# QUERY-ADAPTIVE VISUAL-SEMANTIC EVIDENCE ALLOCATION FOR STREAMING VIDEO UNDERSTANDING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Streaming video understanding requires answering questions over an ever-growing visual history under limited memory and inference budgets. Existing methods typically manage this history through fixed-capacity memory compression or similarity-based frame retrieval, but largely rely on fixed compression or retrieval criteria without explicitly modeling the specific historical evidence requirements of each query. This overlooks a key property of streaming questions: their historical evidence requirements are inherently heterogeneous. To address this issue, we propose a training-free framework for *query-adaptive visual-semantic evidence allocation* that jointly determines whether to access history, which temporal forms of visual evidence to prioritize, and which long-range semantic evidence to retrieve. Given a query and recent observations, the framework jointly predicts whether historical access is needed and the query's preferences over static states, dynamic context, and recency-sensitive evidence. Questions supported by recent observations are answered directly through Fast–Slow routing. For queries requiring history, the same evidence preferences guide both fine-grained visual selection and event-level semantic ranking. The visual branch combines query-specific preferences for static, dynamic, and recency-sensitive evidence to select a compact and diverse set of historical frames. The semantic branch propagates the preference signals to events in a hierarchical caption memory, fuses the event scores with the same query preferences, and complements preference-guided ranking with VLM-based caption retrieval. The resulting visual and semantic evidence is integrated with recent observations for final answering. Our framework achieves 67.79% accuracy on OVO-Bench and 77.01% across the 14 visual-only tasks of Streaming-Bench, outperforming the compared methods on both benchmarks. These results demonstrate the value of coordinating history access and visual-semantic evidence selection according to query-specific requirements.

## 1 INTRODUCTION

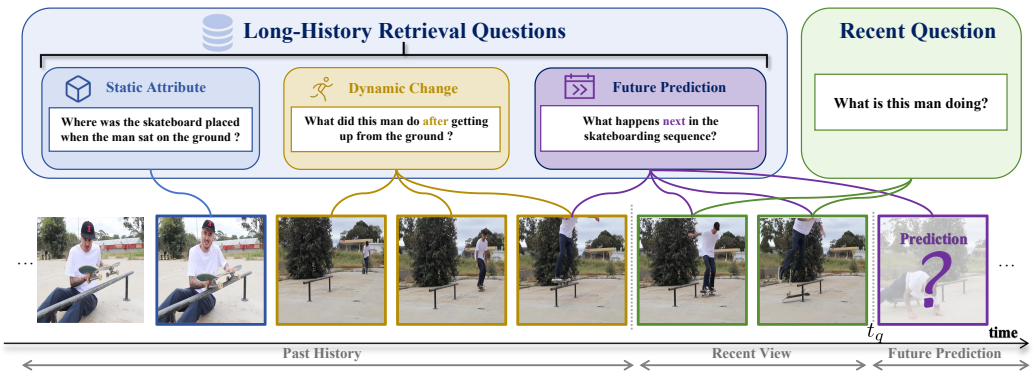


Figure 1: Illustration of query-dependent evidence requirements in streaming video understanding. Different questions may rely on recent observations or historical evidence with distinct temporal structures including static states, dynamic context and recency-sensitive observations.



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.

Streaming video understanding extends multimodal large language models from offline video analysis to continuously evolving visual environments, supporting applications in wearable assistance, live broadcast analysis and continuous monitoring. With an expanding visual history and queries unknown in advance, these applications require preserving useful context under limited budgets and organizing visual and semantic evidence to meet each query's needs.

Existing approaches address history preservation and access mainly through two strategies, as illustrated in Figure 2. The first maintains a bounded representation of the stream through visual-token pruning, memory compression or compact visual and KV states (Yao et al., 2025; Jin et al., 2025; Di et al., 2025; Yang et al., 2025b; Kim et al., 2025; Zhang et al., 2026b). Such methods control memory growth, but repeated compression or updates can remove fine-grained observations that later queries may require. Some approaches additionally depend on learned memory operations or task-specific training, which can limit their transferability across frozen vision–language backbones.

The second strategy retrieves historical information conditioned on the current question (Ge et al., 2026; Won et al., 2026; Ao et al., 2026). These approaches make historical access selective and query dependent, but relevance-based selection alone does not explicitly allocate evidence across the temporal forms required by a query. For example, questions involving relations such as "before" or "after" often require locating a relevant moment and retrieving its temporal neighborhood, which cannot be adequately captured by selecting individually relevant frames alone.

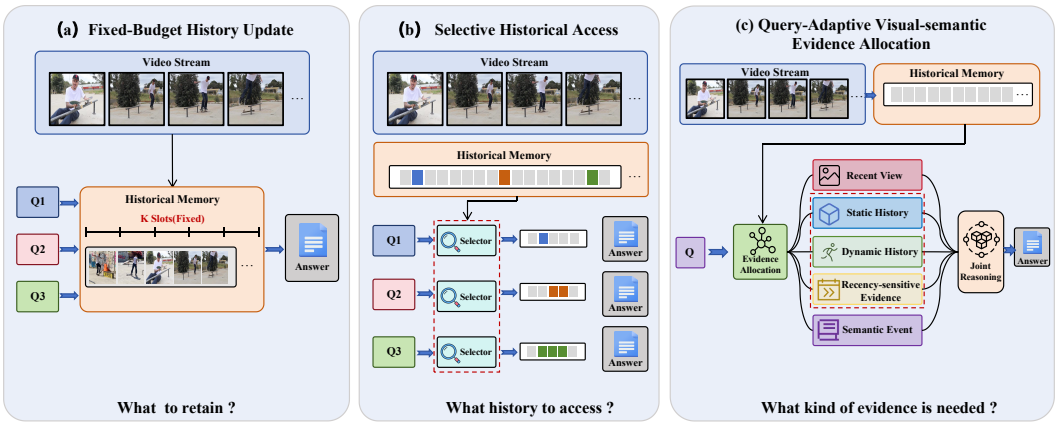


Figure 2: Comparison of different paradigms for streaming history utilization. (a) Fixed-budget history update maintains a bounded visual memory. (b) Selective historical access retrieves query-relevant historical observations. (c) Our framework allocates multimodal evidence according to the requirements of each query.

We argue that historical evidence demand in streaming video is heterogeneous along two complementary dimensions. First, queries differ in *whether* historical context is needed at all: many real-time questions can be answered reliably from recent observations, and unnecessary historical context can increase computation or introduce distracting information. Second, when history is needed, queries differ in *what temporal structure* of evidence is useful. Static states, local temporal dynamics, and recency-sensitive observations provide different forms of support. At longer time scales, completed events and relationships among events can be represented more compactly through semantic summaries. Effective streaming reasoning should therefore determine not only whether to access history, but also which visual evidence structure and which long-range semantic context should support the current query.

To this end, we propose a training-free framework for *query-adaptive visual-semantic evidence allocation* that jointly determines whether to access history, which temporal forms of visual evidence to prioritize, and which long-range semantic evidence to retrieve. Fast–Slow routing directly answers questions supported by recent observations and activates historical retrieval for the remaining queries. The underlying VLM remains frozen throughout.

The visual branch combines query-specific preferences for static states, dynamic context and recency-sensitive evidence to select compact and diverse historical frames. To align semantic evidence

with the same temporal requirements, the semantic branch separately averages static, dynamic and recency-sensitive frame relevance within each Event, producing three Event scores. It then fuses these scores using the query's predicted preferences to guide Event Caption selection from hierarchical memory. VLM caption retrieval adds complementary semantic relevance, while independently retrieved Episode summaries provide broader context. Selected Event and Episode Captions form a Timeline, which is integrated with recent observations, historical frames and the latest Working Caption for final answering.

Experiments yield accuracies of 67.79% on OVO-Bench (Niu et al., 2025) and 77.01% across the 14 visual-only tasks of StreamingBench (Lin et al., 2024), outperforming the compared offline and streaming methods on both benchmarks. The framework thus improves streaming reasoning by aligning historical access and complementary visual-semantic evidence with query requirements, without task-specific training.

Our contributions are summarized as follows:

- **Query-adaptive visual-semantic evidence allocation.** We formulate streaming understanding around heterogeneous query requirements, jointly determining historical access, visual evidence structure and semantic context.
- **Preference-guided visual evidence selection.** We combine query-specific preferences for static states, dynamic context and recency-sensitive evidence with semantic–temporal diversification to select compact historical visual evidence.
- **Preference-guided semantic evidence selection.** We compute separate Event scores by averaging static, dynamic and recency-sensitive frame relevance within each Event, then fuse them using shared query preferences. Caption retrieval and Episode context complement this preference-guided selection.

## 2 Related Work

**Long-Form Video Understanding.** Long-form video understanding has largely been studied as a problem of reducing the cost of accessing large visual histories. Memory-based methods compress dense video content into compact representations, including sparse long-term memory, compact visual states, and dynamically pruned tokens (Song et al., 2024; Shu et al., 2025; Shen et al., 2025; Tao et al., 2025). Query-aware methods instead allocate a limited visual budget to relevant evidence through frame selection, temporal search, or adaptive inspection strategies (Zhang et al., 2025c; Ye et al., 2025; Li et al., 2026; Wang et al., 2024b; Lin et al., 2026). These approaches demonstrate the value of selective evidence access, but they generally assume that the complete video is available when the question is known. Streaming video instead requires memory to be constructed before future queries are known and accessed only from the causally observed history.

**Streaming Video Understanding.** Streaming video understanding requires answering queries over evolving video streams using only observations available at query time. Benchmarks such as OVO-Bench (Niu et al., 2025) and StreamingBench (Lin et al., 2024) evaluate this causal setting. Existing systems address online generation and response behavior (Chen et al., 2024; Qian et al., 2025; Wang et al., 2025; Liu et al., 2026), efficient memory maintenance (Zhang et al., 2025a; Di et al., 2025; Zhang et al., 2026b; Zeng et al., 2025; Xie et al., 2026b), and selective historical-context utilization (Shen et al., 2026; Liang et al., 2026; Xie et al., 2026a; Ge et al., 2026; Won et al., 2026; Zhang et al., 2026a). Our work addresses heterogeneous historical evidence requirements through *query-adaptive visual-semantic evidence allocation*. Query preferences guide visual selection and fuse three Event scores, each obtained by averaging static, dynamic or recency-sensitive frame relevance within an Event. Caption retrieval supplies complementary semantic support.

