BEYOND SALIENCY: TOKEN MEMORY VIA MARGINAL CONTRIBUTION FOR STREAMING VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming video understanding requires processing an unbounded video stream within bounded memory while answering queries whose timing and content are unknown in advance. Because evicted tokens cannot be recovered, memory management should preserve information not already covered by the current memory rather than simply retain individually salient tokens. Existing saliency-based methods do not explicitly model such redundancy, while age-stratified methods constrain competition across temporal tiers. We propose SMART, a training-free, query-independent streaming token memory that retains tokens according to their marginal contribution beyond what the current memory already covers. SMART groups consecutive frames into content-based events to distinguish within-event redundancy from observations across events. Within each event, cross-frame nearest-match weights estimate how much a token is already covered by other retained frames, and modulate an intra-frame determinantal point process (DPP) that measures token complementarity. Because marginal contribution changes as memory evolves, SMART rescores surviving tokens whenever a new frame arrives. An age-augmented DPP kernel further combines contribution with recency for unprotected tokens, while preserving the most recent frame for real-time perception. With Qwen3-VL-8B, SMART achieves 68.6% OVO-Avg and 82.4% accuracy on StreamingBench, the highest aggregate results among the streaming methods evaluated.

1 INTRODUCTION

Streaming video understanding arises in first-person assistants (Grauman et al., 2022; Damen et al., 2018), surveillance (Sultani et al., 2018), and autonomous driving (Caesar et al., 2020; Geiger et al., 2012; Sun et al., 2020), where video is not a fixed recording but a continuously growing stream. A streaming system answers queries about past and present content at any time while keeping its visual token memory within a fixed memory capacity. As the stream grows, the memory overflows and the system must evict tokens, which cannot be recovered afterwards. Because future queries are unknown, the system must decide in advance which tokens to retain.

The retention decision therefore depends less on how salient a token is than on whether it provides information not already covered by the current memory. A person standing at a doorway for ten consecutive frames produces many individually salient yet near-duplicate tokens, whereas the same person appearing again later provides new information that should be retained. Existing streaming methods mainly organize memory around temporal recency or content scores: sliding windows discard all content outside a fixed horizon (Xu et al., 2026); age-stratified methods compress older tiers more aggressively (Xie et al., 2026; Liu et al., 2026); and other methods use attention weights (Zhang et al., 2026a), Value norms (Kim et al., 2025), or similarity-based event merging (Zeng et al., 2025). Attention weights and Value norms measure individual saliency, which does not reflect whether a token is already covered by the current memory, so redundant tokens occupy capacity while tokens with new information are evicted, as Figure 1 illustrates. Similarity alone does not determine redundancy either: similar tokens within the same event repeat the same information, whereas similar tokens from different events record separate occurrences. Age-stratified methods further rank tokens only within the same tier, so an old token cannot compete directly against a new arrival, even when

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.

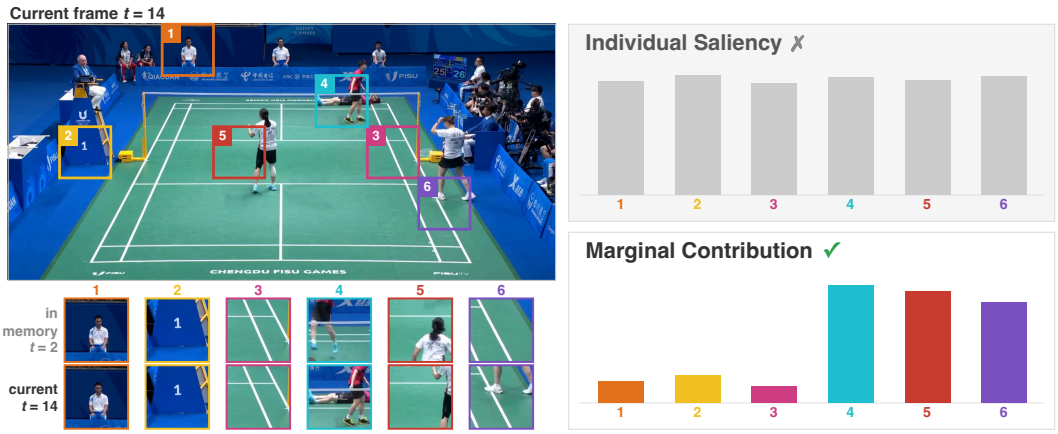


Figure 1: Individual saliency versus marginal contribution. Six token regions are annotated in the current frame at t=14 and compared against a memory frame at t=2. Tokens 1–3 already have near-duplicates in memory, and tokens 4–6 carry new information. Individual saliency assigns comparable scores to all six and cannot separate redundant tokens from new ones. Marginal contribution scores tokens 1–3 low and tokens 4–6 high.

the old token carries more information. These limitations motivate a different retention principle: each token should be retained according to its marginal contribution, the information it adds beyond what the current memory already covers.

We propose SMART, a training-free, query-independent streaming visual token memory that realizes this principle by addressing both the scope and the dynamics of redundancy. For the scope, SMART groups consecutive frames into content-based events, and only similarity within an event counts as redundancy. Within each event, a cross-frame deduplication weight estimates how much of each token is already covered by other frames in memory. This weight scales the token in the kernel of an intra-frame determinantal point process (DPP) (Kulesza & Taskar, 2012), whose greedy ordering measures complementarity among tokens of the same frame and assigns each token its marginal contribution score. For the dynamics, because marginal contribution depends on the current memory, SMART rescores all surviving tokens whenever a new frame arrives. Finally, SMART keeps the most recent frame in full for real-time perception. For all other tokens, an age-augmented DPP kernel embeds exponential age discounting into the quality term, so tokens of all ages compete in a single eviction pool by marginal contribution and recency. We prove that the resulting time-discounted score equals the conditional variance of each token under greedy ordering on the augmented kernel, so the joint weighting of recency and marginal contribution corresponds to greedy ordering on a single DPP kernel.

Our contributions are:

- We propose SMART, a training-free streaming visual token memory that retains tokens by their marginal contribution to the current memory. Event-aware cross-frame deduplication and intra-frame DPP complementarity jointly determine this contribution, and SMART recomputes it at every frame arrival.
- We construct an age-augmented DPP kernel that embeds exponential age discounting into the quality term, and we prove that the resulting time-discounted score equals the conditional variance of each token under greedy ordering on the augmented kernel.
- We evaluate SMART on OVO-Bench and StreamingBench with Qwen2.5-VL-7B and Qwen3-VL-8B backbones. On both backbones, SMART achieves the highest OVO-Avg and StreamingBench accuracy among the evaluated streaming methods, reaching 68.6% and 82.4% with Qwen3-VL-8B. Removing event-aware deduplication, marginal contribution scoring, age discounting, or recent-frame protection lowers both metrics on both backbones.

2 Related Work

2.1 Multimodal Large Language Models

Multimodal large language models (MLLMs) pair visual encoders (Radford et al., 2021; Dosovitskiy et al., 2021) with large language models to reason over image and video content. Image-oriented

MLLMs such as LLaVA (Liu et al., 2023) and InternVL (Chen et al., 2024c) convert images into visual token sequences and feed them alongside text into the language model. Video-oriented MLLMs such as Qwen2-VL (Wang et al., 2024), Qwen2.5-VL (Bai et al., 2025), LLaVA-OneVision (Li et al., 2025a), LLaVA-Video (Zhang et al., 2025c), LongVU (Shen et al., 2025), and VideoLLaMA (Zhang et al., 2023a) sample multiple frames, encode them, and concatenate the resulting tokens into a long sequence. As video length grows, the token count increases linearly, straining both the context window and GPU memory of attention-based (Vaswani et al., 2017) language models. These methods assume the entire video is available before inference and control token count through uniform sampling or a fixed frame budget. Neither strategy applies to real-time streams, where video arrives continuously and the total frame count is unknown. A streaming system therefore manages visual tokens dynamically within a bounded memory capacity.

