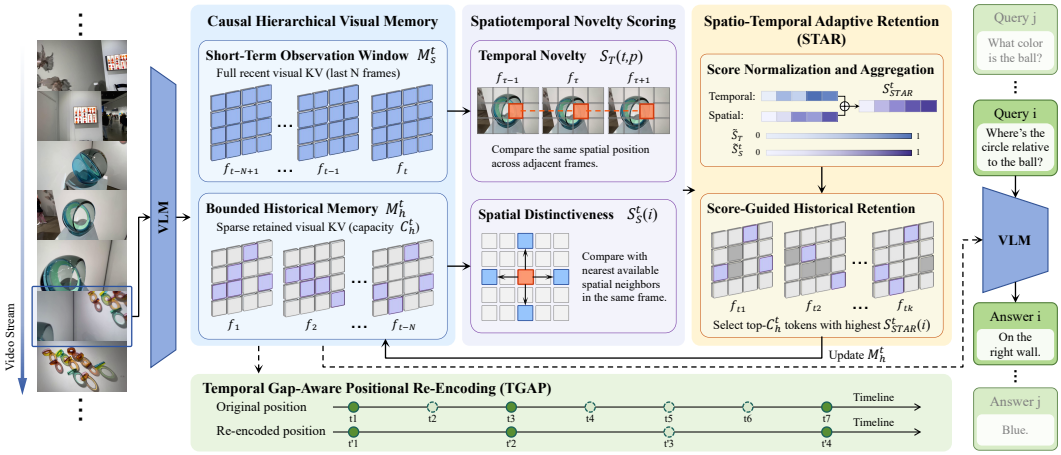


Figure 2: Overview of TRACE for query-triggered streaming video understanding. Incoming frames are encoded into visual KV states and organized in a causal hierarchical memory, where recent observations are preserved in the short-term memory $\mathcal{M}_s^t$ and earlier evidence is sparsely retained in the bounded historical memory $\mathcal{M}_h^t$. TRACE computes temporal novelty S_T and spatial distinctiveness S_S^t, which are jointly aggregated by STAR into S_{STAR}^t to retain historical visual states under capacity C_h^t. TGAP re-encodes the temporal positions of retained visual keys to preserve coarse temporal structure over long streams. At any query time, the current visual memory is provided to the VLM for answering while continuing to evolve with incoming frames.

As illustrated in Fig. 2, TRACE follows an observe–score–retain procedure for causal visual memory management. Recent frames are first preserved at full granularity in a short-term observation window, while older evidence is maintained in a bounded historical memory. Before a frame leaves the short-term window, TRACE computes the temporal novelty of its visual tokens. Their spatial distinctiveness is further evaluated with respect to the evolving historical memory. These complementary signals are jointly used by Spatio-Temporal Adaptive Retention (STAR) to determine which visual states should remain in historical memory.

For long video streams, TRACE further applies Temporal Gap-Aware Positional Re-Encoding (TGAP) to the retained visual keys. TGAP compresses the growing temporal coordinate range while preserving temporal order and coarse temporal separation between retained observations. Together, these components enable TRACE to maintain query-agnostic visual evidence under a bounded memory budget throughout streaming inference.

3.2 Causal Hierarchical Visual Memory

TRACE organizes visual evidence at two temporal scales under a fixed visual-token budget: short-term memory $\mathcal{M}_s$ preserves recent observations at full granularity, while historical memory $\mathcal{M}_h$ stores a sparse subset of earlier visual states. $\mathcal{M}_s$ also serves as a causal observation window before irreversible long-term retention decisions. Let f_t denote the frame observed at streaming step t. After visual encoding and multimodal projection, f_t produces P_t visual tokens, where $P_t = h_t \times w_t$.

For token p in frame t, let $K_{t,p}^{(l)}$ and $V_{t,p}^{(l)}$ denote its key and value states at transformer layer $l \in \{1, \ldots, L\}$. We use $\mathbf{K}_{t,p}$ and $\mathbf{V}_{t,p}$ to denote the respective collections of these states across all transformer layers. Thus, one visual-token index corresponds to its KV states across all layers. We use C to denote the maximum number of visual-token indices maintained in memory.

Short-Term Observation Window. The short-term memory keeps the complete visual KV states of the most recent N frames, with token occupancy C_s^t:

$$\mathcal{M}_s^t = \{(\mathbf{K}_{\tau,p}, \mathbf{V}_{\tau,p}) \mid \max(1, t-N+1) \le \tau \le t,\ 1 \le p \le P_\tau\}, \tag{1}$$

where $C_s^t = |\mathcal{M}_s^t| = \sum_{\tau=\max(1,t-N+1)}^{t} P_\tau$ denotes the current short-term token occupancy. When the short-term window would exceed N frames, the oldest frame is evaluated before being removed from $\mathcal{M}_s$. By this time, the local observations required to finalize its temporal novelty have al-

ready arrived. The subsequent retention decision therefore uses only information observed up to the current streaming step.

Bounded Historical Memory. The historical memory stores selected visual tokens from frames that have left the short-term observation window. Under the total visual-token budget C, its available capacity at streaming step t is $C_h^t = C - C_s^t$. Let $\mathcal{I}_t$ denote the indices of visual tokens retained in historical memory. We define:

$$\mathcal{M}_h^t = \{(\mathbf{K}_{\tau,p}, \mathbf{V}_{\tau,p}) \mid (\tau, p) \in \mathcal{I}_t\}, \qquad |\mathcal{I}_t| \le C_h^t. \tag{2}$$

Thus, $|\mathcal{M}_s^t| + |\mathcal{M}_h^t| \le C$, ensuring that the visual KV memory remains bounded as the stream grows. The same retained token indices are applied to the KV states of all transformer layers, preserving token alignment across the model. When a frame leaves the short-term window, the candidate set $\mathcal{C}_t$ combines all its visual-token indices with those already retained in historical memory. The candidates in $\mathcal{C}_t$ are evaluated using the spatiotemporal novelty scores defined in Sec. 3.3, after which at most C_h^t tokens are selected to update the historical memory. The transition procedure is detailed in Appendix B.1.

3.3 Spatiotemporal Novelty Scoring

TRACE measures visual evidence from two complementary aspects: *temporal novelty*, which captures changes across nearby observations, and *spatial distinctiveness*, which measures how different a visual token is from the visual evidence currently retained around it. Since future queries are unknown, both scores are computed directly from visual representations without relying on query-conditioned attention. Temporal novelty is computed within the short-term observation window and then cached, while spatial distinctiveness is reassessed as the historical memory evolves.