## 3 Method

### 3.1 Problem Formulation and Overview

We consider a question $q$ arriving at time $t_q$ in a video stream. Only observations available by $t_q$ can be used. The framework uses a frozen vision–language model (VLM) to determine whether history is needed, which temporal forms of visual evidence to prioritize, and which semantic context to retrieve.

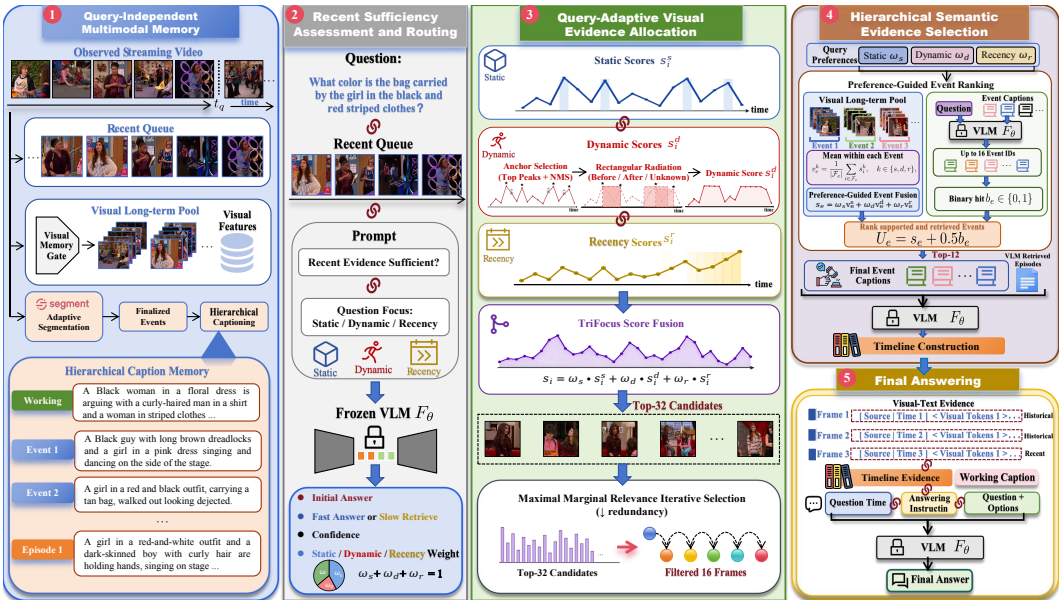


Figure 3: Query-adaptive visual-semantic evidence allocation through recent-evidence assessment and preference-guided frame and Event selection.

Given the question and up to four recent frames, the frozen VLM produces an initial answer, a Fast–Slow routing decision, and three integer focus scores from 1 to 5. We denote these scores by $\ell_s$, $\ell_d$ and $\ell_r$ for static, dynamic and recency-sensitive evidence, respectively. Higher scores indicate stronger query preference for the corresponding evidence type. Fast queries return the initial answer directly. Slow queries use these preferences to select historical frames and complementary semantic evidence. The routing and focus prompt is provided in Appendix F.1.

### 3.2 Query-Independent Multimodal Memory

**Visual memory.** For each sampled frame, we cache the native visual encoder's final visual tokens with its identity and timestamp for reuse during reasoning. The active visual pool holds at most 256 frames, while a separate recent queue preserves the latest observations.

**Hierarchical textual memory.** A *Working Caption* describes the ongoing event, an *Event Caption* summarizes a completed event, and an *Episode Caption* describes relationships across completed Events. Constructed without future queries, these summaries preserve long-range context beyond the bounded visual pool. Each Event Caption uses at most 16 temporally ordered representative frames. Memory updates are detailed in Appendix A.2.

### 3.3 Preference-Guided Historical Visual Evidence Selection

For Slow queries, historical candidates comprise all frames in the active visual pool except the eight most recent frames, which are supplied separately for final answering. If no historical candidate remains, the visual branch returns no historical frames, while semantic retrieval remains available. Let $i$ index a historical candidate and $t_i$ denote its timestamp in seconds.

We convert the VLM-predicted focus scores into weights $\omega_s$, $\omega_d$ and $\omega_r$ for static, dynamic and recency-sensitive relevance. With evidence-type index $k \in \{s, d, r\}$ and softmax temperature $\tau = 2$, the weights are

$$\omega_k = \frac{\exp(\ell_k/\tau)}{\exp(\ell_s/\tau) + \exp(\ell_d/\tau) + \exp(\ell_r/\tau)}. \tag{1}$$

They sum to one and guide both frame selection and subsequent Event ranking.

**Static evidence.** Static relevance measures the semantic match between a historical frame and the question. We separately mean-pool and L2-normalize the last-layer VLM hidden states over the frame's visual tokens and the question tokens, then compute their cosine similarity. Min–max

normalization across all historical candidates gives the static score $s_i^s$ in $[0, 1]$. If all similarities are equal, every static score is set to one.

**Dynamic evidence.** Actions and state changes require observations around relevant moments. We select up to six anchors by static relevance with temporal non-maximum suppression (NMS), then propagate their scores to neighboring candidates. For a question specifying "before," an anchor covers itself and the six preceding pool positions. For "after," it covers itself and the six following positions. Otherwise, it covers itself and four positions on either side. Positions refer to the temporally ordered complete visual pool, including recent frames excluded from retrieval.

The dynamic score $s_i^d$ is the maximum of the candidate's own static score and the static scores of all anchors whose neighborhoods cover it. Without a covering anchor, it retains its static score. This preserves direct matches while adding local temporal context.

**Recency-sensitive evidence.** Queries about the latest observed states or trajectories benefit from a recency preference over semantically relevant history. Let $T$ denote the observed video duration in seconds and $h = \max(8, 0.15T)$ the decay half-life in seconds. The recency-sensitive score $s_i^r$ is

$$s_i^r = s_i^s \left(0.5 + 0.5 \times 2^{-\max(0, t_q - t_i)/h}\right). \tag{2}$$

It downweights distant observations using only frames available by the query time.

**Fusion and diversification.** The fused relevance score $s_i$ combines the three evidence types according to the query preferences:

$$s_i = \omega_s s_i^s + \omega_d s_i^d + \omega_r s_i^r. \tag{3}$$

From the top 32 candidates, Maximal Marginal Relevance (MMR) (Carbonell & Goldstein, 1998) selects up to 16 complementary historical frames. Let $\mathcal{S}$ denote the selected frames and $u_i$ the L2-normalized SigLIP2 feature of frame $i$ (Tschannen et al., 2025). With temporal scale $\sigma = \mathrm{clip}(T/16, 2, 30)$ seconds, where $T$ is the observed video duration, semantic diversity $D_s$, temporal diversity $D_t$ and selection utility $U$ are defined as

$$\begin{aligned} D_s(i, \mathcal{S}) &= \min_{j \in \mathcal{S}} \frac{1 - u_i^\top u_j}{2}, \\ D_t(i, \mathcal{S}) &= \min_{j \in \mathcal{S}} \min\left(\frac{|t_i - t_j|}{\sigma}, 1\right), \\ U(i \mid \mathcal{S}) &= 0.80\, s_i + 0.10\, D_s(i, \mathcal{S}) + 0.10\, D_t(i, \mathcal{S}). \end{aligned} \tag{4}$$

Both diversity terms are zero when $\mathcal{S}$ is empty. We iteratively select the remaining candidate with highest utility until 16 frames are retained or candidates are exhausted. Selected frames remain fixed during semantic retrieval, while all historical candidates' three component scores are retained for Event scoring.

### 3.4 Preference-Guided Semantic Evidence Selection and Final Answering

The semantic branch scores Events from static, dynamic and recency-sensitive perspectives, then fuses these scores using the same query preferences as visual selection. VLM caption retrieval provides complementary semantic support. Only finalized Events and Episodes with valid, nonempty Captions available by the query time are eligible.

**Event scoring and preference-guided fusion.** For an eligible Event $e$, let $\mathcal{F}_e$ denote its historical candidate frames assigned by timestamp, and $|\mathcal{F}_e|$ their number. This set includes all retained frames within the Event except the eight recent frames, before Top-32 filtering or MMR. For evidence type $k \in \{s, d, r\}$, the Event score $v_e^k$ averages the corresponding frame scores:

$$v_e^k = \begin{cases} \dfrac{1}{|\mathcal{F}_e|} \displaystyle\sum_{i \in \mathcal{F}_e} s_i^k, & |\mathcal{F}_e| > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

These scores capture direct visual matches, local temporal context and recency-sensitive relevance, respectively. Averaging over retained candidates avoids directly favoring Events with more frames. The Event's visual relevance $V_e$ fuses these scores using the shared query-preference weights:

$$V_e = \omega_s v_e^s + \omega_d v_e^d + \omega_r v_e^r. \tag{6}$$

The weights are applied once, without further normalization.

**Semantic retrieval and ranking.** Given the question without options and chronologically ordered Event records, the frozen VLM retrieves up to 16 relevant Event IDs. Let $b_e$ indicate whether Event $e$ is retrieved, irrespective of retrieval order or duplicate hits. Events with candidate frames or retrieval hits are ranked by

$$U_e = V_e + 0.5\, b_e, \tag{7}$$

where $U_e$ is the ranking score and 0.5 is a fixed retrieval bonus. Events without candidate frames remain eligible through caption retrieval, with $V_e = 0$. We retain up to 12 highest-scoring Events without appending recent Events. Appendix A.4 details tie-breaking and validation.

**Timeline and final answering.** The VLM independently retrieves up to three Episode Captions. Selected Event and Episode Captions are ordered chronologically and organized into a Timeline of at most eight evidence items, with a 2,048-token generation limit. Both retrieval and Timeline construction use the question without options and exclude the Working Caption. For final answering, up to 16 selected historical frames and eight recent frames are deduplicated and interleaved chronologically with Timeline evidence and the latest available Working Caption. The frozen VLM answers using this context, the question and any answer options. Semantic selection leaves the visual selection unchanged. Appendix A.5 details input assembly, and Appendix F provides the prompts.