2.2 Streaming Video Understanding

Streaming video understanding requires systems to process unbounded video within a fixed memory capacity while responding to queries at any time (Chen et al., 2024a; Zhang et al., 2025a; Qian et al., 2025; Yao et al., 2025; Xia et al., 2026; Zhang et al., 2026a). Early work bounded memory through dual buffers of short-term and long-term memory (Song et al., 2024; He et al., 2024) or attention-sink mechanisms (Xiao et al., 2024). Subsequent methods introduced four kinds of eviction strategies: age-stratified quotas (Xie et al., 2026; Liu et al., 2026; Wang et al., 2025a) divide memory into time segments with fixed slot allocations, and sliding windows (Xu et al., 2026) keep only the most recent segment; per-frame quotas (Chen et al., 2026; Dorovatas et al., 2025) retain a fixed number of tokens per frame; guard-band methods (Han et al., 2026) unconditionally preserve recent frames and rank the rest by content scores; weighted-sum methods (Zeng et al., 2025) linearly combine recency with similarity-based event-merging scores. Age-stratified and per-frame quotas fix the capacity of each time segment or frame, so old tokens are compressed or removed regardless of whether their content is unique. Guard-band and weighted-sum methods rank tokens by content scores, but these scores measure individual saliency or similarity rather than marginal contribution: saliency does not reflect what the current memory already covers, and similarity alone does not separate repetition within an event from recurrence across events. Recency enters these methods either only through the preserved recent frames or as a separately weighted term. Like guard-band methods, SMART protects the most recent frame, but it scores all other tokens by their marginal contribution to the current memory, treats similarity as redundancy only within the same event, and embeds age discounting into the same DPP kernel, so tokens of all ages compete directly in a single eviction pool.

2.3 Visual Token Reduction

Visual token reduction methods select or compress tokens to control computational cost. Training-time approaches introduce merging or pruning during model training (Bolya et al., 2023; Shang et al., 2025). Training-free approaches select tokens at inference time: attention-magnitude methods (Zhang et al., 2023b; Li et al., 2024; Chen et al., 2024b; Dorovatas et al., 2025; Wang et al., 2025b; Zhang et al., 2025d) score tokens by the attention weights they receive; Value-norm methods (Kim et al., 2025) score tokens by the norm of the Value vector; similarity-based methods (Han et al., 2026) score tokens by average similarity to other memory entries. Video-oriented compression (Tao et al., 2025) further exploits temporal redundancy to merge repeated tokens across frames during decoding. Most of these measures evaluate each token independently. CDPruner (Zhang et al., 2025b) instead uses a DPP to optimize conditional diversity within a single image. Apart from the streaming methods discussed above (Dorovatas et al., 2025; Han et al., 2026), these methods target fixed images or offline video and do not address the streaming setting, where evicted tokens cannot be recovered, content arrives incrementally, and temporal decay is needed to balance recency against marginal contribution.

3 Method

SMART estimates each token's marginal contribution to the current memory and weighs it jointly with recency through three components, shown in Figure 2. Adaptive event segmentation groups consecutive frames into events by content similarity and defines the scope of redundancy, so only similarity within an event counts as redundancy (§3.2). Marginal contribution scoring modulates intra-frame DPP greedy ordering with a cross-frame deduplication weight computed within each

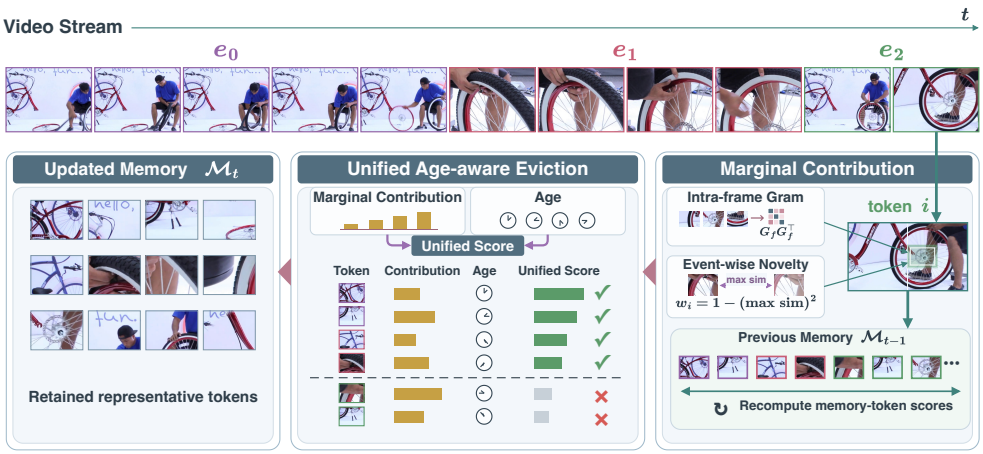


Figure 2: SMART framework. As frames arrive, adaptive event segmentation assigns event labels, intra-frame DPP greedy ordering modulated by deduplication weights within each event computes each token's marginal contribution score s_i, and exponential age discounting turns s_i into the time-discounted score κ_i that decides eviction.

event, and rescores all surviving tokens whenever a new frame arrives (§3.3). The eviction stage protects the most recent frame and uses an age-augmented DPP kernel to combine the scores of all other tokens with exponential age discounting into a time-discounted score κ_i, so tokens of all ages compete directly in a single eviction pool (§3.4).

3.1 Problem Setup

The video stream arrives frame by frame. A visual encoder maps frame t to n unit-normalized tokens arranged as $G_t \in \mathbb{R}^{n \times d}$, where row i, denoted $f_i \in \mathbb{R}^d$, is the feature vector of token i. The system maintains a memory $\mathcal{M}$ with fixed capacity $|\mathcal{M}| \leq B$. Whenever a new frame causes overflow, tokens are evicted until the constraint is restored, and an evicted token cannot be recovered. Queries arrive at unknown times with unknown content, so eviction decisions cannot depend on the query.

3.2 Adaptive Event Segmentation

Consecutive frames from the same scene share many near-duplicate tokens, and retaining all of them wastes capacity. Tokens from different scenes, however, record separate occurrences that should each be retained, even when their features are similar. To distinguish these two cases, the system groups consecutive frames into events by content similarity, and only similarity within an event counts as redundancy.

A frame belongs to the current event when most of its tokens find a close match among the event's tokens in memory. When this match quality drops, a new event begins. Let e denote the current open event and C_e the set of all tokens in memory belonging to event e. For each token i in the arriving frame t, we measure its coverage as the cosine similarity to the closest token in C_e:

$$m_i = \max_{j \in C_e} G_t[i]^\top f_j \tag{1}$$

A value m_i near 1 means token i has a near-duplicate in the current event, and m_i near 0 means it carries unseen content. The frame-level match $\phi_t = \text{median}_i\ m_i$ aggregates token-level coverage into a single scalar. The event assignment rule is

$$e_t = \begin{cases} e, & \phi_t \geq \theta \text{ or } t - b < L_{\min} \\ e + 1, & \text{otherwise (and set } b \leftarrow t) \end{cases} \tag{2}$$

where $\theta \in (0, 1)$ is the similarity threshold, b is the starting frame of the current event, and $L_{\min}$ is the minimum event length that prevents over-segmentation from inter-frame noise. Event labels are assigned causally at frame arrival and are never retroactively modified.