Visual Representation for Novelty Scoring. We compute novelty scores using post-RoPE keys from a fixed anchor transformer layer $l^\star$. For visual token p in frame t, let $\bar{\mathbf{k}}_{t,p}^m \in \mathbb{R}^{d_h}$ denote its key representation at KV head m, where $m \in \{1, \dots, H_{\text{kv}}\}$. We average the key representations across all KV heads to obtain one token-level scoring representation $\tilde{\mathbf{k}}_{t,p} \in \mathbb{R}^{d_h}$. The Euclidean distance between two visual tokens i and j is defined as:

$$D(i,j) = \left\| \tilde{\mathbf{k}}_i - \tilde{\mathbf{k}}_j \right\|_2. \tag{3}$$

For each token retained in the historical memory, its post-RoPE key at the anchor layer is cached as auxiliary metadata and reused for subsequent spatial reassessment.

Temporal Novelty. Temporal novelty measures the representation change of a visual token across adjacent frames. For token p in frame t, let $\mathcal{T}_t = \{t-1, t+1\}$ denote its adjacent frames once they have been observed. Assuming a consistent spatial token grid, we compare tokens at the same spatial position p. Once both neighboring frames have been observed, the temporal novelty score is

$$S_T(t,p) = \frac{1}{|\mathcal{T}_t|} \sum_{\tau \in \mathcal{T}_t} D\big((t,p), (\tau,p)\big). \tag{4}$$

The temporal score is accumulated within the short-term observation window. The resulting score is cached and is not recomputed after the token enters the historical memory. At video boundaries, only the available temporal neighbor is used. Since each comparison is performed only after the corresponding neighboring frame has been observed, the procedure remains causal.

Spatial Distinctiveness. Unlike temporal novelty, spatial distinctiveness depends on the sparse within-frame layout induced by historical retention. As tokens from a retained frame are removed, the spatial neighborhood of the remaining tokens may change. A token that is redundant under one memory state may therefore become distinctive after nearby visual evidence is discarded. For a candidate token $i = (\tau, p) \in \mathcal{C}_t$, we define $\mathcal{N}_S^t(i)$ as the nearest available token from the same frame in each of the four cardinal directions. Only directions containing an available token are included in $\mathcal{N}_S^t(i)$. The spatial distinctiveness score is

$$S_S^t(i) = \frac{1}{|\mathcal{N}_S^t(i)|} \sum_{j \in \mathcal{N}_S^t(i)} D(i,j). \tag{5}$$

A high $S_S^t(i)$ indicates that the token differs strongly from the visual evidence currently retained around it and is therefore less spatially redundant. Since historical retention changes the within-frame neighborhood structure, $S_S^t(i)$ is recomputed before each historical-memory update.

3.4 Spatio-Temporal Adaptive Retention

Given temporal novelty $S_T(i)$ and spatial distinctiveness $S_S^t(i)$, we introduce Spatio-Temporal Adaptive Retention (STAR) to determine which visual tokens should remain in the bounded historical memory. Rather than pruning based on either signal alone, STAR first normalizes and combines the two signals, then makes a single joint retention decision over the current candidate set $\mathcal{C}_t$.

Score Normalization and Aggregation. Although $S_T(i)$ and $S_S^t(i)$ are both derived from visual representation differences, their score distributions may differ. We therefore normalize them independently according to their ranks within $\mathcal{C}_t$. Let $N_t = |\mathcal{C}_t|$, and let $r_T^t(i)$ and $r_S^t(i)$ denote the ascending ranks of candidate i according to $S_T(i)$ and $S_S^t(i)$, respectively, where rank 1 corresponds to the smallest score and rank N_t to the largest. The normalized scores are

$$\widetilde{S}_T^t(i) = \frac{r_T^t(i) - 1}{N_t - 1}, \qquad \widetilde{S}_S^t(i) = \frac{r_S^t(i) - 1}{N_t - 1}. \tag{6}$$

Tied values are assigned their average rank, so both normalized scores lie in $[0, 1]$. Rank normalization makes the two signals comparable without requiring their raw score distributions to share the same scale. We then compute the joint spatiotemporal retention score as:

$$S_{\mathrm{STAR}}^t(i) = \alpha \widetilde{S}_T^t(i) + (1 - \alpha)\widetilde{S}_S^t(i), \qquad \alpha \in [0, 1], \tag{7}$$

where α controls the relative contribution of temporal novelty and spatial distinctiveness. No candidate is removed before both signals have been normalized and jointly evaluated, ensuring that neither temporal nor spatial evidence alone determines the retention decision.

Score-Guided Historical Retention. Given the historical-memory capacity C_h^t available at streaming step t, STAR retains the candidates with the highest joint scores:

$$\mathcal{I}_t = \mathrm{TopK}_{C_h^t}\left(\left\{ S_{\mathrm{STAR}}^t(i) \mid i \in \mathcal{C}_t \right\}\right). \tag{8}$$

If $|\mathcal{C}_t| \leq C_h^t$, all candidates are retained. The resulting index set $\mathcal{I}_t$ defines the updated historical memory according to Eq. 2, and the same retention indices are applied to the visual KV states across all transformer layers. Since spatial distinctiveness and score rankings are updated with the evolving historical memory, the effective retention threshold adapts as new visual evidence arrives. In contrast, previously computed temporal novelty scores are reused without repeated temporal matching.

3.5 Temporal Gap-Aware Positional Re-Encoding

As the video stream grows, the visual states retained in memory may span an increasingly wide temporal range. Directly preserving their original temporal coordinates causes the positional range to grow with stream duration, whereas assigning consecutive positions to retained frames removes the temporal gaps between visual events. We therefore introduce Temporal Gap-Aware Positional Re-Encoding (TGAP) to compress temporal coordinates while preserving temporal order and coarse temporal separation. Let $\{t_1, t_2, \ldots, t_J\}$ denote the ordered timestamps of the distinct frames represented in the current visual memory $\mathcal{M}_h^t \cup \mathcal{M}_s^t$, where $t_1 < t_2 < \cdots < t_J$. For two consecutive retained frames, their original temporal gap and its compressed counterpart are defined as:

$$\Delta_j = t_j - t_{j-1}, \qquad \widetilde{\Delta}_j = \ln(1 + \Delta_j), \quad j = 2, \ldots, J. \tag{9}$$

Because $\widetilde{\Delta}_j > 0$, accumulating these compressed gaps preserves the temporal order of retained frames. The logarithmic mapping compresses large temporal gaps more strongly than small ones, thereby reducing the positional range without collapsing retained frames into uniformly consecutive positions. The temporal coordinate construction and the corresponding re-encoding of cached visual keys are detailed in Appendix B.2.

Table 1: Performance on OVO-Bench for Real-Time Visual Perception and Backward Tracing. Best results within each backbone group are shown in **bold**, and † denotes reproduced results.