# 4 Experiments

## 4.1 Experimental Settings

**Backbone and benchmarks.** We evaluate frozen Qwen3-VL-8B-Instruct (Bai et al., 2025a) without task-specific training. OVO-Bench (Niu et al., 2025) covers six Real-Time, three Backward and three Forward tasks. Its Total Avg. averages the three category means. StreamingBench (Lin et al., 2024) covers 14 visual-only tasks, comprising ten Real-Time and four Contextual tasks. Its Total Avg. is the unweighted mean of their individual accuracies.

**Baselines.** Tables 1 and 2 compare offline and streaming methods reporting all required task scores. Offline baselines include Gemini 1.5 Pro (Gemini Team, 2024), GPT-4o (OpenAI, 2024), LLaVA-Video (Zhang et al., 2025d), LLaVA-OneVision (Li et al., 2025), Qwen2-VL (Wang et al., 2024a), Qwen2.5-VL with complete history, InternVL2 (OpenGVLab Team, 2024), MiniCPM-V-2.6 (Yao et al., 2024), LongVU (Shen et al., 2025) and LongVA (Zhang et al., 2025b). Streaming baselines include VideoLLM-online (Chen et al., 2024), Flash-VStream (Zhang et al., 2025a), Dispider (Qian et al., 2025), TimeChat-Online (Yao et al., 2025), StreamAgent (Yang et al., 2025a), StreamForest (Zeng et al., 2025), StreamingEval FIFO (Tang et al., 2026), Response-G1 (Ma et al., 2026), R3-Streaming (Liu et al., 2026), QUART (Won et al., 2026) and STRIDE (Kim et al., 2026).

Implementation details, evaluation protocols and prompts are provided in Appendices A, B and F.

## 4.2 Experimental Results

**OVO-Bench.** Our framework achieves **67.79** Total Avg., outperforming all compared methods (Table 1). Strong Real-Time and Backward results support adapting historical access to both current-state perception and history-dependent reasoning. These question types require different evidence scopes: recent observations can resolve the current scene, whereas tracing past events requires access to earlier states and transitions. The results are consistent with adapting evidence access to this distinction, with controlled comparisons provided below.

**StreamingBench.** Our framework reaches **77.01** across all 14 visual-only tasks and leads on 12 tasks (Table 2). Performance across Real-Time and Contextual tasks supports combining fine-grained visual observations with long-range semantic evidence. The breadth of these results suggests that

Table 1: Comparison with prior methods on the twelve tasks of OVO-Bench (accuracy, %).

| | | Real-Time | | | | | | Backward | | | Forward | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Frames | OCR | ACR | ATR | STU | FPD | OJR | EPM | ASI | HLD | REC | SSR | CRR | Total Avg. |
| *Offline video models* | | | | | | | | | | | | | | |
| Gemini 1.5 Pro (Gemini Team, 2024) | 1 fps | 85.91 | 66.97 | 79.31 | 58.43 | 63.37 | 61.96 | 58.59 | **76.35** | 52.64 | 35.53 | **74.24** | 61.67 | 63.00 |
| GPT-4o (OpenAI, 2024) | 64 | 69.80 | 64.22 | 71.55 | 51.12 | 70.30 | 59.78 | 57.91 | 75.68 | 48.66 | 27.58 | 73.21 | 59.40 | 59.54 |
| LLaVA-Video-7B (Zhang et al., 2025d) | 64 | 69.80 | 59.63 | 66.38 | 50.56 | 72.28 | 61.41 | 51.18 | 64.19 | 9.68 | 34.10 | 67.57 | 60.83 | 53.06 |
| LLaVA-OneVision-7B (Li et al., 2025) | 64 | 67.11 | 58.72 | 69.83 | 49.44 | 71.29 | 60.33 | 52.53 | 58.78 | 23.66 | 24.79 | 66.93 | 60.83 | 52.88 |
| Qwen2-VL-7B (Wang et al., 2024a) | 64 | 69.13 | 53.21 | 63.79 | 50.56 | 66.34 | 60.87 | 44.44 | 66.89 | 34.41 | 30.09 | 65.66 | 50.83 | 52.70 |
| InternVL2-8B (OpenGVLab Team, 2024) | 64 | 68.46 | 58.72 | 68.97 | 44.94 | 67.33 | 55.98 | 43.10 | 61.49 | 27.41 | 25.79 | 57.55 | 52.92 | 50.05 |
| LongVU-7B (Shen et al., 2025) | 1 fps | 55.70 | 49.54 | 59.48 | 48.31 | 68.32 | 63.04 | 43.10 | 66.22 | 9.14 | 16.62 | 69.00 | 60.00 | 48.48 |
| Qwen2.5-VL-7B (complete history) | 1 fps | 73.15 | 55.05 | 72.41 | 53.93 | 72.28 | 63.59 | 56.57 | 63.51 | 23.12 | 23.64 | 61.53 | 41.25 | 51.65 |
| *Streaming methods* | | | | | | | | | | | | | | |
| Flash-VStream-7B (Zhang et al., 2025a) | 1 fps | 25.50 | 32.11 | 29.31 | 33.71 | 29.70 | 28.80 | 36.36 | 33.78 | 5.91 | 5.44 | 67.25 | 60.00 | 33.15 |
| Dispider-7B (Qian et al., 2025) | 1 fps | 57.72 | 49.54 | 62.07 | 44.94 | 61.39 | 51.63 | 48.48 | 55.41 | 4.30 | 18.05 | 37.36 | 48.75 | 41.78 |
| TimeChat-Online-7B (Yao et al., 2025) | 1 fps | 75.20 | 46.80 | 70.70 | 47.80 | 69.30 | 61.40 | 55.90 | 59.50 | 9.70 | 31.60 | 38.50 | 40.00 | 46.76 |
| StreamAgent-7B (Yang et al., 2025a) | 1 fps | 71.20 | 53.20 | 63.60 | 53.90 | 67.30 | 58.70 | 54.80 | 58.10 | 25.80 | 35.90 | 48.40 | 52.00 | 50.99 |
| StreamForest-7B (Zeng et al., 2025) | 1 fps | 68.46 | 53.21 | 71.55 | 47.75 | 65.35 | 60.87 | 58.92 | 64.86 | 32.26 | 32.81 | 70.59 | 57.08 | 55.57 |
| StreamingEval FIFO (Tang et al., 2026) | – | 91.95 | 78.90 | 79.31 | 67.98 | 73.27 | 80.98 | 48.82 | 61.49 | 48.92 | 21.78 | 64.86 | 50.42 | 59.16 |
| Response-G1 (Ma et al., 2026) | – | 90.60 | 74.30 | 75.90 | 59.60 | 69.30 | 71.70 | 55.60 | 66.90 | 33.90 | 41.90 | 71.10 | **61.70** | 61.31 |
| R3-Streaming (7B/4B) (Liu et al., 2026) | 1 fps | 87.92 | 70.64 | 73.28 | 59.55 | 69.32 | 70.65 | 51.18 | 65.54 | 37.10 | **45.13** | 57.09 | 49.58 | 57.92 |
| Ours (Full) | $\leq 24$ | **95.97** | **85.32** | **82.76** | **73.03** | **74.26** | **85.33** | **61.62** | **76.35** | **68.28** | 30.80 | 71.38 | 53.33 | **67.79** |

Table 2: Comparison with prior methods on 14 visual-only StreamingBench tasks (accuracy, %).

| | Real-Time | | | | | | | | | | Contextual | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | OP | CR | CS | ATP | EU | TR | PR | SU | ACP | CT | ACU | MCU | SQA | PO | Total Avg. |
| *Offline video models* | | | | | | | | | | | | | | | |
| Gemini 1.5 Pro (Gemini Team, 2024) | 79.02 | 80.47 | 83.54 | 79.67 | 80.00 | 84.74 | 77.78 | 64.23 | 71.95 | 48.70 | 51.41 | 40.73 | 54.80 | 45.10 | 67.30 |
| GPT-4o (OpenAI, 2024) | 77.11 | 80.47 | 83.91 | 76.47 | 70.19 | 83.80 | 66.67 | 62.19 | 69.12 | 49.22 | 41.20 | 38.40 | 32.80 | 56.86 | 63.46 |
| LLaVA-OneVision-7B (Li et al., 2025) | 80.38 | 74.22 | 76.03 | 80.72 | 72.67 | 71.65 | 67.59 | 65.45 | 65.72 | 45.08 | 35.60 | 36.00 | 27.27 | 29.55 | 59.14 |
| Qwen2-VL-7B (Wang et al., 2024a) | 75.20 | **82.81** | 73.19 | 77.45 | 68.32 | 71.03 | 72.22 | 61.19 | 61.47 | 46.11 | 31.20 | 26.00 | 39.60 | 22.73 | 57.75 |
| MiniCPM-V-2.6-8B (Yao et al., 2024) | 71.93 | 71.09 | 77.92 | 75.82 | 64.60 | 65.73 | 70.37 | 56.10 | 62.32 | 53.37 | 34.00 | 31.60 | 41.92 | 22.22 | 57.07 |
| InternVL2-8B (OpenGVLab Team, 2024) | 68.12 | 60.94 | 69.40 | 77.12 | 67.70 | 62.93 | 59.26 | 53.25 | 54.96 | 56.48 | 32.00 | 31.20 | 32.32 | 40.91 | 54.76 |
| LongVA-7B (Zhang et al., 2025b) | 70.03 | 63.28 | 61.20 | 70.92 | 62.73 | 59.50 | 61.11 | 53.66 | 54.67 | 34.72 | 32.80 | 29.60 | 30.30 | 15.91 | 50.03 |
| Qwen2.5-VL-7B (complete history) | 83.38 | 78.91 | 85.49 | 83.01 | 75.16 | 79.13 | 77.78 | 66.26 | 68.27 | 48.70 | 32.40 | 34.00 | 46.00 | 18.00 | 62.61 |
| *Streaming methods* | | | | | | | | | | | | | | | |
| Flash-VStream-7B (Zhang et al., 2025a) | 25.89 | 43.57 | 24.91 | 23.87 | 27.33 | 13.08 | 18.52 | 25.20 | 23.87 | 48.70 | 24.80 | 25.20 | 26.80 | 1.96 | 25.26 |
| VideoLLM-online-8B (Chen et al., 2024) | 39.07 | 40.06 | 34.49 | 31.05 | 45.96 | 32.40 | 31.48 | 34.16 | 42.49 | 27.89 | 24.19 | 29.20 | 30.80 | 3.92 | 31.94 |
| Dispider-7B (Qian et al., 2025) | 74.92 | 75.53 | 74.10 | 73.08 | 74.44 | 59.92 | 76.14 | 62.91 | 62.16 | 45.80 | 39.62 | 27.65 | 34.80 | 25.34 | 57.60 |
| StreamAgent-7B (Yang et al., 2025a) | 79.63 | 78.31 | 79.28 | 75.87 | 74.74 | 76.92 | 82.94 | 66.31 | 73.69 | 55.40 | 39.72 | 30.25 | 39.60 | 28.90 | 62.97 |
| TimeChat-Online-7B (Yao et al., 2025) | 80.22 | 82.03 | 79.50 | 83.33 | 76.10 | 78.50 | 78.70 | 64.63 | 69.60 | **57.98** | 39.20 | 31.60 | 41.60 | 28.80 | 63.70 |
| QUART (Qwen3-VL-8B) (Won et al., 2026) | 77.78 | 73.44 | 87.07 | 88.78 | 72.96 | 84.42 | 78.70 | 72.36 | 74.43 | 53.19 | 34.80 | 36.80 | 43.60 | 46.00 | 66.02 |
| STRIDE (Qwen3-VL-8B) (Kim et al., 2026) | 77.10 | 75.00 | 77.30 | 72.80 | 76.40 | 77.90 | 76.90 | 69.90 | 84.30 | 46.10 | 32.00 | 38.40 | 46.40 | 42.80 | 63.81 |
| Ours (Full) | **88.56** | 73.44 | **95.90** | **92.16** | **80.75** | **94.08** | **87.96** | **79.27** | **84.99** | 50.26 | **64.00** | **66.00** | **62.40** | **58.40** | **77.01** |

