StreamWeave: CURATING MULTI-TIMESCALE VISUAL EVIDENCE FOR STREAMING VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models have advanced offline video understanding, but streaming video systems must answer questions using only observations available when each question arrives. As the observed prefix grows beyond the model's finite visual context, different questions make different observations relevant. A fixed frame selection cannot reliably preserve all such evidence. Even when past frames are retained, event descriptions written before a question arrives may not reveal which frames contain the visual evidence it seeks. We introduce **StreamWeave**, a training-free agent that curates multi-timescale visual evidence for each question. As video arrives, it builds a searchable visual history linked to original frames. At query time, a recent glance and a broader look-back provide the first-call evidence; the MLLM answers or requests additional source frames through history retrieval. Across three streaming video understanding benchmarks and five frozen MLLMs ranging from an 8B dense model to a 397B mixture-of-experts model, StreamWeave achieves competitive results against published streaming methods. Extensive ablation studies show the effectiveness of its components, temporal layout, and retrieval choices.

1 INTRODUCTION

Multimodal large language models have advanced offline video understanding, from open-ended dialogue and audio-visual reasoning to questions requiring evidence across long recordings (Zhang et al., 2023; Maaz et al., 2024; Wu et al., 2024; Pan et al., 2026). Continuous video assistants (Wang et al., 2026), time-sensitive livestream moderation (Yew et al., 2026), and robots acting on ongoing observations (Driess et al., 2023) need related capabilities while video is still unfolding. Streaming video understanding requires each question to be answered from the prefix observed at its arrival, without relying on future frames (Niu et al., 2025; Lin et al., 2026). As that prefix grows, deciding which observations to place in the MLLM's finite visual context becomes part of the task.

The consequences of this choice are visible in recent streaming results. SimpleStream shows that a small recent window can be a strong baseline, while adding historical context does not necessarily improve real-time perception (Shen et al., 2026). A recent window, however, cannot show every step of a process that has already passed; spreading a fixed number of samples over a longer interval increases temporal reach but reduces the density needed to see short actions (Qu et al., 2025). Figure 1 makes the distinction concrete: *Recent Action* depends on continuity near the question, whereas *Preceding Process* draws on observations distributed over the preceding activity. These are different demands on the same limited input. Repeated views can spend that input without adding distinct evidence, so the visual context must balance temporal reach with detail near the question.

Any bounded selection from the observed video can omit evidence needed by a later question. *Earlier Detail* in Figure 1 illustrates a brief appearance missed by broad sampling, but historical questions are not limited to isolated details: they may ask about several occurrences or relate observations made at different moments. EventMemAgent and OASIS retain visual evidence from the stream, but their semantic search relies on event descriptions produced as the video unfolds (Wen et al., 2026; Liang et al., 2026). A later question may hinge on a visual detail that those descriptions did not

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.

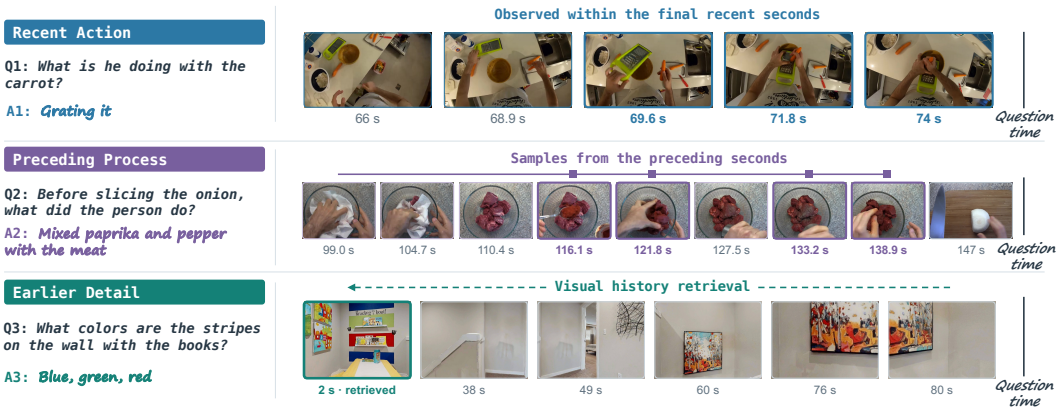


Figure 1: Three OVO-Bench examples from separate videos. A recent action is resolved near the query frontier (top); observations across a longer interval show the preceding process (middle); and a briefly observed detail is recovered from visual history after the question arrives (bottom). All displayed frames lie within the observed prefix.

emphasize. This motivates indexing historical observations by their visual content, so the question can locate relevant records and recover their source frames for reasoning.

These two demands call for organizing visual evidence across time scales as video arrives and when a question is asked. We present **StreamWeave**, a training-free agent that continually builds a coverage-aware visual history: similar observations extend a record's temporal coverage, while distinct observations create source-linked visual records. At query time, StreamWeave assembles a bounded view from a recent glance and a broader look-back, curating redundant glance frames with the MLLM's visual encoder. The MLLM answers from this view or specifies what historical evidence it needs; query-conditioned matching then returns relevant original frames to join the same view for final reasoning. Across five frozen MLLMs, StreamWeave improves OVO-Bench macro accuracy over model-matched 44-frame references by 1.7 to 8.4 points and reaches 86.12 on StreamingBench Real-time.