3.3 Marginal Contribution Scoring

Eviction requires measuring each token's marginal contribution to the current memory, that is, the information it adds beyond what the current memory already covers. As illustrated in Figure 1,

individual saliency is a property of the token alone, whereas marginal contribution depends on the token's relationship with other tokens in memory. We build the scoring function on determinantal point processes (DPPs) (Kulesza & Taskar, 2012). A DPP defines a distribution over subsets of a finite set via a positive semidefinite kernel $L \in \mathbb{R}^{N\times N}$, with $P(S) \propto \det(L_S)$: the diagonal entries L_{ii} encode individual quality, the off-diagonal entries L_{ij} encode pairwise similarity, and $\det(L_S)$ is large when S combines high quality with low redundancy. This property lets the DPP kernel measure within-set redundancy: given L and a reference set, the conditional variance derived from the kernel quantifies how much new information each token contributes. We next define the kernel and then describe how to extract per-token scores.

The kernel L determines which tokens are redundant and which are unique. We define L as a block-diagonal matrix indexed by frames:

$$L = \bigoplus_f \tilde{L}_f \tag{3}$$

where $\bigoplus_f$ denotes the direct sum, i.e., block-diagonal concatenation, so tokens from different frames have no cross-terms. The intra-frame kernel $\tilde{L}_f \in \mathbb{R}^{n\times n}$ is

$$\tilde{L}_f = \mathrm{diag}(\sqrt{w})\, G_f G_f^\top \,\mathrm{diag}(\sqrt{w}) \tag{4}$$

Here $G_f G_f^\top$ is the Gram matrix of frame f's tokens. Because features are unit-normalized, entry (i, j) equals the cosine similarity $f_i^\top f_j$. The off-diagonal entries encode intra-frame similarity: selecting two similar tokens reduces the determinant. The deduplication weight $w_i \in [0, 1]$ is computed from the nearest match of token i among other frames of the same event, measures how much token i differs from that content, and scales the token's effective norm in the kernel:

$$w_i = 1 - \Big(\max_{j\in C_f} f_i^\top f_j\Big)^2, \qquad C_f = \{j \in \mathcal{M} : e_j = e_f,\ j \notin f\} \tag{5}$$

where e_f is the event label of frame f (§3.2) and C_f is the set of memory tokens in the same event but from other frames. A weight $w_i \approx 0$ means token i has a near-duplicate elsewhere in the event, and $w_i \approx 1$ means it is unique. When $C_f = \varnothing$, we set $w_i = 1$. The kernel $\tilde{L}_f$ thus encodes two dimensions of marginal contribution: intra-frame complementarity through the off-diagonal Gram entries, and cross-frame redundancy through the diagonal scaling by w_i.

Because L is block-diagonal, scoring decomposes into independent per-frame subproblems. For frame f, we apply greedy ordering (Chen et al., 2018): starting from an empty selected set $C = \emptyset$, each step selects the token i that contributes the most new information given C and adds it to C. The amount of new information is measured by the conditional variance of token i given C, defined as the Schur complement diagonal of $\tilde{L}_f$:

$$\tilde{L}_{f,\, ii|C} \triangleq \tilde{L}_{f,\, ii} - \tilde{L}_{f,\, iC}\, \tilde{L}_{f,\, CC}^{-1}\, \tilde{L}_{f,\, Ci} \tag{6}$$

where $\tilde{L}_{f,\, ii}$ is the diagonal entry of token i, and $\tilde{L}_{f,\, iC}$, $\tilde{L}_{f,\, CC}$ are the cross and submatrix entries indexed by the selected set C. The conditional variance is large when token i provides information not covered by C, and small when C already covers it. We run the greedy process over all n tokens until every token is ordered and define the marginal contribution score of each token as its conditional variance at the time of selection:

$$s_i = \tilde{L}_{f,\, ii|C_i} \tag{7}$$

where C_i is the set of same-frame tokens selected before token i. Earlier-selected tokens have large s_i, and later-selected tokens, whose information is already covered, have small s_i. Because $\tilde{L}_f$ encodes both intra-frame similarity and cross-frame redundancy, s_i approximates marginal contribution along both dimensions in a single pass. This approximation replaces the conditional variance of token i given the entire memory with two terms that are cheaper to compute: the residual against its nearest match within the same event, and its conditional variance given the same-frame tokens selected before it.

Non-Frozen Rescoring. Marginal contribution is conditional on the current memory and changes as the memory evolves. Whenever a new frame arrives, the system recomputes w_i and s_i for all surviving tokens. As new frames join the same event, older tokens accumulate redundancy, and their w_i and s_i decrease accordingly, so the scores reflect the current memory state rather than a

snapshot from entry time. Rescoring also prevents early frames from keeping the inflated scores they receive from the smaller reference set available at entry time. Because w_i uses single-token maximum matching rather than full-set projection, the deduplication weight of each token depends only on its closest duplicate and does not saturate toward zero as the event grows.

3.4 Age Discount and Eviction

To guarantee real-time perception, we protect the most recent P frames from eviction, so their tokens remain in memory unconditionally. For the remaining tokens, the score s_i measures marginal contribution but ignores time: evicting by s_i alone would let stale but coincidentally unique tokens permanently occupy memory. Inspired by memory decay models in cognitive psychology (Atkinson & Shiffrin, 1968; Ebbinghaus, 1885), we apply an exponential age discount, defining the time-discounted score

$$\kappa_i = 2^{-h \cdot \delta_i} \cdot s_i \tag{8}$$

where $\delta_i = t - t_i$ is the age of token i in frames since its insertion, and $h > 0$ is the decay rate. When $|\mathcal{M}| > B$, unprotected tokens are evicted in ascending κ_i order until $|\mathcal{M}| = B$.

Proposition 1. *Embedding the age discount into the quality term of the global kernel $L = \bigoplus_f \tilde{L}_f$ in Eq. (3) from §3.3, we define the augmented kernel*

$$L^{(\kappa)} = \mathrm{diag}(q) \left(\bigoplus_f G_f G_f^\top \right) \mathrm{diag}(q), \qquad q_i^2 = 2^{-h \cdot \delta_i} \cdot w_i \tag{9}$$

Because tokens in the same frame share the age δ_f, the frame-f block of $L^{(\kappa)}$ is $L_f^{(\kappa)} = 2^{-h \cdot \delta_f} \cdot \tilde{L}_f$. Under greedy ordering on $L^{(\kappa)}$, the conditional variance of token i at the time of selection equals $\kappa_i = 2^{-h \cdot \delta_i} \cdot s_i$.

The proof is in Appendix A. Proposition 1 establishes that evicting by κ_i is equivalent to greedy ordering on $L^{(\kappa)}$, so eviction removes the tokens whose conditional variance is smallest.

3.5 Algorithm and Properties

Computational Cost. Algorithm 1 presents the complete procedure. The dominant per-step cost is Step 3, which computes w_i with one matrix multiplication and a row-wise maximum and runs intra-frame greedy ordering at $O(n^2 d)$ cost for each frame in memory. Let F be the number of frames retaining at least one token, and let N_e be the maximum token count in any single event. The per-step cost is $O(F \cdot n \cdot (N_e \cdot d + n \cdot d))$. Both F and N_e are bounded by the memory capacity, so this cost is independent of stream length.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluated on two streaming video question answering benchmarks. From OVO-Bench (Li et al., 2025b), we used the Real-Time Visual Perception (RT) and Backward Tracing (BT) categories, which ask about the scene at the query time and about evidence that appeared earlier in the stream. From StreamingBench (Lin et al., 2026), we used the Real-Time Visual Understanding subset, which contains multiple-choice questions on real streaming videos.