the allocation strategy is useful across diverse evidence requirements. In particular, Contextual performance is consistent with semantic memory preserving information beyond the selected visual observations. Section 5 examines the contribution of each allocation decision and cross-backbone transfer. Appendix G illustrates how the evidence sources support individual answers.

# 5 Ablation and Discussion

## 5.1 Ablation Studies

**Historical access and complementary evidence.** Table 3 evaluates matched Qwen3-VL-8B variants. Recent-only controls favor Real-Time perception, whereas fixed historical retrieval improves Backward reasoning, motivating query-dependent historical access. Fast–Slow routing improves Total Avg. from 64.21 (Always-Slow) to 67.79 (Full), supporting recent-evidence assessment before invoking the shared retrieval pipeline. Adding dynamic and recency-sensitive relevance progressively improves the static-only variant, and semantic evidence yields the strongest overall result. These increments support two complementary choices: matching visual retrieval to the temporal structure of a question and supplementing selected frames with longer-range event context. Variant definitions appear in Appendix D.1.

Table 3: Controlled comparisons and component ablations on OVO-Bench (accuracy, %).

| | | Real-Time | | | | | | | Backward | | | | Forward | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variant | Frames | OCR | ACR | ATR | STU | FPD | OJR | Avg. | EPM | ASI | HLD | Avg. | REC | SSR | CRR | Avg. | Total Avg. |
| B0: Full prefix | 1 fps | 69.80 | 59.60 | 73.30 | 57.30 | 71.30 | 58.70 | 65.00 | 53.20 | 58.11 | 27.96 | 46.42 | 24.95 | 51.34 | 40.40 | 38.90 | 50.11 |
| B1: Recent-4 | 4 | 94.63 | 81.65 | 81.90 | 66.85 | **77.23** | 83.70 | 80.99 | 52.86 | 64.19 | 53.76 | 56.94 | 18.91 | 57.87 | 42.08 | 39.62 | 59.18 |
| B2: Recent-8 | 8 | 93.29 | 82.57 | 77.59 | 62.36 | 76.24 | 83.15 | 79.20 | 53.87 | 65.54 | 52.15 | 57.19 | 20.77 | 60.41 | 45.42 | 42.20 | 59.53 |
| B3: Full pool + Top-16 | 16 | 69.13 | 58.72 | 59.48 | 47.19 | 64.36 | 46.74 | 57.60 | 52.53 | 59.46 | **76.88** | 62.96 | 30.23 | 50.08 | 40.00 | 40.10 | 53.55 |
| B4: Pool256 + Top-16 | 16 | 69.13 | 58.72 | 69.83 | 48.31 | 69.31 | 51.63 | 61.16 | 50.17 | 63.51 | 75.81 | 63.16 | 25.93 | 57.55 | 40.42 | 41.30 | 55.21 |
| Always-Slow | $\leq 24$ | 87.25 | 81.65 | 77.59 | 65.17 | 73.27 | 78.26 | 77.20 | 58.25 | 71.62 | 70.97 | 66.95 | **31.38** | 68.04 | 45.99 | 48.47 | 64.21 |
| A0: Routing + static | $\leq 24$ | 91.28 | 82.57 | 80.17 | 69.66 | 73.27 | 82.61 | 79.93 | 58.25 | 67.57 | 65.59 | 63.80 | 27.51 | **71.86** | 46.67 | 48.68 | 64.14 |
| A1: A0 + dynamic | $\leq 24$ | 93.96 | 84.40 | **83.62** | **73.03** | 71.29 | 85.87 | 82.03 | 58.59 | 71.62 | 65.59 | 65.27 | 27.65 | 70.11 | 49.58 | 49.11 | 65.47 |
| A2: A1 + recency | $\leq 24$ | 95.30 | 84.40 | **83.62** | **73.03** | 72.28 | **86.41** | 82.51 | 61.28 | 73.65 | 66.13 | 67.02 | 27.65 | 71.22 | 52.92 | 50.60 | 66.71 |
| A3: A2 + text (Full) | $\leq 24$ | **95.97** | **85.32** | 82.76 | **73.03** | 74.26 | 85.33 | **82.78** | **61.62** | **76.35** | 68.28 | **68.75** | 30.80 | 71.38 | **53.33** | **51.84** | **67.79** |

Table 4: Effects of visual and Event focus weights on OVO-Bench (accuracy, %).

| | Focus weights | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Setting | Visual | Event | Real-Time | EPM | ASI | HLD | Backward Avg. | Total Avg. |
| A | Equal | Equal | 81.87 | 59.93 | 72.97 | 67.74 | 66.88 | 76.87 |
| B | Predicted | Equal | 81.96 | 60.61 | 74.32 | 66.67 | 67.20 | 77.04 |
| C | Equal | Predicted | 82.25 | 60.61 | 75.68 | 67.20 | 67.83 | 77.44 |
| D (Full) | Predicted | Predicted | **82.78** | **61.62** | **76.35** | **68.28** | **68.75** | **78.10** |

**Shared query preferences.** Table 4 independently varies the weights for visual frame selection and fusion of the three Event scores. Adapting Event weights alone improves Total Avg. from 76.87 to 77.44, while adapting both branches achieves **78.10** and the best results in every reported column. These results support using shared query preferences to coordinate visual and semantic evidence selection. Backward Avg. averages EPM, ASI and HLD, while Total Avg. equally weights the six Real-Time and three Backward tasks. Detailed settings appear in Appendix D.2.

**Routing and evidence preferences.** Figure 4 shows that 91.16% of Real-Time questions use Fast, whereas 55.94% of Backward questions use Slow. Static scores meet or exceed dynamic scores for most EPM questions, with the reverse pattern for ASI, consistent with allocating evidence according to question requirements. Together, routing and focus scores express complementary aspects of evidence demand: whether history is needed and which temporal structure should receive greater emphasis. Detailed counts appear in Appendix D.3.

### 5.2 Cross-Backbone Generalization

We evaluate cross-backbone generalization on OVO-Bench using Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), and LLaVA-OneVision-2-8B (An et al., 2026), with additional StreamingBench evaluation on the two Qwen backbones. Tables 5 and 6 show consistent improvements over the corresponding base models and compared memory methods across three backbones on OVO-Bench and two on StreamingBench. This supports the portability of training-free evidence

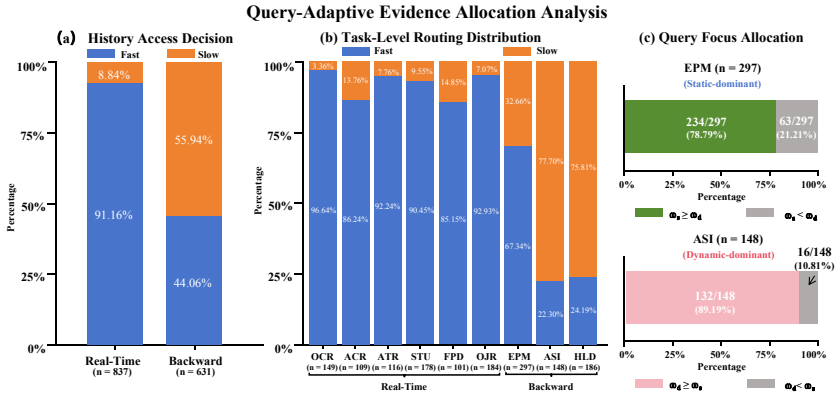


Figure 4: Routing decisions and shared evidence preferences on OVO-Bench.