Our contributions are:

1. We introduce **StreamWeave**, a training-free streaming video agent that organizes multi-timescale visual evidence for each question within a bounded MLLM context.
2. We develop a coverage-aware visual history with searchable visual keys and source-frame links. Question-conditioned retrieval can recover original frames whose relevant details were not described in advance.
3. We evaluate StreamWeave with five frozen MLLMs on three streaming benchmarks, demonstrating consistent OVO-Bench improvements over model-matched 44-frame references and examining the roles of glance curation, look-back coverage, and history retrieval.

2 Related Work

2.1 Streaming Video Understanding

Streaming video understanding evaluates a model on the observed prefix rather than a completed video. OVO-Bench separates backward, real-time, and forward capabilities at explicit causal frontiers (Niu et al., 2025). StreamingBench organizes its tasks into Real-Time Visual Understanding, Omni-Source Understanding, and Contextual Understanding (Lin et al., 2026), while OVBench organizes streaming understanding into perception, memory, verification, and prediction capabilities (Huang et al., 2025). LongVideoBench provides a complementary offline test in which evidence is distributed over long inputs (Wu et al., 2024). Long-video studies also show that sampling density and the temporal locality of a question affect which visual evidence is available to a model (Qu et al., 2025; Lu et al., 2025; Li et al., 2026b; Yuan et al., 2025).

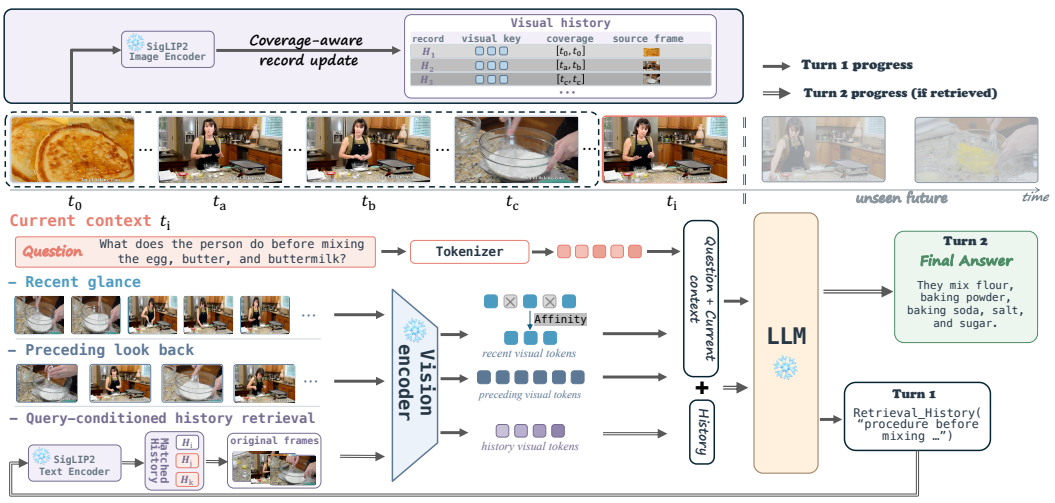


Figure 2: Overview of StreamWeave. Incoming observations update visual history records with a visual key, temporal coverage, and a source-frame pointer. At a query frontier, the glance and look-back form the working context. A history-retrieval request returns original frames that join the same question and working context for final reasoning; future frames remain unavailable.

StreamForest learns persistent event memory alongside fine-grained current perception, and StreamOPD studies streaming-specific post-training (Zeng et al., 2026; Wu et al., 2026b). Training-free work retains the pretrained MLLM while changing which observations reach it: SimpleStream demonstrates the strength of a recent window, whereas SAVEMem, FOLIO, SelectStream, and HERMES explore selective or hierarchical retention (Yuan et al., 2026; Shen et al., 2026; Wu et al., 2026a; Fan et al., 2026; Ge et al., 2026; Zhang et al., 2026). Together, these results show why temporal reach and current-frame detail must be considered jointly.

2.2 Streaming Video Agents

Streaming video agents add decisions about observation, memory, retrieval, or response timing to streaming understanding. StreamBridge couples a memory buffer with a separate activation model for proactive response, while StreamAgent plans when and where to observe future task-relevant evidence (Wang et al., 2026; Yang et al., 2025). StreamRAG introduces retrieval augmentation for real-time video understanding (Xie et al., 2026a). EventMemAgent organizes short- and long-term observations around events and equips the model with adaptive perception tools (Wen et al., 2026). OASIS combines short- and longer-horizon frame buffers with hierarchical event memory for on-demand reasoning (Liang et al., 2026); Visual Agentic Memory and GROVE also develop agentic access to stored video evidence (Li et al., 2026a; Gong et al., 2026). StreamWeave centers its model-visible context on source frames: the first call includes two temporal views, while the historical path adds original frames from visually keyed records when requested.

3 Method

3.1 Problem Setup

Let $V = (x_t)_{t\geq 0}$ be a video stream and let question q_i arrive at frontier τ_i. The agent can use only the observed prefix $V_{\leq \tau_i} = \{x_t \mid t \leq \tau_i\}$. StreamWeave maintains a visual archive $\mathcal{A}_i$ over this prefix and gives a frozen MLLM F_θ a working visual context W_i with the question. The MLLM may answer from that context or request source frames H_i from the archive for a final call. At most B images enter a model call, although the archive can grow with the stream. Figure 2 shows how the current context and historical evidence reach the same MLLM without exposing future frames.