Method	Venue	Frames	Real-Time Visual Perception							Backward Tracing			
			OCR	ACR	ATR	STU	FPD	OJR	Avg.	EPM	ASI	HLD	Avg.
Proprietary Models													
Gemini 1.5 Pro	–	1 fps	85.9	67.0	79.3	58.4	63.4	62.0	69.3	58.6	76.4	52.6	62.5
GPT-4o	–	64	69.8	64.2	71.6	51.1	70.3	59.8	64.5	57.9	75.7	48.7	60.8
Open-source Offline MLLMs													
LongVU-7B	–	1 fps	53.7	53.2	62.9	47.8	68.3	59.8	57.6	43.1	66.2	9.1	39.5
LLaVA-OneVision-7B	–	64	66.4	57.8	73.3	53.4	71.3	62.0	64.0	54.2	55.4	21.5	43.7
InternVL2-8B	–	64	67.1	60.6	63.8	46.1	68.3	56.5	60.4	43.1	61.5	27.4	44.0
Open-source Online MLLMs													
VideoLLM-Online-8B	CVPR24	2 fps	8.1	23.9	12.1	14.0	45.5	21.2	20.8	22.2	18.8	12.2	17.7
Dispider-7B	CVPR25	1 fps	57.7	49.5	62.1	44.9	61.4	51.6	54.6	48.5	55.4	4.3	36.1
Flash-VStream-7B	ICCV25	1 fps	24.2	29.4	28.5	33.7	25.7	28.8	28.4	39.1	37.2	5.9	27.4
TimeChat-Online-7B	ACMMM25	1 fps	74.5	48.6	68.1	48.3	69.3	59.8	61.4	55.9	59.5	9.7	41.7
StreamForest-7B	NeurIPS25	1 fps	68.5	53.2	71.6	47.8	65.4	60.9	61.2	58.9	64.9	32.3	52.0
Within-Backbone													
Qwen2.5-VL-7B†	–	1 fps	79.2	53.2	67.2	51.7	71.3	57.1	63.3	49.5	59.5	21.0	43.4
+ FluxMem	CVPR26	1 fps	81.2	59.6	70.7	53.4	**75.2**	63.0	67.2	50.8	60.1	30.6	47.2
+ HERMES (4k)	ACL26	1 fps	85.2	**64.2**	71.6	53.4	74.3	65.2	69.0	48.5	62.2	**37.6**	49.4
+ TRACE (Ours)	–	1 fps	**89.9**	**64.2**	**75.9**	**53.9**	73.3	**73.9**	**71.5**	**56.6**	**67.6**	28.0	**50.7**
Qwen3-VL-4B†	–	1 fps	77.9	65.1	**82.8**	60.7	71.3	76.6	72.2	49.8	52.0	18.8	41.2
+ TRACE (Ours)	–	1 fps	**89.9**	**76.2**	75.9	**65.2**	**76.2**	**86.4**	**78.5**	**58.6**	**63.5**	**41.4**	**54.7**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Implementation Details. To comprehensively evaluate our method's effectiveness, we use Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-4B (Bai et al., 2025a) as backbones and compare our approach with a range of state-of-the-art streaming video understanding models. For both online and offline benchmarks, videos are sampled at 1 fps. We allocate up to 256 visual tokens per frame for most benchmarks, and reduce this number to 64 for long-video benchmarks. In both settings, the total visual-token capacity is set to 5,120, the short-term memory length to 6 frames, and α to 0.6. The threshold for triggering TGAP is set to 120k input tokens. The final transformer layer is used as the anchor layer in all experiments. All experiments are conducted on NVIDIA GeForce RTX 4090 GPUs.

Benchmarks. We evaluate TRACE on diverse streaming and offline video understanding benchmarks. For streaming understanding, we use StreamingBench (Lin et al., 2026) and OVO-Bench (Niu et al., 2025). For offline video evaluation, we adopt MVBench (Li et al., 2024b), MLVU (Zhou et al., 2025) and VideoMME (Fu et al., 2025), which collectively cover videos of varying durations.

4.2 MAIN RESULTS

4.2.1 STREAMING VIDEO EVALUATION

Table 1 compares TRACE with existing methods on OVO-Bench. TRACE improves both evaluated VLM backbones on Real-Time Visual Perception and Backward Tracing. On Qwen2.5-VL-7B, the average scores increase from 63.3 to 71.5 and from 43.4 to 50.7, respectively. On Qwen3-VL-4B, the corresponding scores increase from 72.2 to 78.5 and from 41.2 to 54.7. The gains are observed on both evaluation tracks for both backbones, with the largest improvement of 13.5 points on Backward Tracing with Qwen3-VL-4B. Table 2 reports the results on the Real-Time Visual Understanding track of StreamingBench. TRACE improves the average score of Qwen2.5-VL-7B from 73.9 to 79.7 and that of Qwen3-VL-4B from 78.3 to 82.3. The latter achieves the highest average score among the open-source methods reported in Table 2. These results show that TRACE improves overall real-time visual understanding across both evaluated backbones.

Table 2: Performance comparison on the Real-Time Visual Understanding track of StreamingBench. The best results within each backbone group are shown in **bold**, and † denotes reproduced results.

Method	Venue	Frames	Real-Time Visual Understanding										
			OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	Avg.
Proprietary Models													
Gemini 1.5 Pro	–	1 fps	79.0	80.5	83.5	79.7	80.0	84.7	77.8	64.2	72.0	48.7	75.7
GPT-4o	–	64	77.1	80.5	83.9	76.5	70.2	83.8	66.7	62.2	69.1	49.2	73.3
Open-source Offline MLLMs													
LongVA-7B	–	128	70.0	63.3	61.2	70.9	62.7	59.5	61.1	53.7	54.7	34.7	60.0
LLaVA-OneVision-7B	–	32	80.4	74.2	76.0	80.7	72.7	71.7	67.6	65.5	65.7	45.1	71.1
InternVL2-8B	–	16	68.1	60.9	69.4	77.1	67.7	62.9	59.3	53.3	55.0	56.5	63.7
Open-source Online MLLMs													
VideoLLM-Online-8B	CVPR24	2 fps	39.1	40.1	34.5	31.1	46.0	32.4	31.5	34.2	42.5	27.9	36.0
Dispider-7B	CVPR25	1 fps	74.9	75.5	74.1	73.1	74.4	59.9	76.1	62.9	62.2	45.8	67.6
Flash-VStream-7B	ICCV25	1 fps	25.9	43.6	24.9	23.9	27.3	13.1	18.5	25.2	23.9	48.7	23.2
TimeChat-Online-7B	ACMMM25	1 fps	80.8	79.7	80.8	83.3	74.8	78.8	78.7	64.2	68.8	58.0	75.3
StreamForest-7B	NeurIPS25	1 fps	83.1	82.8	82.7	84.3	77.5	78.2	76.9	69.1	75.6	54.4	77.3
Within-Backbone													
Qwen2.5-VL-7B	–	1 fps	78.3	80.5	79.8	82.4	75.5	80.4	74.1	62.6	67.6	51.1	73.9
+ FluxMem	CVPR26	1 fps	80.2	81.1	81.4	85.3	78.0	83.8	80.6	65.9	69.6	**52.1**	76.4
+ HERMES (4k)	ACL26	1 fps	**83.7**	**81.3**	88.0	**87.5**	76.7	**86.6**	**82.4**	76.0	73.9	46.6	79.4
+ TRACE (Ours)	–	1 fps	**83.7**	76.6	**89.0**	**87.5**	**79.9**	83.8	**82.4**	**77.2**	**75.0**	48.4	**79.7**
Qwen3-VL-4B†	–	1 fps	81.6	76.6	88.6	87.2	81.8	85.7	83.3	72.0	71.0	44.7	78.3
+ TRACE (Ours)	–	1 fps	**86.2**	**78.9**	**93.4**	**88.1**	**82.4**	**87.9**	**85.2**	**79.3**	**77.8**	**49.5**	**82.3**