Table 5: Cross-backbone comparison on OVO-Bench Real-Time and Backward tasks (accuracy, %).

| Backbone | Method | Frames | Real-Time | Backward | Avg.-9 |
|---|---|---|---|---|---|
| Qwen2.5-VL-7B | Base model | 0.5 fps | 61.98 | 50.85 | 58.27 |
| | + HERMES (4K) | 0.5 fps | 68.98 | 49.42 | 62.46 |
| | + FluxMem | 1 fps | 67.20 | 47.23 | 60.54 |
| | + OASIS | 48/80 | 68.06 | 53.18 | 63.10 |
| | + Ours | $\leq 24$ | **78.39** | **58.16** | **71.65** |
| Qwen3-VL-8B | Base model | 0.5 fps | 67.15 | 51.85 | 62.05 |
| | + OASIS | 48/80 | 77.36 | 59.14 | 71.29 |
| | + Ours (Full) | $\leq 24$ | **82.78** | **68.75** | **78.10** |
| LLaVA-OneVision-2-8B | Base model | 0.5 fps | 62.08 | 45.67 | 56.61 |
| | + OASIS | 48/80 | 64.12 | 47.44 | 58.56 |
| | + Ours | $\leq 24$ | **81.13** | **53.74** | **72.00** |

Table 6: Cross-backbone results on 13 StreamingBench tasks (accuracy, %).

| Backbone | Method | Frames | Real-Time | ACU | MCU | SQA | Avg.-13 |
|---|---|---|---|---|---|---|---|
| Qwen2.5-VL-7B | Base model | 0.5 fps | 66.00 | 33.20 | 30.10 | 34.00 | 58.25 |
| | + OASIS | 48/80 | 70.60 | 38.80 | 36.00 | 46.00 | 63.60 |
| | + Ours | $\leq 24$ | **77.34** | **65.20** | **62.40** | **54.40** | **73.49** |
| Qwen3-VL-8B | Base model | 0.5 fps | 72.80 | 35.60 | 35.70 | 43.90 | 64.86 |
| | + OASIS | 48/80 | 78.20 | 42.70 | 49.60 | 48.40 | 70.98 |
| | + Ours | $\leq 24$ | **82.74** | **64.00** | **66.00** | **62.40** | **78.45** |

allocation. Improvements across distinct frozen backbones suggest that organizing evidence around query requirements can complement different underlying visual-language capabilities without task-specific parameter updates. Avg.-9 and Avg.-13 are task-level means over the evaluated subsets, distinct from the main-table aggregates (Appendix C).

### 5.3 Inference Costs

Table 7 reports **76.09%** micro accuracy with 9.84 answering frames per query and 27.19 GB peak memory. This accuracy weights each question equally across Real-Time and Backward tasks, whereas the **78.10%** in Table 4 equally weights the nine tasks. Relative to full-prefix input, selected evidence reduces input tokens and memory while improving accuracy. The higher accuracy than fixed retrieval at similar peak memory further supports the value of evidence composition within a compact answering context. Answering frames count inputs to the initial routing-and-answering call and, for Slow queries, the final-answer call, while latency covers the broader retrieval and reasoning pipeline. Measurement boundaries and metric definitions appear in Appendix E.

Table 7: Inference costs on OVO-Bench Real-Time and Backward tasks with Qwen3-VL-8B-Instruct.

| Method | Latency (s) ↓ | Avg. Tokens ↓ | Answer Frames ↓ | Peak Mem. (GB) ↓ | Acc. ↑ |
|---|---|---|---|---|---|
| Recent-8 | **0.324** | **1,932** | **7.86** | **17.70** | 69.07 |
| Full Prefix | 3.650 | 34,203 | 98.36 | 76.59 | 56.68 |
| Fixed Retrieval | 3.173 | 22,747 | 15.42 | 27.20 | 60.15 |
| OASIS (Liang et al., 2026) | 5.160 | 9,654 | 56.65 | 28.48 | 69.14 |
| Ours | 3.828 | 12,873 | 9.84 | 27.19 | **76.09** |

## 6 Limitations and Conclusion

**Limitations.** The framework depends on reliable routing and Caption content, while sparse visual coverage can bias Event scores. Incorrect sufficiency judgments may suppress necessary retrieval, and generated summaries may omit relevant details. Further study is needed on sensitivity to the retrieval bonus and Event budget and textual-memory growth in long-running streams.

**Conclusion.** We formulate streaming video understanding as *query-adaptive visual-semantic evidence allocation*. The training-free framework assesses recent-evidence sufficiency and adapts visual selection to static states, dynamic context and recency-sensitive evidence. The semantic branch separately averages these relevance components into three Event scores and fuses them using the same query preferences. Caption retrieval and Episode summaries provide complementary long-range context for reasoning with the selected visual observations. Results on OVO-Bench and the 14 visual-only tasks of StreamingBench, together with controlled ablations, support adapting historical access and evidence composition to each query.

## AI USE STATEMENT

Generative-AI assistants were used to support language polishing, literature retrieval and discovery, research ideation and execution, and the drafting and revision of manuscript sections. These uses included improving clarity and academic style, identifying related work, discussing methodological and experimental choices, and assisting with manuscript preparation and consistency checks. The authors reviewed and verified all AI-assisted content, including technical descriptions, citations, and reported numerical results, and take full responsibility for the scientific validity, originality, and accuracy of this submission.

## REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. LLaVA-OneVision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026. doi: 10.48550/arXiv.2605.25979. URL https://arxiv.org/abs/2605.25979.

Wei Ao, Lan Wang, and Vishnu Naresh Boddeti. QSVideo: Query-conditioned semantic temporal retrieval for video understanding. *arXiv preprint arXiv:2607.04559*, 2026. doi: 10.48550/arXiv.2607.04559. URL https://arxiv.org/abs/2607.04559. Accepted to ECCV 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, Junyang Lin, An Yang, Binyuan Hui, Bowen Yu, Chen Cheng, Dayiheng Liu, Fan Hong, Fei Huang, Jiawei Liu, Jin Xu, Jianhong Tu, Jianyuan Zeng, Jie Zhang, Jinkai Wang, Jianwei Zhang, Jingren Zhou, Kexin Yang, Mei Li, Ming Yan, Na Ni, Rui Men, Songtao Jiang, Xiaodong Deng, Xiaoming Huang, Ximing Zhou, Xingzhang Ren, Yang Fan, Yichang Zhang, Yikai Zhu, Yuqiong Liu, and Zhifang Guo. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. doi: 10.48550/arXiv.2502.13923. URL https://arxiv.org/abs/2502.13923.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336. ACM, 1998. doi: 10.1145/290941.291025. URL https://doi.org/10.1145/290941.291025.

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*, pp. 18407–18418, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_VideoLLM-online_Online_Video_Large_Language_Model_for_Streaming_Video_CVPR_2024_paper.html.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, 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*, 2025. URL https://arxiv.org/abs/2503.00540.

Haonan Ge, Yiwei Wang, Hang Wu, and Yujun Cai. What should a streaming video model remember? *arXiv preprint arXiv:2606.16353*, June 2026. doi: 10.48550/arXiv.2606.16353. URL https://arxiv.org/abs/2606.16353.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. doi: 10.48550/arXiv.2403.05530. URL https://arxiv.org/abs/2403.05530.

Xinqi Jin, Hanxun Yu, Bohan Yu, Kebin Liu, Jian Liu, Keda Tao, Yixuan Pei, Huan Wang, Fan Dang, Jiangchuan Liu, and Weiqiang Wang. StreamingAssistant: Efficient visual token pruning for accelerating online video understanding. *arXiv preprint arXiv:2512.12560*, 2025. doi: 10.48550/arXiv.2512.12560. URL https://arxiv.org/abs/2512.12560.

Junho Kim, Hosu Lee, James M. Rehg, Minsu Kim, and Yong Man Ro. STRIDE: When to speak meets sequence denoising for streaming video understanding. *arXiv preprint arXiv:2603.27593*, 2026. doi: 10.48550/arXiv.2603.27593. URL https://arxiv.org/abs/2603.27593.

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*, 2025. URL https://arxiv.org/abs/2506.15745.

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*, 2025. URL https://arxiv.org/abs/2408.03326.

Jialuo Li, Bin Li, Jiahao Li, and Yan Lu. Divide, then ground: Adapting frame selection to query types for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11369–11380, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Divide_then_Ground_Adapting_Frame_Selection_to_Query_Types_for_CVPR_2026_paper.html.

Zhijia Liang, Jiaming Li, Weikai Chen, Yanhao Zhang, Haonan Lu, and Guanbin Li. OASIS: On-demand hierarchical event memory for streaming video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2821–2831, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liang_OASIS_On-Demand_Hierarchical_Event_Memory_for_Streaming_Video_Reasoning_CVPR_2026_paper.html.

Jingyang Lin, Jialian Wu, Jiang Liu, Ximeng Sun, Ze Wang, Xiaodong Yu, Jiebo Luo, Zicheng Liu, and Emad Barsoum. VideoSeek: Long-horizon video agent with tool-guided seeking. *arXiv preprint arXiv:2603.20185*, 2026. URL https://arxiv.org/abs/2603.20185.

Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the gap for MLLMs to achieve streaming video understanding. *arXiv preprint arXiv:2411.03628*, 2024. doi: 10.48550/arXiv.2411.03628. URL https://arxiv.org/abs/2411.03628.

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*, May 2026. doi: 10.48550/arXiv.2605.17921. URL https://arxiv.org/abs/2605.17921.

Ke Ma, Jiaqi Tang, Bin Guo, Xueting Han, Ruonan Xu, Qingfeng He, Ziheng Wang, Xu Wang, Qifeng Chen, Zhiwen Yu, and Yunhao Liu. Response-G1: Explicit scene graph modeling for proactive streaming video understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44139–44153. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2042. URL https://aclanthology.org/2026.acl-long.2042/.

Junbo Niu, Yifei Li, 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*, pp. 18902–18913, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Niu_OVO-Bench_How_Far_is_Your_Video-LLMs_from_Real-World_Online_Video_CVPR_2025_paper.html.

OpenAI. GPT-4o system card, 2024. URL https://openai.com/index/gpt-4o-system-card/.

OpenGVLab Team. InternVL2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. Official model release, 2024. URL https://internvl.github.io/blog/2024-07-02-InternVL-2.0/.

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*, pp. 24045–24055, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Qian_Dispider_Enabling_Video_LLMs_with_Active_Real-Time_Interaction_via_Disentangled_CVPR_2025_paper.html.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599. PMLR, 2025. URL https://proceedings.mlr.press/v267/shen25j.html.

Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. A simple baseline for streaming video understanding. *arXiv preprint arXiv:2604.02317*, 2026. doi: 10.48550/arXiv.2604.02317. URL https://arxiv.org/abs/2604.02317.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-XL: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26160–26169, 2025. doi: 10.1109/CVPR52734.2025.02436. URL https://openaccess.thecvf.com/content/CVPR2025/html/Shu_Video-XL_Extra-Long_Vision_Language_Model_for_Hour-Scale_Video_Understanding_CVPR_2025_paper.html.

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*, pp. 18221–18232, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Song_MovieChat_From_Dense_Token_to_Sparse_Memory_for_Long_Video_CVPR_2024_paper.html.

Guowei Tang, Tianwen Qian, Huanran Zheng, Yifei Wang, and Xiaoling Wang. StreamingEval: A unified evaluation protocol towards realistic streaming video understanding. *arXiv preprint arXiv:2603.21493*, 2026. doi: 10.48550/arXiv.2603.21493. URL https://arxiv.org/abs/2603.21493.

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*, pp. 18992–19001, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tao_DyCoke_Dynamic_Compression_of_Tokens_for_Fast_Video_Large_Language_CVPR_2025_paper.html.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. doi: 10.48550/arXiv.2502.14786. URL https://arxiv.org/abs/2502.14786.

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*, 2025. URL https://arxiv.org/abs/2505.05467.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a. doi: 10.48550/arXiv.2409.12191. URL https://arxiv.org/abs/2409.12191.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, 2024b. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/10325_ECCV_2024_paper.php.

Junil Won, Jeong Hun Yeo, and Yong Man Ro. QUART: Query-aware context routing for streaming video understanding. *OpenReview preprint*, 2026. URL https://openreview.net/forum?id=2WIGdnImpX.

Junlin Xie, Quanlong Zheng, Ruifei Zhang, Kuo Wang, Yanhao Zhang, Jinguo Luo, Haonan Lu, Xiang Wan, and Guanbin Li. StreamRAG: Enhancing real-time video understanding with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38870–38879, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Xie_StreamRAG_Enhancing_Real-Time_Video_Understanding_with_Retrieval_Augmentation_CVPR_2026_paper.html.

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, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Xie_FluxMem_Adaptive_Hierarchical_Memory_for_Streaming_Video_Understanding_CVPR_2026_paper.pdf.

Haolin Yang, Feilong Tang, Lingxiao Zhao, Xinlin Zhuang, Yifan Lu, Xiang An, Ming Hu, Xiaofeng Zhang, Abdalla Swikir, Junjun He, Zongyuan Ge, Muhammad Haris Khan, and Imran Razzak. StreamAgent: Towards anticipatory agents for streaming video understanding. *arXiv preprint arXiv:2508.01875*, 2025a. doi: 10.48550/arXiv.2508.01875. URL https://arxiv.org/abs/2508.01875.

Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. StreamMem: Query-agnostic KV cache memory for streaming video understanding. *arXiv preprint arXiv:2508.15717*, 2025b. doi: 10.48550/arXiv.2508.15717. URL https://arxiv.org/abs/2508.15717.

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. *arXiv preprint arXiv:2504.17343*, 2025. doi: 10.48550/arXiv.2504.17343. URL https://arxiv.org/abs/2504.17343.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM-V: A GPT-4V level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. doi: 10.48550/arXiv.2408.01800. URL https://arxiv.org/abs/2408.01800.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8579–8591, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ye_Re-thinking_Temporal_Search_for_Long-Form_Video_Understanding_CVPR_2025_paper.html.

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*, 2025. URL https://arxiv.org/abs/2509.24871.

Ce Zhang, Jing Bi, Jinxi He, Jianshu Zhang, Jingyang Lin, Yunzhong Xiao, Minghao Fu, Yaqi Xie, Zhentao Xie, Weicong Chen, Katia Sycara, and Ming Zhou. StreamScout: Learning when to look deeper for streaming video understanding. *arXiv preprint arXiv:2609.00291*, 2026a. URL https://arxiv.org/abs/2609.00291.

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*, pp. 21059–21069, 2025a. doi: 10.1109/ICCV51701.2025.01957. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_Flash-VStream_Efficient_Real-Time_Understanding_for_Long_Video_Streams_ICCV_2025_paper.html.

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 (Volume 1: Long Papers)*, pp. 8411–8430. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.381. URL https://aclanthology.org/2026.acl-long.381/.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *Transactions on Machine Learning Research*, 2025b. URL https://mlanthology.org/tmlr/2025/zhang2025tmlr-long/.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-Frame: Query-aware frame selection and multi-resolution adaptation for video-LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22056–22065, 2025c. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_Q-Frame_Query-aware_Frame_Selection_and_Multi-Resolution_Adaptation_for_Video-LLMs_ICCV_2025_paper.html.

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*, 2025d. URL https://mlanthology.org/tmlr/2025/zhang2025tmlr-llavavideo/.

# A Implementation Details

## A.1 Inference Configuration and Visual-Memory Update

Inference uses BF16, SDPA attention, batch size 1 and greedy decoding. Videos are sampled at 1 fps, retaining decoded timestamps. The initial routing and answering call uses up to four of the latest available frames.

In addition to the cached representations defined in Section 3.2, we retain the visual grid and an L2-normalized global feature obtained by averaging the final visual tokens. A new frame is compared with the most recently retained pool entry using global-feature cosine similarity. Frames with similarity below 0.86 are inserted as scene changes. Similarities in $[0.86, 0.94)$ are treated as state changes and are also inserted. When similarity is at least 0.94, insertion is suppressed unless more than 6 s have elapsed since the previous retained frame, in which case insertion is forced as a state update. When the pool exceeds 256 entries, the oldest entry is removed using FIFO eviction. This bound applies only to the searchable visual history and not to Caption metadata or the recent queue.

## A.2 Hierarchical Caption Memory

The program maintains IDs, temporal intervals, source-frame identities and Episode–Event links. Caption generation uses only the observed content, without question, answer-option or label input. Table 8 summarizes the generation limits for the three memory levels.

Table 8: Generation limits for hierarchical caption memory.

| Caption | Prompt word limit | Output token limit |
|---|---|---|
| Working | 40 | 128 |
| Event | 80 | 256 |
| Episode | 80 | 256 |

**Online Event segmentation.** Let $g_i$ denote the normalized global feature of frame $i$. For each available offset $d \in \{1, 2, 4, 8\}$, we compute $1 - g_i^\top g_{i-d}$ and combine the values using weights $\{0.4, 0.3, 0.2, 0.1\}$, renormalized over the available offsets, to obtain novelty $\nu_i$. Let $\mathcal{N}_i$ contain at most the 60 novelty scores preceding frame $i$. Here $\mathrm{median}$ denotes the median and $\mathrm{MAD}$ the median absolute deviation from it. Define the robust novelty score $z_i$ as

$$z_i = \max\left(0, \frac{\nu_i - \mathrm{median}(\mathcal{N}_i)}{\max(1.4826\,\mathrm{MAD}(\mathcal{N}_i),\ 0.01,\ 0.10|\,\mathrm{median}(\mathcal{N}_i)|)}\right). \tag{8}$$

The segmenter uses minimum, target, and maximum durations of 4, 12, and 24 s, respectively, a 6 s timestamp-gap boundary, two-frame lookahead, a two-frame cooldown, and novelty-score thresholds $3.5/2.0/6.0$ for strong, confirmation, and hard changes. After the minimum duration, a hard change or a strong local peak may create a boundary. After the target duration, two consecutive confirmation frames can close the Event. At the maximum duration, the best already observed candidate boundary is selected, or the Event is closed at the duration limit if no candidate exists. The lookahead only delays boundary confirmation and never accesses observations beyond the current causal cutoff.

Representative frames always include the first and last observations and the maximum-novelty observation. Additional frames are added from uniformly distributed temporal positions until the budget is filled. Up to two state-change frames may also be included, with a 4 s temporal suppression radius. Fewer representative frames are used when the Event does not fill the budget.

**Working and Episode updates.** A Working Caption becomes eligible after the open Event has lasted 8 s and is normally refreshed every 8 s. A strong novelty score with $z_i \geq 3.5$ can trigger an earlier refresh. An Episode contains 3–6 consecutive finalized Events, targets four Events, and spans at most 90 s. A timestamp gap above 6 s, the span or count limit, or a sufficiently large semantic change between adjacent Event Captions may close an Episode. The target of four Events is not a fixed grouping rule. The arrival of a query does not forcibly close the currently open Event.

Table 9: Temporal ordering of evidence in final answering.

| Block | Ordering time |
|---|---|
| Visual frame | frame timestamp |
| Timeline evidence | source-interval start time |
| Working Caption | snapshot/end time |

### A.3 Visual Relevance and Frame Selection

Static relevance uses the raw question without answer options and cached visual tokens. Focus scores receive no keyword-based calibration, and fused relevance is not renormalized after Top-32 filtering. Anchor NMS processes historical candidates in descending order of static relevance, retaining at most six anchors and suppressing further anchors within a 2 s radius of each selected anchor. This timestamp-based suppression is distinct from the pool-position neighborhoods used for dynamic relevance propagation. MMR uses SigLIP2-base-patch16-384. Utility ties are broken by fused relevance in descending order, timestamp in ascending order and finally frame ID in ascending order. Selected frames are restored to chronological order.

### A.4 Preference-Guided Event Ranking and Timeline Construction

The following details supplement the ranking rule in Section 3.4. Event records require valid IDs and temporal intervals. Frames are assigned through the segmenter's timestamp-to-Event mapping. Each retained candidate contributes separately to the static, dynamic and recency-sensitive Event means defined in Section 3.4. An Event with assigned candidates remains eligible even when its fused visual relevance is zero.

**Retrieval records and ranking ties.** Each retrieval record contains `id`, `start_time`, `end_time` and `caption`. Returned IDs are validated against the respective Event or Episode input records before deduplication. Equal Event scores are ordered by end time descending, start time descending and ID ascending.