Baselines. We compared against nine published streaming methods: FluxMem (Xie et al., 2026), WeaveTime (Zhang et al., 2026b), HERMES (Zhang et al., 2026a), StreamForest (Zeng et al., 2025), Streamo (Xia et al., 2026), TimeChat-Online (Yao et al., 2025), Dispider (Qian et al., 2025), VideoLLM-online (Chen et al., 2024a), and Flash-VStream (Zhang et al., 2025a). Representative offline Video LLMs are listed as reference points. Unless otherwise noted, all baselines were reproduced with the configurations from the original papers or official code.

Implementation Details. We tested SMART on two backbones: Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and Qwen3-VL-8B-Instruct (Qwen Team, 2025). Frames were sampled at 1 fps. Unless otherwise noted, we set the memory capacity to $B = 4n$, which equals the token count of 4 frames. Hyperparameter details are in Appendix C.

Algorithm 1 SMART Streaming DPP Memory Management

Require: Frame feature sequence $(G_1, G_2, \ldots)$, n unit-norm tokens per frame; memory capacity B; decay rate h; similarity threshold θ; minimum event length $L_{\min}$; protected frames P

$\mathcal{M} \leftarrow \emptyset, \quad e \leftarrow 0, \quad b \leftarrow 0$
for each arriving frame t **do**
  // Step 1: Event segmentation (§3.2)
  if $t > 0$ **then**
    $C_e \leftarrow$ all tokens of event e in memory
    $\phi_t \leftarrow \mathrm{median}_i \max_{j \in C_e} G_t[i]^\top f_j$
    if $\phi_t < \theta$ **and** $t - b \geq L_{\min}$ **then**
      $e \leftarrow e + 1, \quad b \leftarrow t$
    end if
  end if
  // Step 2: Insertion
  Insert n tokens of frame t into $\mathcal{M}$, set $e_t \leftarrow e$, mark as protected
  // Step 3: Full-memory rescoring (§3.3)
  for each frame f in $\mathcal{M}$ **do**
    $C_f \leftarrow \{j \in \mathcal{M} : e_j = e_f,\ j \notin f\}$
    $w_i \leftarrow 1 - (\max_{j \in C_f} f_i^\top f_j)^2$ for each token i in frame f, or $w_i \leftarrow 1$ if $C_f = \varnothing$
    $\tilde{L}_f \leftarrow \mathrm{diag}(\sqrt{w})\, G_f G_f^\top \mathrm{diag}(\sqrt{w})$
    Run greedy ordering on $\tilde{L}_f$ to obtain score s_i for each token
  end for
  // Step 4: Eviction (§3.4)
  while $|\mathcal{M}| > B$ **do**
    Compute $\kappa_i \leftarrow 2^{-h \cdot \delta_i} \cdot s_i$ for all unprotected tokens
    Evict the token with smallest κ_i
  end while
  Unmark frames no longer among the newest P frames
end for

Table 1: Main results on OVO-Bench and StreamingBench. OVO-RT and OVO-BT are Real-Time Visual Perception and Backward Tracing accuracy, and OVO-Avg is the arithmetic mean of OVO-RT and OVO-BT. SBench is accuracy on the Real-Time Visual Understanding subset of StreamingBench. All metrics are accuracy in %, and higher is better. Best in **bold**, second best underlined. Per-subtask breakdown is in Appendix F.

				OVO-Bench		
Model	**Backbone**	**#Fr**	**SBench**	**OVO-RT**	**OVO-BT**	**Avg**
Offline Video LLMs						
Qwen2.5-VL	7B	1fps	73.3	59.9	44.7	52.3
Qwen3-VL	8B	1fps	78.9	69.4	47.0	58.2
LLaVA-OV	7B	32	71.1	64.0	43.7	53.9
InternVL2	8B	16	63.7	60.4	43.4	51.9
LLaVA-Video	7B	64	–	63.5	40.4	52.0
Qwen2-VL	7B	64	69.0	60.7	48.6	54.7
LongVU	7B	1fps	–	57.4	39.5	48.5
Online / Streaming Video LLMs						
VideoLLM-online	8B	2fps	35.9	20.8	17.7	19.3
Flash-VStream	7B	1fps	23.3	28.4	27.4	27.9
Dispider	7B	1fps	67.6	54.6	36.1	45.4
TimeChat-Online	7B	1fps	75.3	61.9	41.7	51.8
StreamForest	7B	1fps	77.2	61.2	52.0	56.6
Streamo	7B	1fps	–	66.0	46.1	56.1
FluxMem	7B	1fps	78.1	63.6	48.2	55.9
WeaveTime	7B	1fps	79.1	67.0	50.5	58.8
HERMES	7B	1fps	79.4	69.0	49.4	59.2
SMART (ours)						
Qwen2.5-VL	7B	1fps	81.1	78.2	50.6	64.4
Qwen3-VL	8B	1fps	**82.4**	**78.7**	**58.6**	**68.6**

4.2 Main Results

SMART achieved the highest OVO-Avg and StreamingBench accuracy among streaming methods with both backbones, reaching 68.6% and 82.4% with Qwen3-VL-8B and 64.4% and 81.1% with Qwen2.5-VL-7B. The largest gain was on OVO-RT, where SMART exceeded the strongest stream-

Table 2: Ablation on OVO-Bench and StreamingBench. Each variant changes one component.

	Qwen3-VL-8B				Qwen2.5-VL-7B			
Variant	OVO-Avg	Δ	SBench	Δ	OVO-Avg	Δ	SBench	Δ
SMART (full method)	68.6		82.4		64.4		81.1	
No cross-frame deduplication ($w_i \equiv 1$)	67.3	−1.3	81.3	−1.1	63.1	−1.3	80.0	−1.1
Global deduplication ($C_f = \mathcal{M}$)	67.9	−0.7	81.1	−1.3	63.9	−0.5	80.5	−0.6
Randomized s_i	66.1	−2.5	80.2	−2.2	61.7	−2.7	79.1	−2.0
No age discount ($\kappa_i = s_i$)	67.6	−1.0	81.0	−1.4	63.8	−0.6	79.9	−1.2
No protected frames ($P = 0$)	68.1	−0.5	82.0	−0.4	63.5	−0.9	80.9	−0.2

ing baseline HERMES by more than 9 points with both backbones. On OVO-BT, the Qwen3-VL-8B variant reached 58.6%, 6.6 points above StreamForest, whereas the Qwen2.5-VL-7B variant reached 50.6% and trailed StreamForest by 1.4 points. Because the methods in Table 1 differ in backbone architecture and training data, these absolute gaps reflect both the backbone and the memory management algorithm. The ablation in Table 2 isolates each component on each fixed backbone.

4.3 Ablation Study

Removing any single component lowered both OVO-Avg and StreamingBench (SBench) accuracy on both backbones, and marginal contribution scoring contributed the most. All variants in Table 2 used the memory capacity $B/n = 4$, and within each backbone they differed only in memory management. Pairs of numbers below refer to Qwen3-VL-8B and Qwen2.5-VL-7B in this order.