3.2 COVERAGE-AWARE VISUAL HISTORY

As frames arrive, StreamWeave builds a visual archive without assigning event descriptions in advance. A retained record is

$$m_j = (h_j, \bar{x}_j, [a_j, b_j], f_j, p_j), \tag{1}$$

where h_j identifies the record, $\bar{x}_j$ is its representative observation, and $[a_j, b_j]$ records its temporal coverage. The visual key f_j is a frozen SigLIP2 image embedding of that observation (Tschannen et al., 2025); p_j points to the original source frame. The key makes the record searchable, while the pointer allows the MLLM to inspect the corresponding pixels after retrieval.

Coverage aggregation updates this archive as frames arrive. For each incoming candidate x_t, a low-resolution appearance gate compares it with the latest retained representative. On a match, the record's coverage becomes $[a_j, t]$ while its visual key and source-frame pointer stay fixed; otherwise a new record is appended with coverage $[t, t]$. Scene changes and query-time anchors are retained. Thus repeated observations share a searchable record, while distinct observations preserve separate source-frame references. Coverage is record metadata: retrieval returns one original source frame per selected record, not every frame in its interval. The gate's boundary rule is given in Appendix A. At frontier τ_i, the eligible archive $\mathcal{A}_i$ contains only information from the observed prefix.

3.3 WEAVING MULTI-TIMESCALE VISUAL EVIDENCE

When question q_i arrives at frontier τ_i, the MLLM first receives two views of the observed stream. A recent glance retains the action near the question, while a sparser look-back shows how the scene reached its current state. Let Δ_g and Δ_ℓ be their durations, and let n_g and n_ℓ be their frame budgets. With $\mathcal{U}(I, k)$ denoting up to k uniformly selected source frames from interval I, we form

$$R_i = \mathcal{U}([\tau_i - \Delta_g, \tau_i], n_g), \qquad L_i = \mathcal{U}([\tau_i - \Delta_g - \Delta_\ell, \tau_i - \Delta_g), n_\ell). \tag{2}$$

Here R_i contains glance candidates and L_i contains look-back frames. Both are available in the first model call. The look-back remains unchanged; the glance is curated to avoid spending visual input on nearly identical adjacent observations.

The MLLM's visual encoder represents each glance candidate r_j as a pooled feature v_j. We score an interior frame by its similarity to both neighbors in the original glance:

$$\begin{aligned} v_j &= \mathrm{L2Norm}\big(\mathrm{MeanPool}(E_m^{\mathrm{vis}}(r_j))\big)\,, \\ \rho_j &= \min(v_{j-1}^\top v_j,\ v_j^\top v_{j+1}). \end{aligned} \tag{3}$$

We protect the first and final three glance positions, then greedily remove nonadjacent interior frames with $\rho_j \geq \gamma$ in descending score order, retaining at least $\min(|R_i|, 9)$ glance frames. The retained sequence $R_i^\star$ and look-back frames L_i form $W_i = \mathrm{Pack}(L_i, R_i^\star)$, preserving source timestamps and temporal order.

The frozen MLLM first reasons over (q_i, W_i). It can answer from this context or call the history-retrieval tool to request missing evidence. For questions explicitly about earlier events, the model may reconsider a direct response once before it is finalized. A history request u_i expresses the information sought as text z_i and may specify temporal, evidence-type, scope, and frame-budget constraints. The SigLIP2 text encoder maps z_i into the embedding space of the visual keys f_j stored in Section 3.2:

$$g_i = \mathrm{L2Norm}(E_{\mathrm{S}}^{\mathrm{text}}(z_i)), \qquad s_{ij} = g_i^\top f_j, \quad m_j \in \mathcal{A}_i. \tag{4}$$

The tool ranks eligible records by s_{ij}, applies the request constraints, and excludes source frames already present in W_i. Several records can be selected from different parts of the observed prefix. Each contributes one original source frame, and the returned frames are ordered by timestamp:

$$\begin{aligned} J_i &= \mathrm{SelectRecords}\big(u_i, \{(m_j, s_{ij}) : m_j \in \mathcal{A}_i\}\big), \qquad |J_i| \leq k_i, \\ H_i &= \mathrm{SortTime}\big((\mathrm{Load}(p_j))_{j \in J_i}\big). \end{aligned} \tag{5}$$

When history is retrieved, the same MLLM receives the original question and the complete first-call context together with the retrieval trace and recalled source frames. It produces the final answer from $C_i = \mathrm{Pack}(W_i, H_i)$, where $n_{\mathrm{img}}(C_i) \leq B$. The Pack operation preserves group labels, timestamps,

and within-group order; recalled frames follow the first-call context. History can therefore contribute evidence from anywhere in the observed prefix without becoming another fixed sampling window.

The complementary roles of these views can be expressed as an evidence decomposition. Let Y denote the answer, Q the question, and $R^\star$, L, and H the corresponding visual evidence over queries. The chain rule gives

$$\begin{aligned} I(Y; R^\star, L, H \mid Q) = {} & I(Y; R^\star \mid Q) \\ & + I(Y; L \mid Q, R^\star) \\ & + I(Y; H \mid Q, R^\star, L). \end{aligned} \tag{6}$$

The terms distinguish evidence in the glance, additional evidence from the look-back, and evidence recalled through history retrieval beyond the first-call view.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate on three streaming benchmarks. OVO-Bench has 3,035 questions in Backward, Real-time, and Forward tracks; its overall score is the macro average of the three track accuracies (Niu et al., 2025). We use all 2,500 questions in the StreamingBench Real-time split and report both its official accuracy and ten task scores (Lin et al., 2026). OVBench comprises six task families and 16 subtasks (Huang et al., 2025). We show the six families in the main table and the complete subtask results in Appendix E; the official overall score averages the 16 subtask accuracies, not the six family means. A separate LongVideoBench transfer study appears in Appendix C (Wu et al., 2024).