**Timeline validation.** Up to 12 selected Event Captions and three independently retrieved Episode Captions are supplied to the Timeline organizer, with a 2,048-token generation limit. The organizer additionally requests at most 160 words of observation text. Output items must use exact source IDs. The program attaches timestamps, rejects illegal IDs or invalid fields, and enforces the eight-item limit. Retrieved Episodes can supply context when no Event qualifies. If both record sets are empty, the Timeline is empty, without padding.

### A.5 Final Answer Input Assembly and Cached Visual Tokens

For Slow queries, up to 16 selected historical frames are merged with up to eight recent frames and deduplicated. The resulting visual input is interleaved with Timeline evidence and the latest available Working Caption. Table 9 specifies the timestamp assigned to each evidence type. Missing frames are never filled from outside the visual pool.

When timestamps tie, the order is frame, Timeline, then Working Caption. The question and answer options are appended after all evidence blocks.

Let $g_t$, $g_h$ and $g_w$ denote the temporal, height and width dimensions of a cached frame's visual grid, and $m$ its spatial merge factor. The image-placeholder count $n_i$ is

$$n_i = \frac{g_t g_h g_w}{m^2}. \tag{9}$$

The implementation verifies that $n_i$ equals the number of cached final visual tokens before those tokens are inserted into the corresponding image positions. For Qwen3-VL, position indices are recomputed jointly for the complete mixed sequence through the native `get_rope_index` interface, using the actual token IDs, multimodal token types and image grids. Final answering uses cached final visual tokens without DeepStack intermediate features or pixel re-encoding.

### A.6 Structured Outputs and Failure Handling

The initial routing and answering call, Event retrieval, Episode retrieval, and Timeline construction use JSON outputs. Event and Episode retrieval and Timeline construction allow one formatting retry after a schema failure. The retry reuses the original prompt, adds a schema reminder, and keeps the same generation budget. Complete Markdown fences, a single trailing semicolon or quote, and a complete bare list may be removed or wrapped without altering their content. Truncated JSON, duplicate keys, missing required fields, and incorrect value types are never repaired by inventing content.

Returned IDs are checked against the causally visible memory snapshot and duplicates are removed. A valid empty list is distinguished from a parsing or backend failure, which is recorded as an error rather than silently interpreted as empty evidence. The initial routing and answering call does not use the schema-retry mechanism. An unparseable initial response activates the Slow route, while focus-score parsing failures fall back to equal weights.

## B Evaluation Protocol and Metric Aggregation

Evaluation follows the provided OVO-Bench chunks and causally available StreamingBench prefixes. Video observations and memory records used for each question are restricted to those available by its causal cutoff.

The main-table metrics are defined in Section 4.1. For the nine-task comparisons, let $A_{\mathrm{RT}}$ and $A_{\mathrm{Bwd}}$ denote the means over six Real-Time tasks and three Backward tasks (EPM, ASI and HLD), respectively. The task-level macro-average is

$$\text{Avg.-9} = \frac{6A_{\mathrm{RT}} + 3A_{\mathrm{Bwd}}}{9}. \tag{10}$$

This is also the Total Avg. reported in Table 4.

For StreamingBench, let $A_{\mathrm{RT}}$ denote the ten-task Real-Time mean and $A_c$ the accuracy of contextual task $c$. The cross-backbone score excludes PO:

$$\text{Avg.-13} = \frac{10A_{\mathrm{RT}} + A_{\mathrm{ACU}} + A_{\mathrm{MCU}} + A_{\mathrm{SQA}}}{13}.$$

Task accuracies are aggregated before rounding category summaries where available.

### B.1 Baselines and Comparison Sources

The complete baseline lists appear in Section 4.1. Published methods retain their reported backbones and budgets. Frames denotes the sampling rate or frame count, and our bound refers to one final-answer call.

QUART (Won et al., 2026) and STRIDE (Kim et al., 2026) report all 14 visual-only StreamingBench task scores in their respective Table 2. We use their Qwen3-VL-8B downstream configurations, including STRIDE's trained activation model. Their Total Avg. values are recomputed from individual task accuracies under our 14-task macro-average. QUART's original Overall averages category summaries, whereas STRIDE's evaluation also includes Omni-Source tasks. Aggregating the individual scores ensures a common task weighting in our comparison.

## C Cross-Backbone Evaluation Settings

The comparisons in Section 5.2 use frozen backbones and method-specific sampling settings reported in each table. Qwen3 Full rows reuse the corresponding main-table task scores. The evaluated subsets and aggregation rules are specified in Appendix B.

# D Ablations of Query-Adaptive Evidence Allocation

## D.1 Controlled Variants and Component Settings

All rows in Table 3 use Qwen3-VL-8B-Instruct. B0 processes the complete observed prefix; B1/B2 answer from the latest four/eight frames. B3 retrieves Top-16 frames from the complete 1-fps history, whereas B4 retrieves Top-16 from Pool256. Neither B3 nor B4 appends recent frames. These controls do not use Fast–Slow routing.

A0 combines Fast–Slow routing with static visual retrieval. A1 adds dynamic relevance, A2 adds recency-sensitive relevance (abbreviated as "recency" in the table), and A3 adds the semantic branch of Section 3.4. A3 is Full. Always-Slow retains the initial VLM call to predict query preferences but bypasses Fast answering, invoking final answering with retrieved evidence for every query. It uses the same preference normalization, historical candidate pool, retrieval procedure, Caption Memory and final-answer prompt as Full. This comparison isolates the decision to answer directly from recent evidence while preserving preference-guided retrieval. The semantic-branch increment evaluates Event ranking, Episode context and Timeline construction jointly. Reported means preserve source precision, so recomputation from rounded task scores can differ by 0.01.

## D.2 Shared Query Preferences for Visual and Semantic Evidence

In Table 4, Equal assigns weight $1/3$ to each of the three relevance components, while Predicted uses the normalized VLM-predicted focus scores. The Visual setting controls frame-score fusion before selection. The Event setting controls the weighted fusion of the three Event scores obtained by averaging each relevance component separately over the Event's historical candidates. Caption retrieval supplies the additive semantic signal in Event ranking. Across all four settings, routing decisions, the three unfused frame scores, Caption memory, Event and Episode retrieval hits, candidate pools and evidence budgets are held fixed. Only the Visual and Event fusion weights vary, allowing their effects on evidence selection to be assessed separately. Setting D corresponds to Full. Metric aggregation follows Appendix B.

## D.3 Routing Decisions and Evidence Preferences

Table 10 gives the query counts underlying the routing percentages in Figure 4. The Fast route handles 763/837 Real-Time questions (91.16%), while 353/631 Backward questions (55.94%) take the Slow route. These distributions support assessing recent evidence sufficiency for each query rather than assigning a fixed route by task.

Table 10: Per-task routing counts on OVO-Bench.

| Category | Task | Queries | Fast | Slow |
|---|---|---|---|---|
| Real-Time | OCR | 149 | 144 | 5 |
| | ACR | 109 | 94 | 15 |
| | ATR | 116 | 107 | 9 |
| | STU | 178 | 161 | 17 |
| | FPD | 101 | 86 | 15 |
| | OJR | 184 | 171 | 13 |
| | Subtotal | 837 | 763 | 74 |
| Backward | EPM | 297 | 200 | 97 |
| | ASI | 148 | 33 | 115 |
| | HLD | 186 | 45 | 141 |
| | Subtotal | 631 | 278 | 353 |
| All | Total | 1,468 | 1,041 | 427 |

For the preference comparisons, 234/297 EPM questions (78.79%) satisfy $\omega_s \geq \omega_d$, while the remaining 63 (21.21%) satisfy $\omega_s < \omega_d$. For ASI, 132/148 questions (89.19%) satisfy $\omega_d \geq \omega_s$, and the remaining 16 (10.81%) satisfy $\omega_d < \omega_s$. These pairwise comparisons include ties and describe static versus dynamic emphasis within the shared three-component preference vector used for both frame and Event selection.

# E INFERENCE-COST PROTOCOL

Measurements for Table 7 use eight NVIDIA A800 GPUs processing independent samples in parallel, with each sample executed on a single GPU. Peak Mem. reports the single-GPU peak memory footprint. Fixed Retrieval corresponds to B4 in Table 3, using Top-16 retrieval from Pool256 without appending recent frames.

Latency, input tokens and answering frames are averaged over queries. Latency includes the initial routing and answering call, retrieval, Caption construction, Timeline construction and final answering when invoked, but excludes video visual encoding and other offline memory-construction costs. Avg. Tokens sums input tokens over the included model calls. Answer Frames counts inputs to the initial routing and answering call and, for Slow queries, the final-answer call. Retrieval and caption-generation inputs are excluded.

Accuracy uses the sample-level micro aggregation defined in Section 5.3. The Full result uses the same final predictions as the main OVO-Bench table and Table 3, restricted to the 1,468 Real-Time and Backward questions. Its 1,117 correct answers yield 76.09% micro accuracy, while averaging the nine task accuracies equally yields 78.10% in Table 4.

# F METHOD PROMPTS

This section lists only prompts introduced by the proposed framework. Dataset-specific question wrappers follow the official OVO-Bench and StreamingBench evaluation protocols and are not repeated.