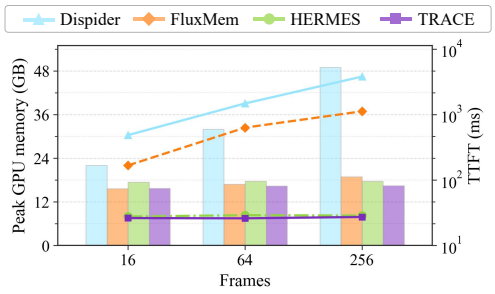


Figure 3: Efficiency comparison on Qwen2.5-VL-7B across different input lengths. Bars denote peak GPU memory (left axis), and lines denote TTFT (right axis, logarithmic scale).

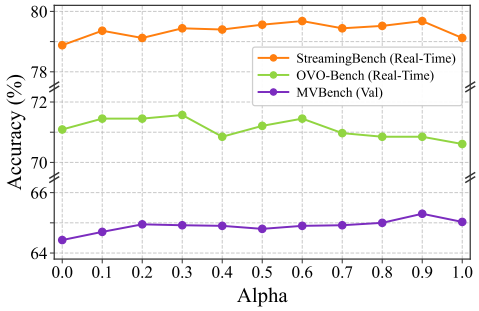


Figure 4: Effect of the aggregation weight α on StreamingBench, OVO-Bench, and MVBench. The y-axis is broken for readability.

4.2.2 Offline Video Evaluation

Table 3 reports the results on MVBench, MLVU, and VideoMME. On MVBench, TRACE increases the performance of Qwen2.5-VL-7B from 63.7 to 64.9, while the score of Qwen3-VL-4B decreases from 65.5 to 63.3. On the longer-video benchmarks MLVU and VideoMME, TRACE retains only 8.2% of the visual tokens while preserving 91.1% of the corresponding backbone performance on average. These results illustrate the trade-off between visual-memory compression and offline video understanding performance.

4.2.3 Efficiency Analysis

To evaluate the efficiency of TRACE built upon Qwen2.5-VL-7B, we measure peak GPU memory consumption and time-to-first-token (TTFT) under different input lengths. As shown in Figure 3, TRACE maintains stable GPU memory usage as the number of frames increases, owing to its bounded visual KV retention rather than continuously accumulating incoming visual states. Meanwhile, TTFT remains relatively stable across different input lengths, indicating that query-time latency does not increase substantially as the stream grows.

Table 3: Performance comparison on MVBench, MLVU, and VideoMME (without subtitles). The best results within each backbone group are shown in **bold**, and † denotes reproduced results.

Method	Venue	Frames	MVBench	MLVU	VideoMME			
			Val	M-Avg	Short	Medium	Long	Avg.
Open-source Offline MLLMs								
LongVA-7B	–	128	–	56.3	61.1	50.4	46.2	52.6
LLaVA-OneVision-7B	–	32	56.7	64.7	70.1	56.4	48.8	58.4
Open-source Online MLLMs								
Dispider-7B	CVPR25	1 fps	–	61.7	–	53.7	49.7	57.2
VideoChat-Online-8B	CVPR25	2 fps	64.9	–	–	–	44.9	52.8
Flash-VStream-7B	ICCV25	1 fps	–	–	72.0	61.1	50.3	61.2
StreamForest-7B	NeurIPS25	1 fps	68.6	69.6	–	–	–	61.9
LiveVLM-7B	DAC26	0.5/0.2 fps	–	66.3	66.7	56.4	48.8	57.3
Within-Backbone								
Qwen2.5-VL-7B†	–	1 fps	63.7	**67.9**	**73.8**	**62.4**	**53.8**	**63.3**
+ TRACE (Ours)	–	1 fps	**64.9**	62.1	72.6	56.0	45.9	58.1
Qwen3-VL-4B†	–	1 fps	**65.5**	**73.0**	72.6	**65.9**	**58.8**	**65.5**
+ TRACE (Ours)	–	1 fps	63.3	66.3	**73.2**	56.8	48.0	59.3

Table 4: Ablation of the hierarchical memory design using Qwen2.5-VL-7B.

Memory $\mathcal{M}_s$	**Memory** $\mathcal{M}_h$	OVO-Bench real-time	MVBench Val	VideoMME w/o sub.	Score Avg.
✓	✗	71.0	63.9	56.5	63.8
✗	✓	68.5	64.4	57.6	63.5
✓	✓	**71.5**	**64.9**	**58.1**	**64.8**

Table 5: Ablation of the position re-encoding strategies using Qwen2.5-VL-7B.

Method	OVO-Bench real-time	VideoMME w/o sub.	Score Avg.
TRACE	–	–	–
+ Contiguous RoPE	71.0	57.3	64.2
+ TGAP	**71.5**	**58.1**	**64.8**

4.3 Ablation Study

To systematically evaluate the effectiveness of each component in TRACE, we conduct ablation studies from three perspectives: hierarchical memory design, position re-encoding strategy, and spatiotemporal score aggregation. Additional ablations are provided in Appendix C.

Hierarchical Memory Design. We ablate each memory bank to evaluate its contribution. As shown in Table 4, combining both memories achieves the best performance compared with using either memory alone. This result suggests that the full-granularity short-term memory and sparse historical memory provide complementary information for streaming video understanding.

Position Re-Encoding Strategy. As shown in Table 5, TGAP consistently outperforms Contiguous RoPE (Xu et al., 2025). This result suggests that preserving compressed temporal gaps between retained frames is more effective than contiguous re-indexing under the evaluated settings.

Score Aggregation Mechanism. We vary α from 0 to 1 to examine the contributions of temporal and spatial scores. As shown in Figure 4, combining temporal novelty and spatial distinctiveness generally outperforms using either signal alone. This result supports our design choice of jointly considering the two signals for visual-token retention.

5 Conclusion

We present TRACE, a training-free framework for novelty-aware spatiotemporal evidence retention in streaming video understanding. TRACE combines a causal hierarchical visual memory with joint temporal novelty and spatial distinctiveness scoring to retain informative visual evidence under a bounded memory budget, while TGAP preserves coarse temporal structure over long streams. Experiments across different VLM backbones show improved streaming video understanding with substantial visual-token reduction, bounded visual KV memory, and stable query-time latency.