The **no cross-frame deduplication** variant ($w_i \equiv 1$) set all deduplication weights to one, so the kernel reduced to the raw Gram matrix $\tilde{L}_f = G_f G_f^\top$. Greedy ordering still captured intra-frame complementarity, but event structure no longer entered the score, because it enters only through w_i. OVO-Avg dropped by 1.3 points and SBench by 1.1 points on both backbones. The **global deduplication** variant ($C_f = \mathcal{M}$) kept the deduplication weight but compared each token against the entire memory rather than the current event. OVO-Avg dropped by 0.7 and 0.5 points and SBench by 1.3 and 0.6 points, so restricting the comparison to the current event contributed beyond deduplication against the entire memory on both backbones. On Qwen3-VL-8B, global deduplication scored 0.2 points below the variant without cross-frame deduplication on SBench, so the entire deduplication gain on this benchmark came from the event-aware scope.

The **randomized** s_i variant replaced s_i with uniform random values and kept age discounting and protected frames. OVO-Avg dropped by 2.5 and 2.7 points and SBench by 2.2 and 2.0 points, the largest drops among all variants, so marginal contribution scoring is the largest single source of the gain over age-discounted eviction. The **no age discount** variant ($\kappa_i = s_i$) removed the exponential decay, so old tokens with high s_i stayed in memory regardless of age. On Qwen3-VL-8B, OVO-BT rose from 58.6% to 60.3% because more old content survived, whereas OVO-RT fell from 78.7% to 74.9% because recent perceptual information was displaced. OVO-Avg dropped by 1.0 and 0.6 points and SBench by 1.4 and 1.2 points, the second largest SBench drops for both backbones.

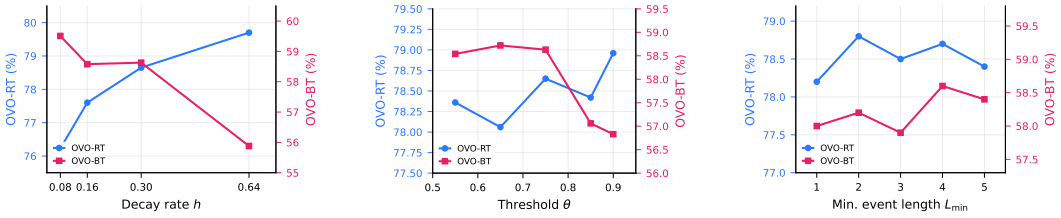
The **no protected frames** variant ($P = 0$) removed the guarantee that the most recent frame stays in memory in full. OVO-Avg dropped by 0.5 and 0.9 points and SBench by 0.4 and 0.2 points. The protected frame secures real-time perception: questions about the present scene need the complete current frame, which marginal contribution scoring alone does not guarantee.

4.4 Sensitivity Analysis

Memory Capacity. The memory capacity B controls how much past visual content the memory can retain. A larger memory capacity raised StreamingBench accuracy and left OVO-Bench nearly unchanged. As B/n grew from 4 to 8 in Table 3, StreamingBench accuracy increased from 82.4 to 83.4. OVO-Avg declined from 68.6 to 68.3 over the same range because OVO-RT fell from 78.7 to 78.0, while OVO-BT stayed at 58.6.

Table 3: Performance under memory capacity B/n, in units of frames.

B/n	SBench	OVO-RT	OVO-BT	OVO-Avg
4	82.4	78.7	58.6	68.6
6	82.5	78.4	58.3	68.4
8	83.4	78.0	58.6	68.3



(a) Decay rate h (b) Threshold θ (c) Min. event length $L_{\min}$

Figure 3: Sensitivity analysis on OVO-Bench.

Decay Rate h. The decay rate h controls the trade-off between recency and content value in eviction. OVO-Avg peaked near the default $h = 0.3$ in Figure 3a. As h increased, old tokens decayed faster, so OVO-RT improved monotonically while OVO-BT declined, most steeply beyond $h = 0.3$.

Event Segmentation Threshold θ. The threshold θ controls the sensitivity of event boundary detection. OVO-BT declined by 1.9 points for $\theta > 0.75$ in Figure 3b, while OVO-RT stayed between 78.1 and 79.0 across $[0.5, 0.9]$. As θ increased, the mean event count per window of at most 64 frames grew from 1.1 to 13.3, and event lengths shrank as Figure 4 shows. Event length was capped at 64 frames, the window length, so an event of 64 frames means that no boundary was placed in the window. At $\theta = 0.55$, 90% of the windows formed a single event, so deduplication compared each token against the entire memory. At $\theta = 0.90$, frequent boundaries created many short events with small comparison sets. The OVO-BT decline coincided with this fragmentation, which suggests that small comparison sets weaken deduplication for backward tracing.

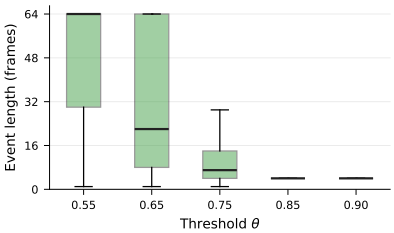


Figure 4: Event length distribution.

Minimum Event Length $L_{\min}$. The minimum event length $L_{\min}$ controls the robustness of event segmentation to inter-frame noise. OVO-RT and OVO-BT both peaked at $L_{\min} = 4$ in Figure 3c and varied by less than 1 point across the tested range, with θ fixed at 0.75. A small $L_{\min}$ admits spurious boundaries from transient similarity drops, and a large one delays real boundaries.

Computational Overhead. Memory management dominated the per-frame latency on a single RTX 4090, taking 739 ms against 58 ms for visual encoding in Table 4. Greedy ordering accounted for 728 ms because it rescores every frame in memory at each arrival. In exchange, SMART compresses all seen frames to at most B tokens, which bounds the LLM forward cost at query time. Batching greedy ordering into one GPU kernel or restricting rescoring to the current event may reduce this cost in future work.

Table 4: Per-frame latency in milliseconds.

Stage	ms
Visual encoding	58
Memory management	739
Event segmentation (Step 1)	<1
Deduplication weight w_i (Step 3)	3
Greedy ordering (Step 3)	728
Eviction (Step 4)	7

5 Conclusion

We present SMART, a training-free streaming visual token memory that retains tokens by their marginal contribution to the current memory and weighs this contribution jointly with recency within a fixed memory capacity. Adaptive event segmentation restricts redundancy to similarity within an event, cross-frame deduplication weights and intra-frame DPP greedy ordering jointly score marginal contribution with rescoring at every frame arrival, and an age-augmented DPP kernel lets all tokens outside the protected recent frame compete in a single eviction pool. SMART achieves the highest OVO-Avg and StreamingBench accuracy among the evaluated streaming methods with both backbones, and removing any component lowers both metrics on both backbones, with marginal contribution scoring contributing the most. Future work may pair this query-independent memory with query-aware retrieval at answer time, so that the tokens passed to the language model also reflect the question.

AI USE STATEMENT

Generative AI tools were used during the preparation of this manuscript for three purposes: (1) aiding and polishing writing, (2) retrieval and discovery of related work, and (3) research ideation or execution. All AI-generated or AI-assisted content has been reviewed by the authors. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. Our method is a training-free inference-time token memory mechanism evaluated on publicly available video question answering benchmarks (OVO-Bench, StreamingBench). No human subjects were involved, and no new data were collected. The benchmark videos may contain people, and we used them only as released by the benchmarks for evaluation, without identifying or analyzing any individual. The benchmarks used are released under their respective licenses for research purposes. We do not foresee direct negative societal impacts from the proposed memory management strategy.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of all hyperparameters and the default configuration in Section 4.1 and Appendix C. The algorithmic procedure is fully specified in Algorithm 1. All experiments use publicly available models (Qwen2.5-VL-7B-Instruct, Qwen3-VL-8B-Instruct) and public benchmarks (OVO-Bench, StreamingBench) with their standard evaluation protocols. The method is training-free and deterministic given a fixed decoding configuration. Source code will be released upon publication.