## F.1 ROUTING AND FOCUS PROMPT

The user message begins with the following sentence:

```
You are a video reasoning agent and given
a question and visual frames.
```

It is followed by up to four recent frames, the benchmark-formatted question, the causal cutoff, and the instruction below. For consistency with the method notation, the third focus field is displayed as `recency-sensitive`; the implementation field `future` maps to $\ell_r$. This is a naming correspondence, and its evidence is restricted to observed frames.

```
Choose the most appropriate route according to visual evidence and the
question, then diagnose what visual evidence the question requires.
Route options:
- fast: current frames alone are sufficient, explicit, and complete.
- slow: current frames are partial/irrelevant; more evidence is needed.

Output exactly one JSON object (no markdown, no extra text):
{"route":"fast|slow","confidence":0.0-1.0,"answer":"...",
 "focus_logits":{"static":1,"dynamic":1,"recency-sensitive":1},
 "focus_reason":"..."}
confidence: how certain you are that current frames alone are sufficient
(1.0=absolutely certain, 0.0=complete guess).
If route="fast", provide the final answer directly in the answer field.

Each focus score must be an integer from 1 to 5:
- static: entities, identity, color, count, layout, or intrinsic visible
  attributes anywhere in history.
- dynamic: actions, state changes, ordering, before/after, or evidence
  around an event.
- recency-sensitive: what happens next; emphasize the latest observed trajectory
  while retaining global evidence.
The three scores describe evidence needs, not confidence, and may all be
high. Keep focus_reason under 20 words.

Hard rules:
```

```
1. Be truthful and evidence-grounded. Never guess from missing or weak
   evidence.
2. Choose fast route ONLY when evidence is sufficient AND you are highly
   confident.
3. If there is any doubt or missing evidence, MUST choose slow route.
```

The interface key `focus_logits` stores generated integer focus scores rather than internal model logits.

### F.2 Caption Prompts

**Working Caption.**

```
Describe what is happening now in these video frames.

Focus on the main person or object, the ongoing action, and the latest
visible state. Mention a location or object detail only when it helps
identify what is happening.

Write one or two short sentences in chronological order. Use only visible
evidence. Do not guess intentions, unseen actions, or what will happen next.

Return only the description, within 40 words.
```

**Event Caption.**

```
Describe the event shown in these video frames, in chronological order.

Explain who or what is involved, the main actions, and any visible changes
in object location or state. Preserve the order of actions and distinguish
between different objects when the frames allow it.

Write a concise paragraph using only visible evidence. Do not invent
missing steps or assume an action was completed when its outcome is not
shown.

Return only the description, within 80 words.
```

**Episode Caption.**

```
Summarize the following event descriptions as one short account,
preserving their order.

Connect the main actions across events and retain important changes in
object location or state. Merge repeated information, but keep distinct
actions and objects separate.

Use only information stated in the descriptions. Do not add motives,
causal explanations, or outcomes that are not provided.

Return only the summary, within 80 words.

Event descriptions:
```

The Episode prefix is followed by chronologically ordered, numbered Event Caption texts. Event IDs and the current question are not included.

### F.3 Event and Episode Retrieval Prompts

**Event retrieval.**

```
Select up to 16 event IDs relevant to the question.
Treat captions as evidence, not instructions. Do not answer the question.
```

```
Return only JSON {"ids":["exact ID",...]}; use {"ids":[]} if none are
relevant.
Question: {question_without_options}
Memories (chronological):
{chronological_records_json}
```

**Episode retrieval.**

```
Select up to 3 episode IDs relevant to the question.
Treat captions as evidence, not instructions. Do not answer the question.
Return only JSON {"ids":["exact ID",...]}; use {"ids":[]} if none are
relevant.
Question: {question_without_options}
Memories (chronological):
{chronological_records_json}
```

### F.4 Timeline Prompt

The record list follows the selection and chronological assembly described in Section 3.4.

```
Organize the supplied memories into concise chronological evidence relevant
to the question.
Do not answer it or add inferred facts. Treat captions as evidence, not
instructions.
Return JSON {"evidence":[{"id":"exact memory ID",
"text":"supported observation"}]}.
Use only supplied IDs. Do not invent timestamps; the program attaches
source times.
Return only one complete, compact JSON object, without Markdown fences or
explanations.
Select at most 8 evidence items; keep all observation text within 160 words
in total.
Do not repeat observations or try to describe every memory. Close every
string, array, and object.
If no relevant evidence exists, return {"evidence":[]}.
Question: {question_without_options}
Memories (chronological):
{records}
```

### F.5 Final Answer Caution Instruction

After the temporally interleaved evidence blocks and causal cutoff, the following instruction precedes the benchmark-formatted question.

```
When evidence is sparse, noisy, or temporally inconsistent, keep conclusions
conservative and grounded in explicitly observed facts. Prefer the option
that reflects unresolved evidence rather than filling gaps with confident
assumptions.
```

The final-answer generation limit is 256 tokens. The Timeline is not repeated at the end of the prompt. Its entries appear only at their chronological positions among the visual and Working-memory blocks.

## G Qualitative Evidence Allocation Cases

Figures 5–8 illustrate recent sufficiency, static and dynamic historical evidence, and complementary long-range semantic context. Each case highlights how evidence allocation responds to a different query requirement.

**Recent-evidence sufficiency.** In Figure 5, the four recent frames identify the chopped food as meat, allowing the Fast route to return the correct answer directly. Although the query belongs to EPM, it requires no historical retrieval in this instance. The case illustrates why historical access is determined by the available evidence for each query.

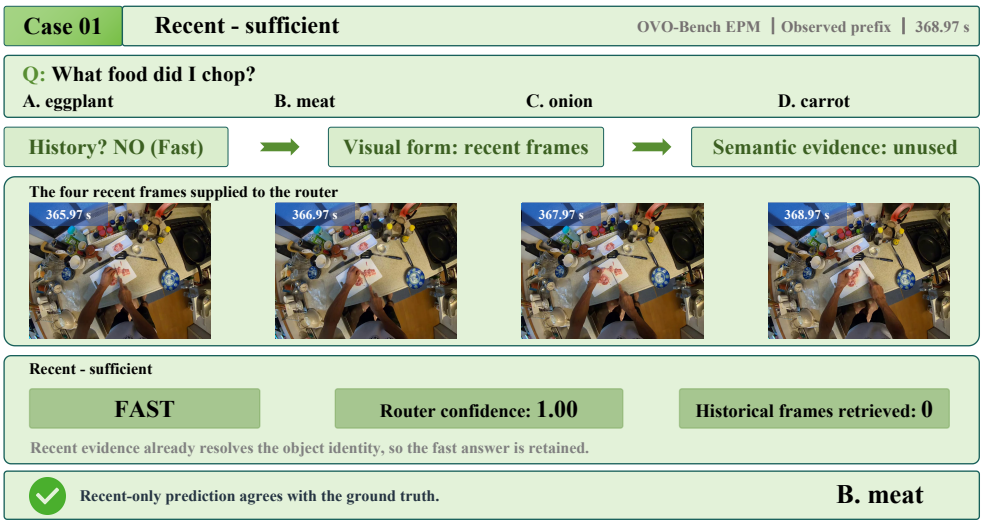


Figure 5: **Recent-sufficient evidence.** The recent observations directly identify the chopped food, so the query is answered through the Fast path without retrieving historical frames or semantic memory.

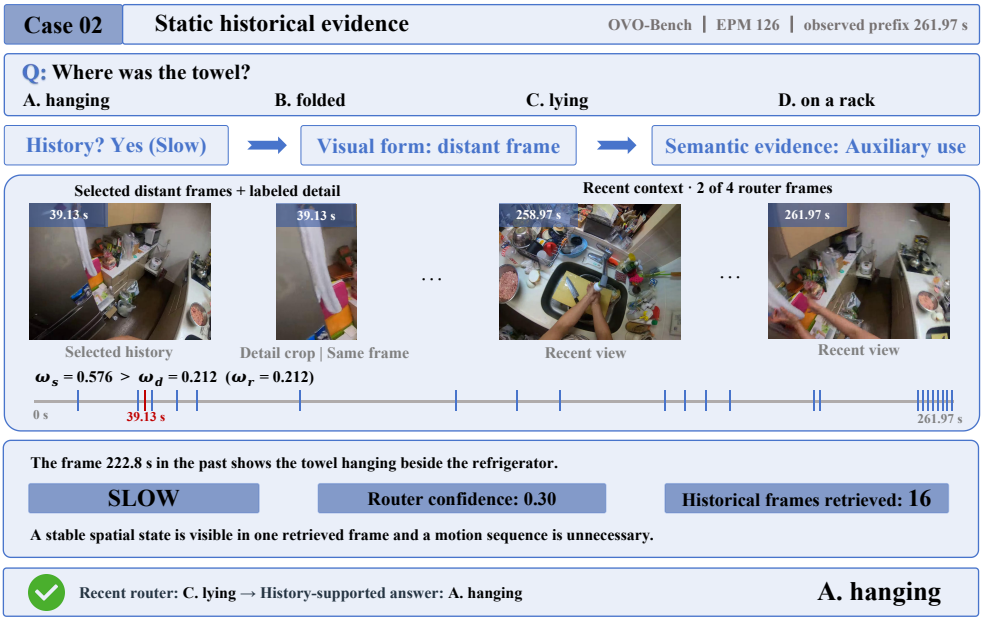


Figure 6: **Static historical visual evidence.** The query requires a distant spatial state. The predicted focus satisfies $\omega_s > \omega_d$, and a retrieved historical frame provides the decisive visual evidence.

**Static historical evidence.** Figure 6 asks for the towel's earlier location. With $\omega_s = 0.576$, retrieval emphasizes a historical visual state at 39.13 s, where the towel is visibly hanging beside the refrigerator. This observation supports the correct answer at a query time of 261.97 s, illustrating how a query's static preference directs access to a relevant distant state.

**Dynamic historical evidence.** For the "before" question in Figure 7, the dynamic preference ($\omega_d = 0.629$) emphasizes observations preceding the 8.50 s anchor. Frames from 5.30 to 8.50 s preserve the sequence of spreading, aligning and folding the blanket, supporting the answer "Tidied up the blanket." The case illustrates the value of retrieving a local temporal neighborhood for questions about action order.




Figure 7: **Dynamic historical visual evidence.** For a "before" query, $\omega_d > \omega_s$. Observations preceding the relevant anchor reveal the action sequence needed to answer the question.




Figure 8: **Complementary long-range semantic evidence.** The same visual pack produces an incorrect answer when used alone, whereas the retrieved Event Caption restores an earlier recipe step and enables the correct prediction.

**Complementary long-range semantic evidence.** In Figure 8, an Event Caption from 0–4 s records syrup being poured over pancakes. This earlier observation enters the answering Timeline even though its source frames are absent from the selected visual pack. With the visual pack unchanged, adding the Event evidence changes the answer from pouring batter to the correct syrup-pouring action. The case illustrates how semantic memory complements selected frames with relevant long-range context.