Models and protocol. We use five frozen MLLMs spanning dense and mixture-of-experts architectures: Qwen3-VL-8B-Instruct, Qwen3.5-9B, Qwen3-VL-32B-Instruct, Qwen3.6-27B, and the FP8 checkpoint of Qwen3.5-397B-A17B. We set the glance duration Δ_g to 8 s with $n_g = 12$ candidates and the look-back duration Δ_ℓ to 40 s with $n_\ell = 8$ frames. The MLLM's visual encoder curates the glance candidates; SigLIP2 provides the keys for historical search (Tschannen et al., 2025). The coverage-aggregation threshold is 0.85. A retrieval can return up to 24 original frames, giving at most 44 images in the final visual context. The same configuration is used across the five MLLMs and three benchmarks without benchmark-specific tuning, allowing the same evidence-curation policy to be evaluated across model scales and architectures.

For evaluation throughput, we cache query-independent visual features and history updates, then restrict the archive to each question's observed prefix. The MLLM always receives original frames rather than index features. Generation uses temperature 0 and at most 256 output tokens; additional protocol details are in Appendix A.

For each MLLM, the base uniformly samples 44 frames from the observed prefix. Tables 1–3 retain the published methods' frame protocols and use dashes for unreported task scores. Paired OVO-Bench intervals use 10,000 source-video cluster-bootstrap replicates.

4.2 OVO-BENCH

OVO-Bench exposes different evidence demands across Backward, Real-time, and Forward questions. Table 1 compares each MLLM with its model-matched 44-frame base and with published streaming systems. Across all five MLLMs, StreamWeave raises the three-track macro average by 1.7 to 8.4 points; every model improves on both Backward and Real-time.

Table 1: OVO-Bench accuracy (%). #Frames follows each source's protocol; counts denote per-query visual context and fps denotes stream sampling. Colored subscripts report percentage-point changes from the corresponding 44-frame reference; bold and underlined scores mark the best and second best.

Method	Params	#Frames	Backward	Real-time	Forward	Avg.
Training-based methods						
StreamForest (Zeng et al., 2026)	7B	1 fps	52.00	61.20	53.50	55.60
EventMemAgent (Wen et al., 2026)	8B	$\leq$ 32	58.03	68.29	55.92	60.75
SelectStream (Ge et al., 2026)	8B	1 fps	62.20	**82.76**	56.13	67.03
Training-free methods						
SimpleStream (Shen et al., 2026)	8B	4	54.00	81.40	–	–
HERMES (Zhang et al., 2026)	7B	1 fps	49.40	69.00	–	–
OASIS (Liang et al., 2026)	8B	–	57.21	78.14	–	–
FOLIO (Fan et al., 2026)	8B	–	69.10	<u>82.00</u>	–	–
GROVE (Gong et al., 2026)	Mixed	0.5 fps	<u>69.90</u>	76.20	56.50	67.50
FluxMem (Xie et al., 2026b)	7B	1 fps	–	67.20	–	53.30
SAVEMem (Wu et al., 2026a)	7B	1 fps	50.44	74.93	–	–
Qwen3-VL-8B	8B	44	51.94	68.25	54.50	58.23
+StreamWeave	8B	$\leq$ 44	58.89 (+7.0)	74.40 (+6.1)	53.69 (-0.8)	62.33 (+4.1)
Qwen3.5-9B	9B	44	51.37	71.66	54.34	59.12
+StreamWeave	9B	$\leq$ 44	60.33 (+9.0)	76.43 (+4.8)	55.83 (+1.5)	64.19 (+5.1)
Qwen3-VL-32B	32B	44	60.77	72.28	54.25	62.43
+StreamWeave	32B	$\leq$ 44	63.00 (+2.2)	77.46 (+5.2)	51.80 (-2.4)	64.09 (+1.7)
Qwen3.6-27B	27B	44	58.22	74.81	49.77	60.94
+StreamWeave	27B	$\leq$ 44	65.64 (+7.4)	80.34 (+5.5)	**62.00** (+12.2)	<u>69.32</u> (+8.4)
Qwen3.5-397B-A17B	397B-A17B-FP8	44	60.96	73.70	<u>59.90</u>	64.85
+StreamWeave	397B-A17B-FP8	$\leq$ 44	**72.00** (+11.0)	81.10 (+7.4)	57.11 (-2.8)	**70.07** (+5.2)

Forward changes vary by model: Qwen3.6-27B gains 12.2 points and Qwen3.5-9B gains 1.5 points, while the other three decrease by 0.8 to 2.8 points. Qwen3.5-397B-A17B achieves the highest macro score, 70.07, and Qwen3.6-27B achieves the largest paired gain, reaching 69.32. The joint Backward and Real-time gains raise the complete three-track average across all models, while Forward remains the main source of variation. GROVE's Mixed label denotes its Qwen3.5-35B-A3B perception model, GPT-4.1-mini memory consolidation, and GPT-5-mini reasoning (Gong et al., 2026). FluxMem reports an overall score for all three OVO-Bench tracks, whereas SAVEMem reports only Backward and Real-time; its two-track average is not entered in the three-track Avg. column. Appendix D reports the 12-task breakdown.