REFERENCES

Richard C. Atkinson and Richard M. Shiffrin. Human memory: A proposed system and its control processes. In *Psychology of Learning and Motivation*, volume 2, pp. 89–195. Academic Press, 1968.

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*, 2025.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations (ICLR)*, 2023.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

Laming Chen, Guoxin Zhang, and Eric Zhou. Fast greedy MAP inference for determinantal point process to improve recommendation diversity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 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 (CVPR)*, 2026.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67 (12):220101, 2024c.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Vaggelis Dorovatas, Soroush Seifi, Gunshi Gupta, and Rahaf Aljundi. Recurrent attention-based token selection for efficient streaming Video-LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Hermann Ebbinghaus. *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot, Leipzig, 1885.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Zifan Han, Hongbo Sun, Jinglin Xu, Canhui Tang, Yulong Lei, Xuchong Zhang, Hongbin Sun, Zhongjiang He, and Hao Sun. WAT: Online video understanding needs watching before thinking. *arXiv preprint arXiv:2603.13412*, 2026.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. InfiniPot-V: Memory-constrained KV cache compression for streaming video understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012.

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

Yifei Li, Junbo Niu, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. OVO-Bench: How far is your Video-LLMs from real-world online video understanding? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jinming Liu, Jianguo Huang, Zhaoyang Jia, Jiahao Li, Xiaoyi Zhang, Zongyu Guo, Bin Li, Wenjun Zeng, Yan Lu, and Xin Jin. An efficient streaming video understanding framework with agentic control. *arXiv preprint arXiv:2605.17921*, 2026.

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

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Xiao Wang, Qingyi Si, Shiyu Zhu, Jianlong Wu, Li Cao, and Liqiang Nie. AdaReTaKe: Adaptive redundancy reduction to perceive longer for video-language understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025b.

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 (CVPR)*, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

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 (CVPR)*, 2026.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. StreamingVLM: Real-time understanding for infinite video streams. In *International Conference on Learning Representations (ICLR)*, 2026.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, Lingpeng Kong, Qi Liu, Yuanxing Zhang, and Xu Sun. TimeChat-Online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2025.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, Yi Wang, and Limin Wang. StreamForest: Efficient online video understanding with persistent event memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, 2023a.

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

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 Annual Meeting of the Association for Computational Linguistics (ACL)*, 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025c.

Yulin Zhang, Cheng Shi, and Sibei Yang. WeaveTime: Streaming from earlier frames into emergent memory in VideoLLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Yunzhu Zhang, Yu Lu, Tianyi Wang, Fengyun Rao, Yi Yang, and Linchao Zhu. FlexSelect: Flexible token selection for efficient long video understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025d.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

A Proof of Proposition 1

The augmented kernel $L^{(\kappa)}$ (Eq. (9)) is

$$L^{(\kappa)} = \operatorname{diag}(q)\left(\bigoplus_f G_f G_f^\top\right)\operatorname{diag}(q), \qquad q_i^2 = 2^{-h\cdot\delta_i}\cdot w_i \tag{10}$$

where w_i is the deduplication weight (Eq. (5)) and $\bigoplus_f$ denotes the direct sum over frames. The block of $L^{(\kappa)}$ for frame f is

$$L_f^{(\kappa)} = \operatorname{diag}(q_f)\, G_f G_f^\top \operatorname{diag}(q_f) \tag{11}$$

Tokens in the same frame share age $\delta_f = t - t_f$, so $2^{-h\cdot\delta_f/2}$ is constant within the frame. Substituting $q_i = 2^{-h\cdot\delta_f/2}\sqrt{w_i}$:

$$L_f^{(\kappa)} = \operatorname{diag}\big(2^{-h\cdot\delta_f/2}\sqrt{w}\big)\, G_f G_f^\top \operatorname{diag}\big(2^{-h\cdot\delta_f/2}\sqrt{w}\big) \tag{12}$$

$$= 2^{-h\cdot\delta_f}\cdot\operatorname{diag}(\sqrt{w})\, G_f G_f^\top \operatorname{diag}(\sqrt{w}) \tag{13}$$

$$= 2^{-h\cdot\delta_f}\cdot\tilde{L}_f \tag{14}$$

where $\tilde{L}_f = \operatorname{diag}(\sqrt{w})\, G_f G_f^\top \operatorname{diag}(\sqrt{w})$ is the intra-frame kernel (Eq. (4)).

Because $L_f^{(\kappa)} = 2^{-h\cdot\delta_f}\cdot\tilde{L}_f$, every conditional Schur complement diagonal entry is scaled by $2^{-h\cdot\delta_f}$ relative to $\tilde{L}_f$. The greedy conditional variance of token i on $L_f^{(\kappa)}$ is therefore

$$2^{-h\cdot\delta_f}\cdot s_i = 2^{-h\cdot\delta_i}\cdot s_i = \kappa_i \tag{15}$$

where s_i is the greedy conditional variance of token i on $\tilde{L}_f$ (Eq. (7)).

Because $L^{(\kappa)}$ is block-diagonal, the scalar factor $2^{-h\cdot\delta_f}$ scales all conditional variances within frame f uniformly, leaving the intra-frame selection order unchanged. The resulting κ_i values can therefore be compared across frames in a single eviction pool, while intra-frame ranking depends only on $\tilde{L}_f$.

B Interpretation of Intra-Event Deduplication Weights

The deduplication weight $w_i = 1 - (\max_{j\in C_f} f_i^\top f_j)^2$ equals the Schur complement residual conditioned on the single best-matching token. Let $j^* = \arg\max_{j\in C_f} f_i^\top f_j$. The conditioned diagonal entry is

$$S_{ii} - S_{ij^*} S_{j^*j^*}^{-1} S_{j^*i} = \|f_i\|^2 - (f_i^\top f_{j^*})^2/\|f_{j^*}\|^2 = 1 - (f_i^\top f_{j^*})^2 = w_i \tag{16}$$

where the last step uses the unit norms $\|f_i\| = \|f_{j^*}\| = 1$.

We prefer single-token maximum matching over the full-set projection residual $\|f_i - \mathrm{proj}_{C_f} f_i\|^2$ for two reasons. (1) The single-token residual depends only on the closest match, so it stays large for a token without a near-duplicate however many frames the event contains. The full projection residual tends toward zero as same-event frames accumulate and suppresses the scores of all old tokens. (2) Maximum matching costs $O(|C_f| \cdot d)$, whereas the full-set Schur complement requires $O(|C_f|^2 d + |C_f|^3)$.

C Implementation Details

Model and Inference Configuration. We used the official checkpoints of Qwen2.5-VL-7B-Instruct and Qwen3-VL-8B-Instruct with the default dynamic resolution, which produces n visual tokens per frame, typically between 256 and 1024 depending on resolution. Inference used bfloat16 precision, greedy decoding with temperature 0, and no system prompt. We adopted the standard prompt templates and answer-parsing rules provided by each benchmark.