AI USE STATEMENT

In this work, we used generative AI tools to improve the grammar and readability of author-written English text and to assist with writing parts of the code used in this work. All AI-assisted text and code were reviewed by the authors, and the code was tested for correctness through the reported experiments. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The supplementary material includes a runnable demo of TRACE containing the core model implementation, together with a video demonstrating its usage.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Daniel Bolya, Cheng-Yang Fu, Xiaoli Dai, Judy Hoffman, and Christoph Feichtenhofer. Token merging: Your vit but faster. In *International Conference on Learning Representations*, 2023.

Jiajun Cao, Qizhe Zhang, Peidong Jia, Xuhui Zhao, Bo Lan, Xiaoan Zhang, Xiaobao Wei, Sixiang Chen, Liyun Li, Xianming Liu, et al. Fastdrivevla: Efficient end-to-end driving via plug-and-play reconstruction-based token pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2571–2579, 2026.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Live: Learning video llm with streaming speech transcription at scale. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29083–29095. IEEE, 2025.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, volume 2024, pp. 35549–35562, 2024.

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.

Zhengyao Fang, Pengyuan Lyu, Chengquan Zhang, Guangming Lu, Jun Yu, and Wenjie Pei. Prune redundancy, preserve essence: Vision token compression in vlms via synergistic importance-diversity. In *International Conference on Learning Representations*, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Hong Gao, Yiming Bao, Xuezhen Tu, Bin Zhong, Linan Yue, and Min-Ling Zhang. Apvr: Hour-level long video understanding with adaptive pivot visual information retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4113–4121, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19959–19973, 2025a.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: Ovbench and videochat-online. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3328–3338. IEEE, 2025b.

Huawei Ji, Yuanhao Sun, Yuan Jin, Cheng Deng, Jiaxin Ding, Luoyi Fu, and Xinbing Wang. Vispco: Visual token pruning configuration optimization via budget-aware pareto-frontier learning for vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 9281–9301, 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*, 2024a.

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

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*, pp. 12147–12151. IEEE, 2026.

Zhenyang Liu, Yongchong Gu, Yikai Wang, Xiangyang Xue, and Yanwei Fu. Activevla: Injecting active perception into vision-language-action models for precise 3d robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8141–8151, 2026.

Xudong Lu, Guan Huankang, Yang Bo, Jinpeng Chen, Xintong Guo, Shuhan Li, Fang Liu, Peiwen Sun, Xueying Li, Wei Zhang, et al. Phostream: Benchmarking real-world streaming for omnimodal assistants in mobile scenarios. In *Forty-third International Conference on Machine Learning*, 2026.

Bingjun Luo, Tony Wang, Chaoqi Chen, and Xinpeng Ding. St-simdiff: Balancing spatiotemporal similarity and difference for efficient video understanding with mllms. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ke Ma, Jiaqi Tang, Bin Guo, Xueting Han, Ruonan Xu, Qingfeng He, Ziheng Wang, Xu Wang, Qifeng Chen, Zhiwen Yu, et al. Response-g1: Explicit scene graph modeling for proactive streaming video understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 44139–44153, 2026.

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

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

Yaxuan Qin, Hefei Li, Wenqi Mu, and Yancheng He. Efficient frame selection for long video understanding via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16944–16953, 2026.

Peiwen Sun, Xudong Lu, Huadai Liu, Yang Bo, Dongming Wu, Huankang Guan, Minghong Cai, Jinpeng Chen, Xintong Guo, Shuhan Li, et al. X-stream: Exploring mllms as multiplexers for multi-stream understanding. *arXiv preprint arXiv:2606.02482*, 2026.

Chao Wang, Xudong Tan, Jianjian Cao, Kangcong Li, and Tao Chen. Curvestream: Boosting streaming video understanding in mllms via curvature-aware hierarchical visual memory management. *arXiv preprint arXiv:2603.19571*, 2026a.

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*, 2025.

Yueqian Wang, Songxiang Liu, Disong Wang, Nuo Xu, Guanglu Wan, Huishuai Zhang, and Dongyan Zhao. Mmduet2: Enhancing proactive interaction of video mllms with multi-turn reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 23257–23270, 2026b.

Hang Wu, Sherin Mary Mathews, Yujun Cai, Ming-Hsuan Yang, and Yiwei Wang. Semantic-aware adaptive visual memory for streaming video understanding. *arXiv preprint arXiv:2605.07897*, 2026.

Junbin Xiao, Jiajun Chen, Tianxiang Sun, Xun Yang, and Angela Yao. Mukv: Multi-grained kv cache compression for long streaming video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11381–11391, June 2026.

Yiweng Xie, Bo He, Junke Wang, Xiangyu Zheng, Ziyi Ye, and Zuxuan Wu. Fluxmem: Adaptive hierarchical memory for streaming video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31272–31282, 2026.

Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. In *International Conference on Learning Representations*, volume 2025, pp. 69332–69351, 2025.

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

Weicai Yan, Yuhong Dai, Qi Ran, Haodong Li, Wang Lin, Hao Liao, Xing Xie, Tao Jin, and Jianxun Lian. Proact-vl: A proactive videollm for real-time ai companions. In *Proceedings of the International Conference on Machine Learning*, 2026.

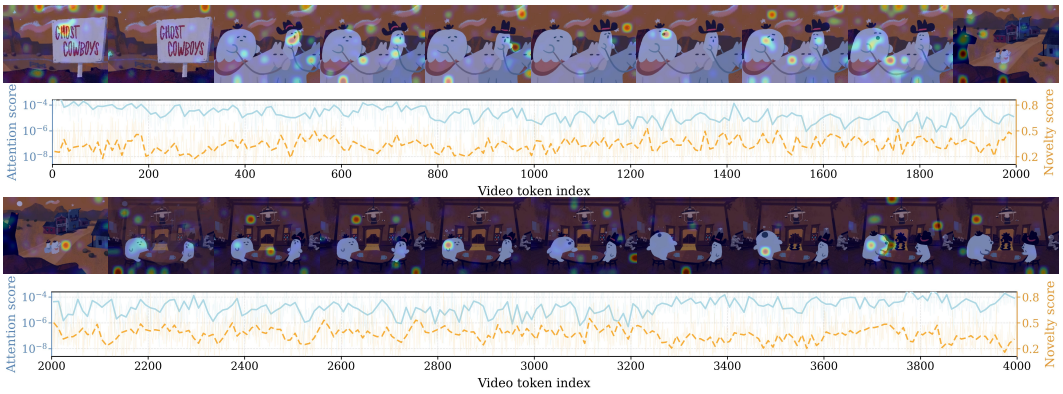
Zhenyu Yang, Kairui Zhang, Yuhang Hu, Bing Wang, Shengsheng Qian, Bin Wen, Fan Yang, Tingting Gao, Weiming Dong, and Changsheng Xu. Livestar: Live streaming assistant for real-world online video understanding. *Advances in Neural Information Processing Systems*, 38:31266–31304, 2026.

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.