4.3 StreamingBench Real-time

StreamingBench Real-time tests whether broader temporal reach comes at the cost of perception at the query frontier. With the same configuration, all five MLLMs improve over their model-matched bases in Table 2. Qwen3.5-397B-A17B reaches 86.12, and Qwen3.6-27B reaches 85.60. The paired gains range from 0.9 points for the 27B model to 5.0 points for the 397B model; the 8B model also gains 3.8 points. The gains therefore do not simply increase with model size.

Clips Summarize and Spatial Understanding improve in every paired comparison, whereas Counting decreases in every one. For Qwen3-VL-8B, Clips Summarize rises from 80.8 to 92.7 and Counting falls from 57.0 to 47.1; the same contrast appears in the 397B model, where these scores change from 89.3 to 93.7 and from 53.9 to 48.7, respectively. The task breakdown therefore locates both the gains and the remaining weakness within real-time perception. Qwen3.8-Omni-Flash reports an overall score of 80.80 without a public ten-task breakdown (Qwen Team, 2026).

4.4 OVBench

Real-time gains alone do not show whether the same visual context helps with other temporal demands.

Table 2: StreamingBench Real-time accuracy (%). Avg. is the official score over 2,500 questions; task scores have one decimal. Colored Avg. subscripts compare with each 44-frame reference. Bold and underline mark the highest and second-highest reported scores in each column; ties share a mark. Task names appear in Appendix A.

Method	Params	#Frames	Real-time tasks										Avg.
			OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	
Training-based methods													
StreamForest	7B	1 fps	83.1	82.8	82.7	84.3	77.5	78.2	76.8	69.1	75.6	54.4	77.26
EventMemAgent	8B	$\leq$ 32	83.9	76.6	87.7	85.3	77.0	79.4	77.8	68.3	72.2	48.7	77.00
StreamOPD	4B	4	–	–	–	–	–	–	–	–	–	–	84.55
SelectStream	8B	1 fps	86.3	87.4	86.0	87.1	**84.7**	84.6	85.8	75.0	78.2	**71.6**	82.67
Training-free methods													
SimpleStream	8B	4	–	–	–	–	–	–	–	–	–	–	80.59
HERMES	7B	1 fps	83.7	81.3	88.0	87.5	76.7	86.6	82.4	76.0	73.9	46.6	79.44
OASIS	8B	–	–	–	–	–	–	–	–	–	–	–	78.22
FOLIO	8B	–	–	–	–	–	–	–	–	–	–	–	81.20
GROVE	Mixed	0.5 fps	–	–	–	–	–	–	–	–	–	–	77.00
FluxMem	7B	1 fps	80.2	81.1	81.4	85.3	78.0	83.8	80.6	65.9	69.6	52.1	76.40
SAVEMem	7B	1 fps	–	–	–	–	–	–	–	–	–	–	76.00
Off-the-shelf MLLMs													
Qwen3.8-Omni-Flash	–	–	–	–	–	–	–	–	–	–	–	–	80.80
Qwen3-VL	8B	44	81.2	71.1	80.8	84.0	74.5	81.0	81.5	65.5	73.9	57.0	76.08
+StreamWeave	8B	$\leq$ 44	84.5	74.2	92.7	86.9	75.8	84.4	80.6	72.0	80.2	47.1	79.84 (+3.8)
Qwen3.5	9B	44	84.7	88.3	88.6	89.2	77.0	86.6	80.6	76.0	80.5	57.5	81.96
+StreamWeave	9B	$\leq$ 44	88.8	86.7	91.2	90.8	77.6	86.9	83.3	**82.1**	81.9	47.1	83.20 (+1.2)
Qwen3-VL	32B	44	82.3	88.3	87.1	85.6	78.3	84.1	83.3	69.9	76.8	54.9	79.52
+StreamWeave	32B	$\leq$ 44	85.6	83.6	93.1	87.3	80.8	86.3	85.2	76.8	83.6	49.2	82.44 (+2.9)
Qwen3.6	27B	44	87.5	**92.2**	89.6	91.5	82.0	89.4	84.3	76.0	82.7	64.8	84.68
+StreamWeave	27B	$\leq$ 44	91.0	88.3	92.7	**92.8**	82.6	89.4	84.3	81.7	**85.8**	51.8	85.60 (+0.9)
Qwen3.5	**397B-A17B-FP8**	44	87.2	66.4	89.3	87.3	80.8	85.4	84.3	74.4	82.4	53.9	81.12
+StreamWeave	**397B-A17B-FP8**	$\leq$ 44	**92.4**	88.3	**93.7**	**92.8**	84.5	**91.3**	**86.1**	**82.1**	85.5	48.7	**86.12** (+5.0)

OVBench provides a complementary test through 16 subtasks organized into six families. Table 3 reports the family means and the official 16-subtask average.

The unchanged configuration improves four of five MLLMs in Table 3; Qwen3-VL-8B changes by -0.04 points. Qwen3.6-27B reaches the highest overall score, 62.51, and Qwen3.5-9B has the largest paired increase, 3.05 points. For Qwen3.5-9B, Spatial Perception and Temporal Hallucination Verification gain 8.4 and 7.0 points, respectively, while Past Memory drops by 6.6 points. FP, THV, PM, SP, STP, and TP denote Future Prediction, Temporal Hallucination Verification, Past Memory, Spatial Perception, Spatio-Temporal Perception, and Temporal Perception. The full 16-subtask comparison is in Appendix E.