Hyperparameters. The default configuration used memory capacity $B = 4n$, decay rate $h = 0.3$, similarity threshold $\theta = 0.75$, minimum event length $L_{\min} = 4$, and $P = 1$ protected frame. The same hyperparameters were used for both backbones and on both benchmarks. SMART is training-free and has no learned parameters. Section 4.4 reports the sensitivity to each hyperparameter: accuracy varied by less than 1 point across the tested range of $L_{\min}$, whereas OVO-BT declined by 1.9 points for $\theta > 0.75$.

Reference Initialization of h. We derive a reference starting point for h. Under the discount $2^{-h\cdot\delta}$, requiring the geometric series of steady-state retention to sum to B/n gives $\sum_{\delta=0}^{\infty} 2^{-h\cdot\delta} = 1/(1-2^{-h}) = B/n$, which yields $h_0 = \log_2(B/n/(B/n-1))$. At this value, the expected retention at each age decays as $2^{-h_0\cdot\delta}$ and sums to the capacity.

Computing Environment. All experiments ran on a single NVIDIA GeForce RTX 4090 24GB GPU, with PyTorch 2.3 and CUDA 12.1.

D Memory Age Distribution

Figure 5 shows the age distribution of tokens in steady-state memory, computed over all RT+BT queries of OVO-Bench. Statistics excluded the protected frame, which retains all of its tokens unconditionally, and were renormalized.

Figure 5(a) varies memory capacity with h fixed at 0.3. Smaller capacity produced steeper decay: at $B/n = 2$, frames with age ≥ 6 had nearly zero retention, while at $B/n = 8$, age 15 still held approximately 3%. Figure 5(b) varies the decay rate with B/n fixed at 5. At $h = 0.64$, retention concentrated on recent frames and approached a sliding window, and at $h = 0.08$, the distribution flattened and distant frames received comparable shares.

The distribution reflects the interplay between the age discount $2^{-h\cdot\delta}$ and the capacity constraint. Larger h concentrates retention on recent frames, and smaller h increases the competitiveness of

distant frames. Even at $h = 0.08$, age-15 frames held approximately 4% of competing capacity, so distant content is not excluded entirely.

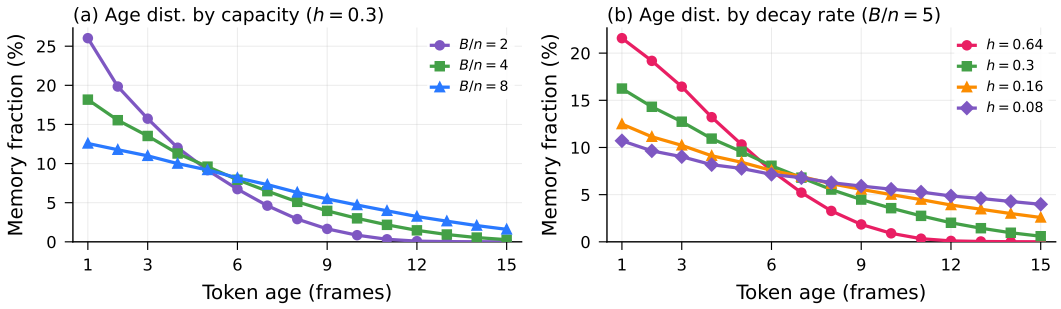


Figure 5: Age distribution of tokens in steady-state memory, excluding the protected frame. (a) Different memory capacities with $h = 0.3$. (b) Different decay rates with $B/n = 5$. Statistics are computed over all RT+BT queries of OVO-Bench.

E Event Segmentation Visualization

Figures 6–8 visualize adaptive event segmentation on three OVO-Bench videos. Each row is one event, border color indicates event membership, and frame numbers are annotated above each frame. OVO #1393 is a badminton match whose 4 events follow the transitions between match footage, close-ups, replay analysis, and the scoreboard. OVO #1264 is a football match whose 7 events follow multiple camera switches. OVO #483 is a tire repair video whose 3 events follow the action transitions from preparation to repair and inflation.

F Additional Experimental Results

Table 5: Per-subtask accuracy on OVO-Bench in %, where higher is better. RT tasks include OCR, ACR, ATR, STU, FPD, and OJR; BT tasks include EPM, ASI, and HLD. Best results in **bold**.

			Real-Time Visual Perception							Backward Tracing				
Model	**Backbone**	**#Fr**	**OCR**	**ACR**	**ATR**	**STU**	**FPD**	**OJR**	**Avg**	**EPM**	**ASI**	**HLD**	**Avg**	**OVO-Avg**
Offline Video LLMs														
Qwen2.5-VL	7B	1fps	67.8	55.1	67.2	42.1	66.3	60.9	59.9	51.5	58.8	23.7	44.7	52.28
LLaVA-OV	7B	32	66.4	57.8	73.3	53.4	71.3	62.0	64.0	54.2	55.4	21.5	43.7	53.85
InternVL2	8B	16	67.1	60.6	63.8	46.1	68.3	56.5	60.4	48.2	57.4	24.7	43.4	51.90
LLaVA-Video	7B	64	69.1	58.7	68.8	49.4	74.3	59.8	63.5	56.2	57.4	7.5	40.4	51.95
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	54.62
LongVU	7B	1fps	55.7	49.5	59.5	48.3	68.3	63.0	57.4	43.1	66.2	9.1	39.5	48.45
Online / Streaming Video LLMs														
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	19.26
Flash-VStream	7B	1fps	24.2	29.4	28.5	33.7	25.7	28.8	28.4	39.1	37.2	5.9	27.4	27.90
Dispider	7B	1fps	57.7	49.5	62.1	44.9	61.4	51.6	54.6	48.5	55.4	4.3	36.1	45.35
TimeChat-Online	7B	1fps	75.2	46.8	70.7	47.8	69.3	61.4	61.9	55.9	59.5	9.7	41.7	51.80
StreamForest	7B	1fps	68.5	53.2	71.6	47.8	65.4	60.9	61.2	58.9	64.9	32.3	52.0	56.60
Streamo	7B	1fps	79.2	57.8	75.0	49.4	64.4	70.1	66.0	54.6	52.0	31.7	46.1	56.05
FluxMem	7B	1fps	75.5	55.5	71.0	49.5	67.5	62.5	63.6	53.5	61.5	29.5	48.2	55.90
WeaveTime	7B	1fps	81.5	60.0	73.0	51.5	71.5	64.5	67.0	54.5	63.0	34.0	50.5	58.75
HERMES	7B	1fps	85.2	64.2	71.6	53.4	74.3	65.2	69.0	48.5	62.2	37.6	49.4	59.20
SMART (ours)														
SMART	Qwen2.5-VL-7B	1fps	**94.0**	69.7	**81.9**	**65.2**	75.3	**83.2**	78.2	53.2	64.2	34.4	50.6	64.40
SMART	Qwen3-VL-8B	1fps	91.3	**77.1**	81.0	62.4	**79.2**	81.0	**78.7**	**56.6**	**68.2**	**51.1**	**58.6**	**68.64**