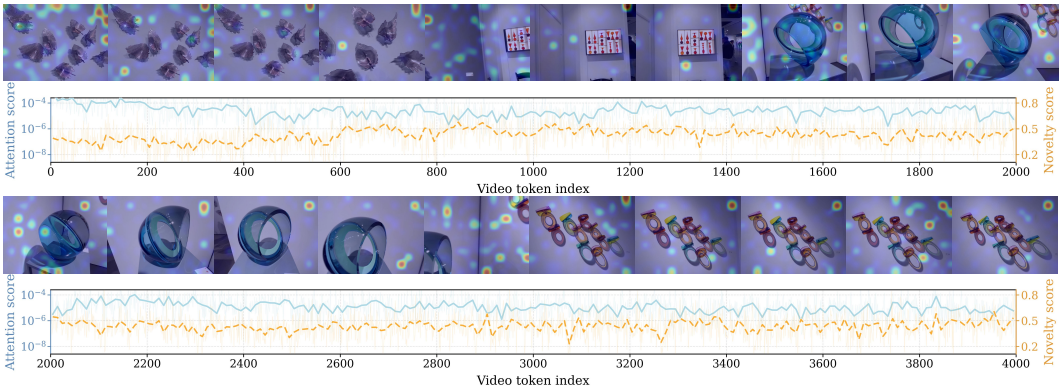
Haowei Zhang, Shudong Yang, Jinlan Fu, See Kiong Ng, and Xipeng Qiu. Hermes: Kv cache as hierarchical memory for efficient streaming video understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 8411–8430, 2026.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025.

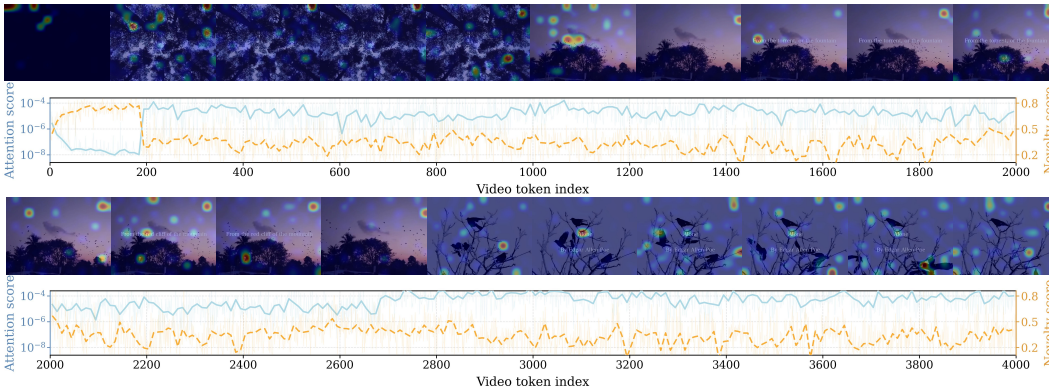
A Additional Analysis of Attention and Visual Novelty



(a) Art Exhibition

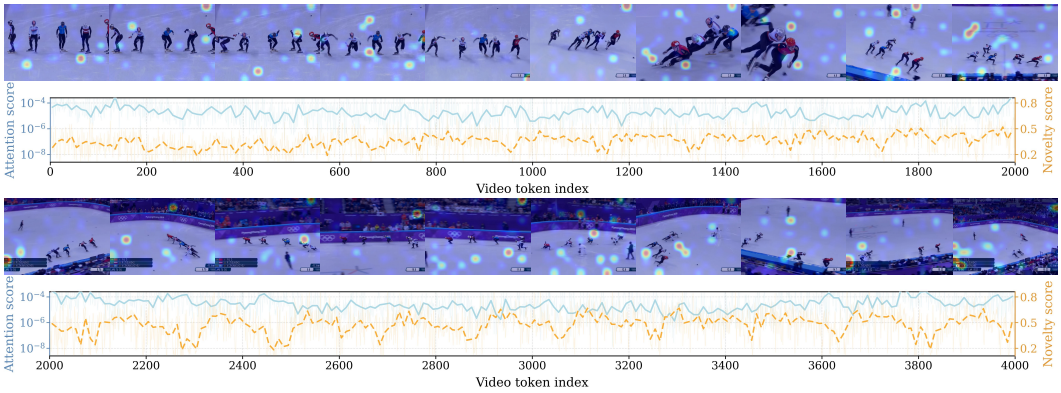


(b) Film & Animation

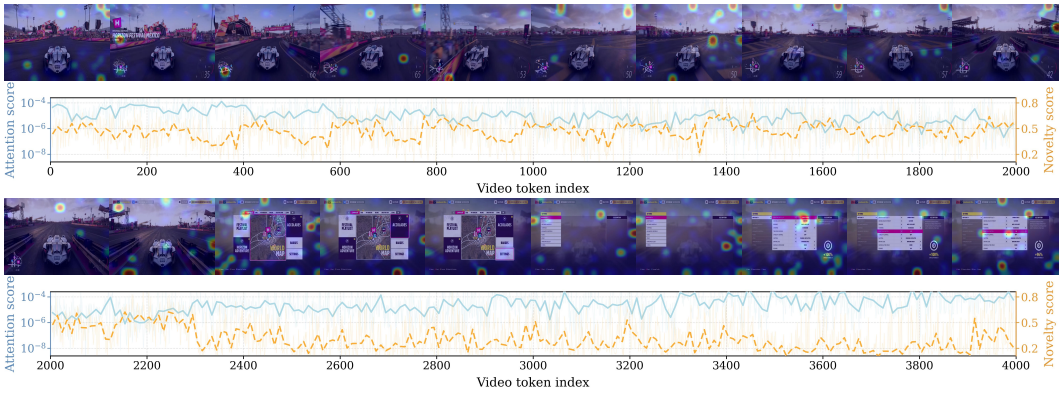


(c) Natural Landscapes

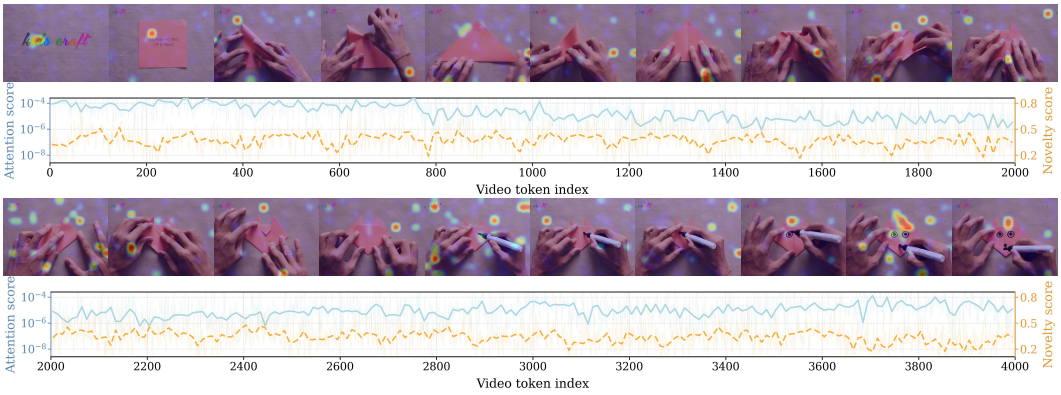
Figure 5: Additional analysis of the relationship between attention and visual novelty in Qwen2.5-VL-7B across different VideoMME categories.