Across the three benchmarks, the same configuration improves all five model-matched bases on OVO-Bench and StreamingBench Real-time and four on OVBench. The gains extend from the 8B model to the 397B mixture-of-experts model, showing that the evidence-curation strategy remains effective across the evaluated model scales and architectures.

4.5 Ablation Studies

Component contributions. We use Qwen3.5-9B on complete OVO-Bench. Table 4 starts from the full method and removes one component at a time. Without the look-back, the macro score falls from 64.19 to 60.59, including an 8.36-point drop on Forward questions. Removing historical retrieval lowers the macro score to 63.18 and Forward accuracy by 4.09 points; removing glance curation lowers the macro score to 63.26. Coverage aggregation raises the macro score from 63.29

Table 3: OVBench accuracy (%). The six task-family columns average their constituent subtasks; Avg. is the official mean of all 16 subtask scores. #Frames follows each source's protocol. Colored Avg. subscripts compare with each 44-frame reference; bold and underline mark the highest and second-highest reported scores in each column.

Method	Params	#Frames	FP	THV	PM	SP	STP	TP	Avg.
Training-based methods									
VideoChat-Online (Huang et al., 2025)	4B	2 fps	46.8	61.4	55.7	39.5	68.6	56.9	54.90
StreamForest (Zeng et al., 2026)	7B	1 fps	–	–	–	–	–	–	60.50
Off-the-shelf MLLMs									
Gemini-1.5-Flash (Team et al., 2024)	–	≤ 64	49.0	52.6	49.0	34.9	68.5	50.8	50.70
LLaVA-OV-7B (Li et al., 2024a)	7B	≤ 64	55.5	51.7	49.4	28.6	68.1	43.0	49.50
Qwen2-VL-7B (Wang et al., 2024)	7B	≤ 64	49.5	52.5	57.2	31.9	<u>73.2</u>	35.9	49.70
Qwen3-VL-8B	8B	44	50.2	60.8	53.2	42.1	71.5	54.3	55.14
+StreamWeave	8B	≤ 44	54.1	59.9	52.3	40.6	70.4	53.6	55.10 (-0.04)
Qwen3.5-9B	9B	44	51.3	55.6	60.3	42.1	64.3	56.2	55.18
+StreamWeave	9B	≤ 44	56.4	62.6	53.7	**50.5**	67.4	59.3	58.23 (+3.05)
Qwen3-VL-32B	32B	44	49.8	59.5	60.2	46.9	66.9	55.7	56.45
+StreamWeave	32B	≤ 44	55.4	62.4	56.4	46.8	**76.0**	54.2	58.18 (+1.73)
Qwen3.6-27B	27B	44	51.8	**68.2**	**66.0**	40.4	72.8	62.7	60.80
+StreamWeave	27B	≤ 44	**60.7**	<u>67.4</u>	59.4	49.1	71.8	<u>65.2</u>	**62.51** (+1.70)
Qwen3.5-397B-A17B	397B-A17B-FP8	44	54.2	62.4	<u>61.0</u>	37.3	71.7	63.5	58.84
+StreamWeave	397B-A17B-FP8	≤ 44	<u>59.4</u>	64.4	<u>61.0</u>	<u>49.3</u>	70.1	**65.6**	<u>61.86</u> (+3.02)

Table 4: Component ablations on complete OVO-Bench (Qwen3.5-9B). Base uses 44 uniformly sampled frames. Each w/o row removes one component; Δ Avg. is relative to the full method.

Setting	Backward	Real-time	Forward	Avg.	Δ Avg.
Base	51.37	71.66	54.34	59.12	-5.07
StreamWeave (full)	60.33	76.43	55.83	64.19	–
w/o look-back	56.59	77.70	47.47	60.59	-3.60
w/o glance curation	60.91	75.90	52.98	63.26	-0.93
w/o historical retrieval	59.81	77.98	51.74	63.18	-1.01
w/o coverage aggregation	60.96	75.79	53.12	63.29	-0.90

to 64.19 while reducing returned historical frames per question from 6.76 to 4.13. The separate glance-curation comparison is in Appendix B.

Table 5: Temporal coverage on complete OVO-Bench (Qwen3.5-9B, %). Glance and look-back give frame counts; – means none. Δ Avg. is relative to each block's reference row.

Glance	Look back	Backward	Real-time	Forward	Avg.	Δ Avg.
Glance size						
4	–	55.87	76.26	50.39	60.84	-1.17
8	–	56.25	77.30	49.68	61.08	-0.93
12	–	57.94	77.30	50.79	**62.01**	–
16	–	57.55	77.88	50.00	61.81	-0.20
Look-back coverage						
12	–	57.94	77.30	50.79	62.01	–
12	4	60.08	76.60	52.35	63.01	+1.00
12	8	60.91	76.49	56.01	**64.47**	+2.46
12	12	61.68	76.39	54.90	64.32	+2.31
Matched current-view size						
20	–	59.67	77.82	50.45	62.65	–
12	8	60.12	77.82	55.13	**64.36**	+1.71

Temporal placement. Table 5 isolates placement from glance curation. Adding eight look-back frames to a 12-frame glance raises Avg. from 62.01 to 64.47. With 20 frames in both views, the mixed layout still beats glance alone (64.36 versus 62.65; video-cluster interval: $[0.31, 3.15]$). Similar model calls and image counts tie the gain to placement.

Table 6: History-cap accuracy (%) on complete OVO-Bench with Qwen3.5-9B. Δ Avg. uses six frames as reference.