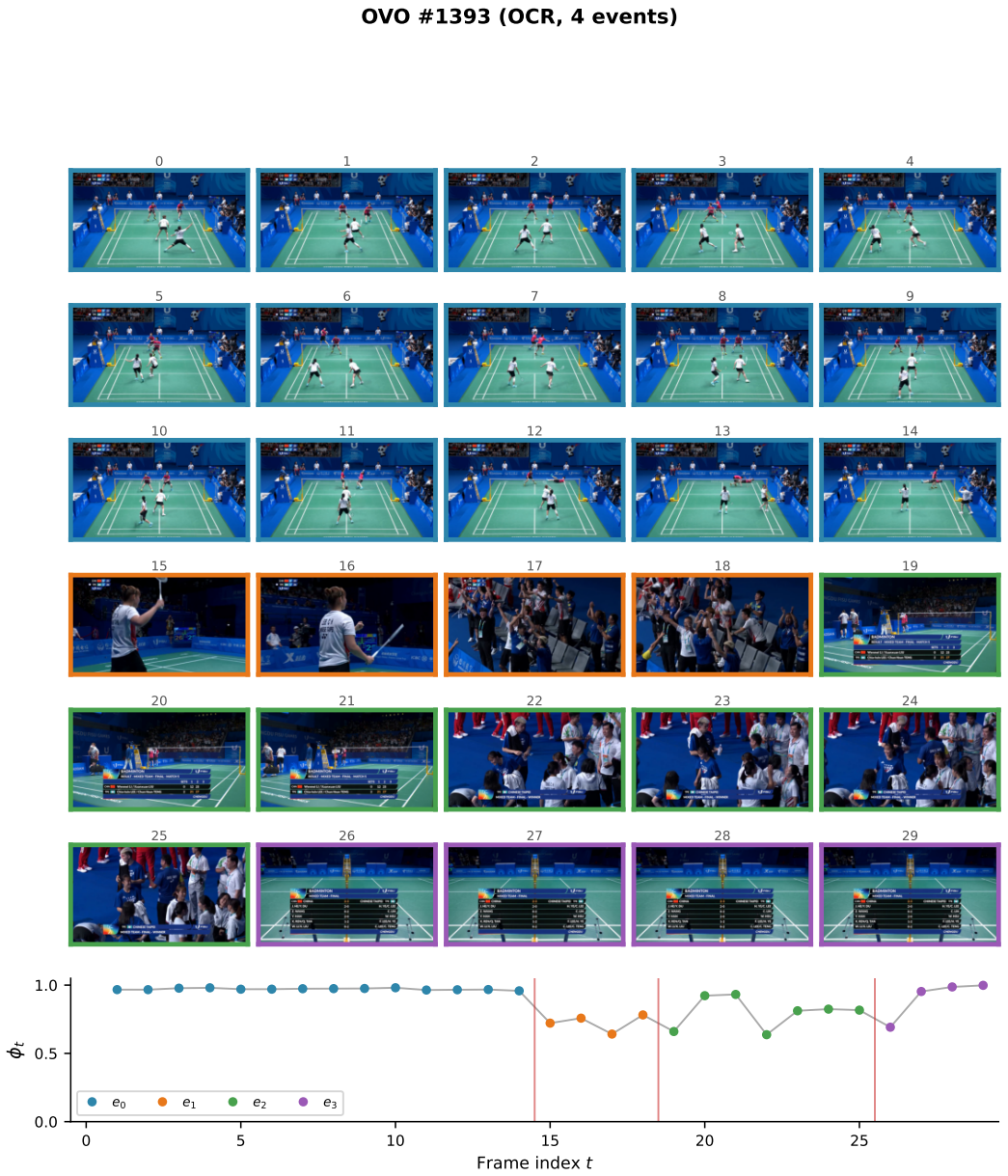


Figure 6: Event segmentation results for OVO #1393, an OCR sample with 4 events. e_0: match footage, 15 frames; e_1: close-up transition, 4 frames; e_2: replay and analysis, 7 frames; e_3: scoreboard, 4 frames.



Figure 7: Event segmentation results for OVO #1264, an ACR sample with 7 events. The event boundaries follow the camera switches in the video.

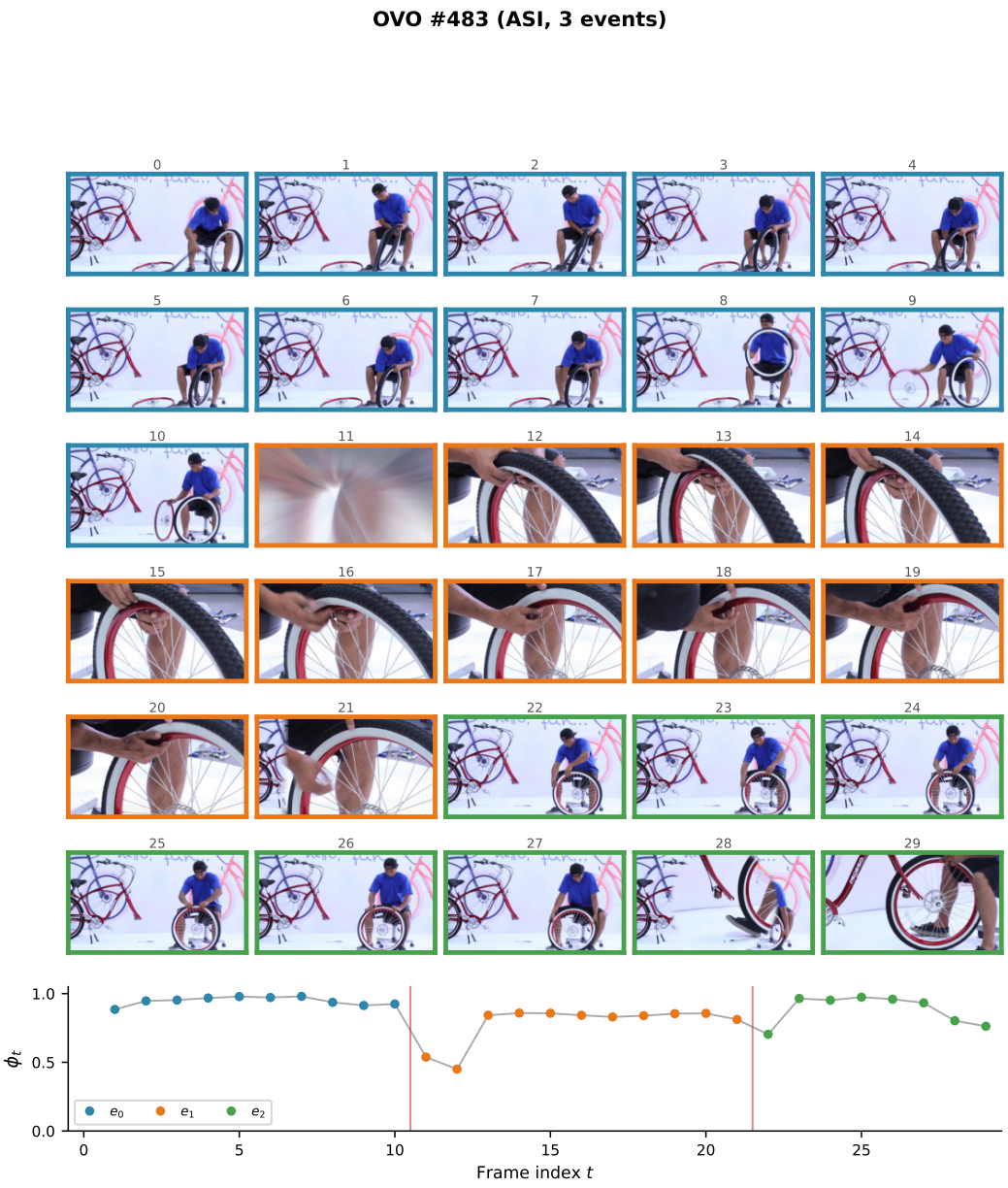


Figure 8: Event segmentation results for OVO #483, an ASI sample with 3 events. e_0: preparation, e_1: tire repair, e_2: inflation.