(d) Sports Competition



(e) Game Play



(f) How-To Guides

Figure 5: Additional analysis of the relationship between attention and visual novelty in Qwen2.5-VL-7B across different VideoMME categories (continued).

To further examine the relationship between attention and visual novelty across diverse video content, we provide additional visualizations for six videos from different categories in VideoMME (Fu et al., 2025). These videos exhibit substantially different temporal dynamics and motion intensities, ranging from relatively static scenes with gradual appearance changes to fast-changing content involving frequent object motion and scene transitions. The diversity of these examples allows us to examine whether the observed attention-novelty mismatch persists across different types of video content.

As illustrated in Fig. 5, similar discrepancies between attention and visual novelty can be observed across these examples. In relatively static segments, attention often remains concentrated on similar

visual regions even when consecutive observations introduce little new visual information. In more dynamic videos, changes in visual content are not always accompanied by proportionally higher attention. Consequently, highly attended tokens do not necessarily correspond to visually novel observations, and visual changes can also occur at locations receiving only moderate attention.

Importantly, this does not imply that attention is uninformative for model inference. Rather, the two signals characterize different properties of visual tokens. Attention primarily reflects their role in the model's current computation, whereas visual novelty characterizes the new visual evidence introduced relative to neighboring observations. In query-triggered streaming video understanding, retention decisions must be made before future queries are known. Therefore, current attention does not necessarily coincide with the visual evidence that should be preserved for unknown future queries. This distinction motivates the use of visual novelty as a query-agnostic retention signal rather than relying solely on attention-based importance.

B Additional Method Details

B.1 Transition to Historical Retention

Suppose frame f_τ is about to leave the short-term memory at streaming step t. The visual-token indices of the outgoing frame are

$$\mathcal{O}_\tau = \{(\tau, p) \mid 1 \le p \le P_\tau\}. \tag{10}$$

Let $\mathcal{I}_{t-1}$ denote the indices already retained in historical memory. Before updating $\mathcal{M}_h$, the historical candidate set is formed as

$$\mathcal{C}_t = \mathcal{I}_{t-1} \cup \mathcal{O}_\tau. \tag{11}$$

For the outgoing tokens in $\mathcal{O}_\tau$, their temporal novelty scores have already been finalized within the short-term observation window. The temporal scores of previously retained historical tokens are directly reused. In contrast, spatial distinctiveness is recomputed over the current candidate set $\mathcal{C}_t$, since the sparse within-frame neighborhood may change as historical tokens are removed or retained.

STAR then rank-normalizes the temporal and spatial scores and computes the joint retention score $S^t_{\mathrm{STAR}}(i)$ for each candidate $i \in \mathcal{C}_t$ according to Sec. 3.4. If $|\mathcal{C}_t| \le C^t_h$, all candidates are retained. Otherwise, the C^t_h candidates with the highest joint scores are selected to form the updated historical index set $\mathcal{I}_t$. The historical memory $\mathcal{M}^t_h$ is subsequently updated according to Eq. 2.

Since the outgoing frame is evaluated only after the observations required to finalize its temporal novelty have arrived, all quantities used in this transition depend only on information available up to streaming step t. The update therefore preserves the causal streaming constraint.

B.2 TGAP Implementation

Given the compressed temporal gaps defined in Eq. (9), the re-encoded temporal coordinates are constructed cumulatively:

$$\widetilde{t}_1 = 0, \qquad \widetilde{t}_j = \widetilde{t}_{j-1} + \widetilde{\Delta}_j, \quad j = 2, \dots, J. \tag{12}$$

For a retained visual token p from frame t_j, its original and re-encoded multimodal positions are

$$\boldsymbol{\pi}_{j,p} = (t_j, h_p, w_p), \qquad \widetilde{\boldsymbol{\pi}}_{j,p} = (\widetilde{t}_j, h_p, w_p), \tag{13}$$

where (h_p, w_p) denotes its intra-frame spatial coordinates. TGAP modifies only the temporal coordinate, leaving the intra-frame spatial organization unchanged.

To apply this coordinate change to the KV cache, we re-encode the retained visual keys rather than merely update their position indices, since these keys have already been rotary-encoded at their original positions. Let $K^{(l)}_{t_j,p}$ denote the cached key of token p from frame t_j at transformer layer l, and let $\mathbf{R}^{(l)}(\boldsymbol{\pi})$ denote the rotary transformation associated with multimodal position $\boldsymbol{\pi}$. TGAP re-encodes the cached key from its original position $\boldsymbol{\pi}_{j,p}$ to the compressed position $\widetilde{\boldsymbol{\pi}}_{j,p}$ as

$$\widetilde{K}^{(l)}_{t_j,p} = \mathbf{R}^{(l)}(\widetilde{\boldsymbol{\pi}}_{j,p})\mathbf{R}^{(l)}(\boldsymbol{\pi}_{j,p})^{-1} K^{(l)}_{t_j,p}. \tag{14}$$

The re-encoding is applied to the retained visual keys at all transformer layers, including the selected historical keys and all keys in the short-term memory. The corresponding value states remain unchanged because TGAP only modifies the positional encoding of the cached keys. TGAP is triggered when the cumulative number of input tokens reaches a predefined threshold. At the trigger point, all currently retained visual keys are re-encoded using their compressed temporal coordinates, after which subsequent streaming inference proceeds in the compressed temporal coordinate space. This operation changes only the positional representation of the retained keys and does not alter their retention status.

C Additional Ablation Studies

C.1 Scoring Criteria for Visual-Token Retention

Table 6: Comparison of scoring criteria for visual-token retention. Best results are shown in bold.

Scoring Criteria	OVO-Bench real-time	MVBench Val	VideoMME w/o sub.	Score Avg.
Attention	64.6	61.9	51.9	59.5
Value	68.8	**65.2**	**59.2**	64.4
Pre-RoPE Key	71.3	64.9	57.7	64.6
Post-RoPE Key	**71.5**	64.9	58.1	**64.8**

We compare four scoring bases for visual-token retention in Table 6. Attention directly uses attention scores for token selection, whereas Value, Pre-RoPE Key, and Post-RoPE Key compute temporal novelty and spatial distinctiveness from their respective representations. All variants use the same memory budget and retention configuration, differing only in the signal used to score candidate visual tokens.

All three novelty-based variants consistently outperform attention-based selection across the evaluated benchmarks. In terms of the average score, Value, Pre-RoPE Key, and Post-RoPE Key achieve 64.4, 64.6, and 64.8, respectively, compared with 59.5 for attention. This result supports our use of representation-based visual novelty rather than attention as the primary criterion for query-agnostic evidence retention.