History cap	Backward	Real-time	Forward	Avg.	Δ Avg.
6 frames	61.51	76.52	52.74	63.59	–
12 frames	61.31	76.68	55.20	64.40	+0.81
18 frames	61.43	76.62	55.21	64.42	+0.83
24 frames	61.31	76.49	56.00	**64.60**	+1.01

History capacity. Table 6 varies the retrieved history cap under SigLIP2 glance curation. Raising it from six to 24 frames improves Avg. from 63.59 to 64.60, with the largest change on Forward questions.

Table 7: History access on complete OVO-Bench with Qwen3.5-9B. Scores are accuracy (%); Calls and Images are per-question means.

Setting	Accuracy (%)				Per-question usage	
	Backward	Real-time	Forward	Avg.	Calls	Images
History availability and tool use						
No history	60.03	77.39	51.72	63.05	1.00	19.43
No evidence reconsideration	61.61	76.16	48.04	61.94	1.06	21.05
Adaptive retrieval	62.12	76.42	55.83	64.79	1.67	39.06
Always retrieve	62.08	78.35	55.51	**65.31**	2.15	53.50
Historical frame selection with fixed tool decisions						
Uniform sampling	61.29	76.22	52.94	63.48	1.67	39.06
Visual-key matching	62.53	76.41	55.86	**64.93**	1.67	39.06

Retrieval and source-frame selection. Table 7 shows that adaptive retrieval raises Avg. from 63.05 without history to 64.79. Without evidence reconsideration, history calls fall from 36.4% to 5.5% of questions and Avg. to 61.94. With tool decisions and image counts fixed, visual-key matching outperforms uniform sampling (64.93 versus 63.48; paired video-cluster interval: $[0.69, 2.12]$).

Retrieval cost. Always retrieving raises Avg. from 64.79 to 65.31 but increases model calls from 1.67 to 2.15 and images from 39.06 to 53.50 per question, counting repeated first-call context. The 0.52-point gap has a video-cluster interval spanning zero; adaptive retrieval uses less inference without a conclusive accuracy loss.

5 Limitations and Conclusion

StreamWeave improves the evidence supplied to a frozen MLLM but cannot overcome its perceptual or reasoning limits; joint training may raise this ceiling. It also permits only one retrieval round. A ReAct-style agent could refine retrieval after inspecting returned frames, with added latency to manage.

Overall, StreamWeave curates multi-timescale visual evidence through a glance, look-back, and agent-gated history retrieval. Across five frozen MLLMs, it reaches 70.07 on OVO-Bench, 86.12 on StreamingBench Real-time, and 62.51 on OVBench. Ablations support the complementary roles of current-view curation and historical recall.

AI USE STATEMENT

In this work, we independently formulated the research question, developed the method, designed and conducted the experiments, analyzed the results, and selected and verified the references. We used generative AI tools only for language polishing and LaTeX formatting. We reviewed all AI-assisted edits for technical accuracy and consistency with the experimental results and cited sources. We take full responsibility for the final content of this work, including any text or artifacts prepared with AI assistance.

REFERENCES

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Haoyang Fan, Dhruv Parikh, Anvitha Ramachandran, Sameh Gobriel, Nilesh Jain, Rajgopal Kannan, and Viktor Prasanna. Folio: Focused semantic memory for streaming video understanding. *arXiv preprint arXiv:2607.13298*, 2026.

Haonan Ge, Yiwei Wang, Hang Wu, and Yujun Cai. What should a streaming video model remember? *arXiv preprint arXiv:2606.16353*, 2026.

Sitong Gong, Caixin Kang, Tianyu Yan, Guo Chen, Bo Zheng, Kaipeng Zhang, Yunzhi Zhuge, Xiang Ruan, Huchuan Lu, and Yifei Huang. Grove: Growing and reasoning over temporally stratified memory from streaming video experience. *arXiv preprint arXiv:2608.02392*, 2026.

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

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

Aiden Yiliu Li, Nels Numan, and Anthony Steed. Visual agentic memory: Enabling online long video understanding via online indexing, hierarchical memory, and agentic retrieval. *arXiv preprint arXiv:2605.16481*, 2026a.

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

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, et al. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024b.

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, 2026b.

Zhijia Liang, Jiaming Li, Weikai Chen, Yanhao Zhang, Haonan Lu, and Guanbin Li. Oasis: On-demand hierarchical event memory for streaming video reasoning. *arXiv preprint arXiv:2604.17052*, 2026.

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

Yujie Lu, Yale Song, William Wang, Lorenzo Torresani, and Tushar Nagarajan. Vited: Video temporal evidence distillation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8501–8511. IEEE, 2025.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

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

Hewen Pan, Cong Wei, Dashuang Liang, Zepeng Huang, Pengfei Gao, Ziqi Zhou, Lulu Xue, Pengfei Yan, Xiaoming Wei, Minghui Li, et al. Ufvideo: Towards unified fine-grained video cooperative understanding with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31833–31844, 2026.

Tianyuan Qu, Longxiang Tang, Bohao Peng, Senqiao Yang, Bei Yu, and Jiaya Jia. Does your vision-language model get lost in the long video sampling dilemma? In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20889–20899. IEEE, 2025.

Qwen Team. Qwen3.8-omni-flash: Omni senses. agentic delivery., September 2026. URL `https://qwen.ai/blog?id=qwen3.8-omni-flash`.

Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. A simple baseline for streaming video understanding. *arXiv preprint arXiv:2604.02317*, 2026.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

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

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.