Among the novelty-based variants, no single representation is uniformly best on every benchmark. Value features obtain the highest scores on MVBench and VideoMME, whereas key-based representations perform better on OVO-Bench. Pre-RoPE and Post-RoPE keys show similar overall performance, with average scores of 64.6 and 64.8, respectively. We therefore adopt Post-RoPE keys as the default scoring representation in TRACE, as they achieve the highest average score among the evaluated alternatives and can be directly reused from the visual KV states maintained during streaming inference.

C.2 Joint vs. Sequential Spatiotemporal Retention

Table 7: Comparison of joint and sequential spatiotemporal retention.

Strategy	OVO-Bench real-time	MVBench Val	VideoMME w/o sub.	Score Avg.
Temporal $\rightarrow$ Spatial	71.3	64.7	57.3	64.4
Spatial $\rightarrow$ Temporal	71.2	64.7	57.1	64.3
Joint	**71.5**	**64.9**	**58.1**	**64.8**

We compare joint spatiotemporal retention with temporal-first and spatial-first sequential variants within TRACE. In the sequential variants, one criterion is applied before the other, so candidates removed at the first stage cannot be reconsidered using the complementary signal. In contrast, the joint strategy evaluates temporal novelty and spatial distinctiveness together before making the retention decision.

As shown in Table 7, the joint strategy achieves the best performance on all three benchmarks, with an average score of 64.8, compared with 64.4 for Temporal-to-Spatial and 64.3 for Spatial-to-Temporal retention. Although the margins are modest, the consistent improvements support our design choice of jointly evaluating temporal novelty and spatial distinctiveness before removing any candidate visual tokens. Joint evaluation reduces the risk of prematurely discarding evidence that appears redundant according to one signal but remains informative according to the other.

C.3 Layer Selection for Score Computation

Table 8: Ablation of the anchor layer for novelty scoring.

Layer Index	OVO-Bench real-time	MVBench Val	VideoMME w/o sub.	Score Avg.
0	69.5	65.1	**58.6**	64.4
14	70.0	**65.2**	58.4	64.5
27	**71.5**	64.9	58.1	**64.8**

We compare early, intermediate, and final transformer layers as the anchor layer for novelty scoring. As shown in Table 8, the choice of anchor layer leads to relatively small differences on MVBench and VideoMME. The intermediate layer achieves the highest MVBench score, while the early layer performs best on VideoMME. In contrast, the final layer obtains the highest score on OVO-Bench and achieves the highest overall average score of 64.8. Based on this overall performance, we use the final transformer layer as the default anchor layer for novelty scoring in TRACE.

C.4 Spatial Distinctiveness Scoring Strategies

Table 9: Different strategies for spatial distinctiveness scoring.

Strategy	OVO-Bench real-time	MVBench Val	VideoMME w/o sub.	Score Avg.
TRACE	–	–	–	–
+ Adjacent-k	71.2	64.7	57.6	64.5
+ Nearest-k	**71.5**	**64.9**	**58.1**	**64.8**

We compare Adjacent-k and Nearest-k for spatial distinctiveness scoring. Nearest-k follows the neighborhood construction in Sec. 3.3, using the nearest available candidate from the same frame in each cardinal direction. As shown in Table 9, these modest improvements support the use of Nearest-k on the evaluated comparisons. This strategy allows spatial distinctiveness to be evaluated against available visual evidence even when immediately adjacent tokens have been removed from historical memory.

C.5 Sensitivity to Memory Window and Budget

Table 10: Sensitivity to the short-term observation window length N and total visual-token budget C. Best results within each parameter sweep are shown in bold.

Capacity		OVO-Bench	MVBench	VideoMME	Score
N	C	real-time	Val	w/o sub.	Avg.
2	5120	67.0	**65.0**	57.8	63.3
6	5120	**71.5**	64.9	**58.1**	**64.8**
10	5120	71.2	64.5	57.3	64.3
6	2560	**73.1**	63.6	55.1	63.9
6	5120	71.5	64.9	58.1	**64.8**
6	10240	68.5	**65.4**	**59.7**	64.5

We examine the effects of the short-term observation window length N and the total visual-token budget C. As shown in Table 10, with the total budget fixed at $C = 5120$, the default six-frame window achieves the highest average score among the tested window lengths. A shorter window ($N = 2$) reduces performance on OVO-Bench, while increasing the window from six to ten frames does not provide further gains. Since a larger short-term window preserves more recent observations at full granularity but correspondingly leaves less capacity for historical memory, these results indicate that the allocation between recent full-resolution observations and sparse historical evidence needs to be balanced under a fixed memory budget.

We further vary the total visual-token budget while fixing $N = 6$. Increasing C from 2560 to 10240 improves performance on MVBench and VideoMME, whereas OVO-Bench shows the opposite trend. This result indicates that increasing the retained visual context does not uniformly improve all evaluation settings: longer-video benchmarks benefit from a larger memory budget, while the streaming OVO-Bench setting achieves its highest score with the more compact configuration. Among the evaluated settings, $N = 6$ and $C = 5120$ achieves the highest average score of 64.8, and is therefore adopted as the default configuration in TRACE.

D Scalability to Extended Video Streams

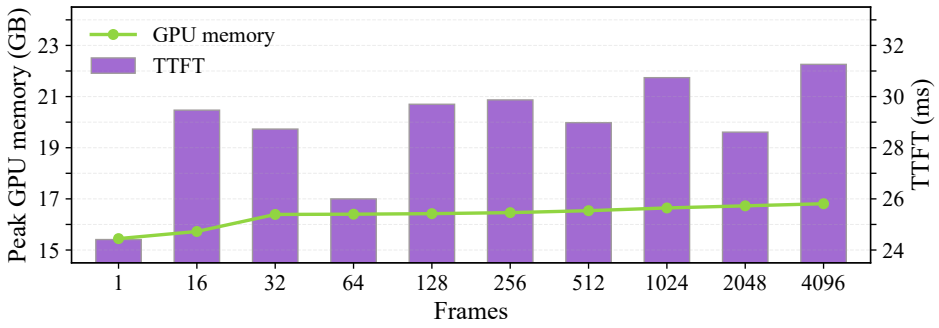


Figure 6: Scalability of TRACE over extended video streams. Peak GPU memory grows only modestly and TTFT shows no consistent upward trend as the stream increases to 4,096 frames, supporting stable resource usage over hour-scale streams.

We further examine the resource behavior of TRACE by processing video streams sampled at 1 fps under a fixed memory configuration. Figure 6 reports peak GPU memory and TTFT as the number of processed visual observations increases from short sequences to 4,096 frames. At 1 fps, the longest evaluated stream corresponds to approximately 68 minutes of video.

Across the evaluated durations, GPU memory usage grows only modestly, while TTFT shows no consistent upward trend as the stream becomes longer. Even at 4,096 frames, the resource footprint remains comparable to that observed for substantially shorter streams. This behavior is consistent with TRACE maintaining a bounded visual KV memory rather than continuously accumulating all incoming visual states as the stream grows. These results support the scalability of TRACE to hour-scale streaming video inference.