Siwei Wen, Zhangcheng Wang, Xingjian Zhang, Lei Huang, and Wenjun Wu. Eventmemagent: Hierarchical event-centric memory for online video understanding with adaptive tool use. *arXiv preprint arXiv:2602.15329*, 2026.

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

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

Keming Wu, Baoyi Wang, Kaichen Zhang, Xiang An, Zuhao Yang, Sudong Wang, Haowei Zhu, Tingxuan Huang, Hongcheng Gao, and Bin Wang. Streamopd: A post-training recipe with spatio-temporal cue gating for streaming video understanding. *arXiv preprint arXiv:2608.16320*, 2026b.

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.

Yiweng Xie, Bo He, Junke Wang, Xiangyu Zheng, Ziyi Ye, and Zuxuan Wu. Fluxmem: Adaptive hierarchical memory for streaming video understanding. *arXiv preprint arXiv:2603.02096*, 2026b.

Haolin Yang, Feilong Tang, Lingxiao Zhao, Xinlin Zhuang, Yifan Lu, Xiang An, Ming Hu, Xiaofeng Zhang, Abdalla Swikir, Junjun He, et al. Streamagent: Towards anticipatory agents for streaming video understanding. *arXiv preprint arXiv:2508.01875*, 2025.

Wei Chee Yew, Hailun Xu, Sanjay Saha, Xiaotian Fan, Hiok Hian Ong, David Yuchen Wang, Kanchan Sarkar, Zhenheng Yang, and Danhui Guan. Dynamic content moderation in livestreams: Combining supervised classification with mllm-boosted similarity matching. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2528–2538, 2026.

Chao Yuan, Yang Yang, Yehui Yang, and Zach Cheng. Date: Dynamic absolute time enhancement for long video understanding. *arXiv preprint arXiv:2509.09263*, 2025.

Chao Yuan, Shimin Chen, Minliang Lin, Limeng Qiao, Guanglu Wan, and Lin Ma. Unicomp: Rethinking video compression through informational uniqueness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18609–18618, 2026.

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

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pp. 543–553, 2023.

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, 2026.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

A Protocol Details

The state machine permits an initial model request, at most one evidence-obligation retry, at most one schema repair, one real retrieval, and a final answer request. The MLLM supplies the retrieval decision and arguments; the validator only requests a well-formed reconsideration and never selects evidence itself. Valid references must exist in the frozen causal index and stay at or before the query frontier. The final set contains at most 20 current candidates before semantic curation and 24 historical images. Missing source videos, protocol failures, empty retrievals, timeouts, and unparseable answers remain in the official denominator.

History updates and requests. For an incoming candidate x_t, the duplicate gate compares its low-resolution appearance with the latest retained representative. Let d_Y and r_Y be the exposure-compensated luminance deviation and changed-pixel ratio, with d_C and r_C their chrominance counterparts. A candidate extends the current record when

$$|\Delta Y| < \epsilon_Y, \quad d_Y < \epsilon_{dY}, \quad r_Y < \epsilon_{rY}, \quad d_C < \epsilon_{dC}, \quad r_C < \epsilon_{rC}. \tag{7}$$

A compact spatial appearance descriptor resolves a narrow boundary region; scene changes and query-time anchors are retained. A history-tool request is $u_i = (z_i, \delta_i, \eta_i, \sigma_i, k_i)$, where z_i names the missing visual information, δ_i its temporal relation to known observations, η_i the evidence type, σ_i the search scope, and k_i the frame budget.

Features from the MLLM visual encoder provide recent affinity on the three streaming benchmarks; the LongVideoBench transfer rows use SigLIP2 recent affinity. Recent features are mean-pooled from the final unmerged visual tokens in FP32 before normalization. SigLIP2 embeddings provide history localization in both cases. These features are cached independently of the question. The history cache is filtered by the query frontier before retrieval, and the recent cache is used only for the prescribed recent candidates; it never expands the history candidate set. Frames are sent as JPEG images at quality 90 and at most 151,200 pixels per image. Generation uses temperature 0, seed 0, disabled thinking, and at most 256 output tokens. LongVideoBench follows the official multiple-choice parser with a deterministic query-seeded fallback; malformed outputs remain in the denominator.

All paired StreamingBench references and StreamWeave runs use the same single-letter answer constraint and the official first-character evaluator; all 2,500 questions remain in the denominator.

StreamingBench task labels. OP, CR, CS, ATP, EU, TR, PR, SU, ACP, and CT denote Object Recognition, Causal Reasoning, Clips Summarize, Attribute Recognition, Event Understanding, Text-Rich Understanding, Prospective Reasoning, Spatial Understanding, Action Recognition, and Counting, respectively.

B Glance Curation

Table 8 compares frame-selection rules within the same 12-frame glance and eight-frame look-back layout on complete OVO-Bench. The model-matched ViT used in the main method improves Avg. from 63.26 without curation to 64.19, a 0.93-point gain. Its score remains close to SigLIP2 at 64.60, while retaining 16.68 current frames per question on average instead of 19.43. This choice lets the MLLM's visual encoder curate the frames it subsequently reasons over.

Table 8: Glance-curation accuracy (%) on complete OVO-Bench with Qwen3.5-9B. Δ Avg. is measured against no curation; bold and underline mark the highest and second-highest accuracy in each column.

Curation	Backward	Real-time	Forward	Avg.	Δ Avg.
None	<u>60.91</u>	75.90	52.98	63.26	–
Pixel similarity	60.10	75.14	52.70	62.65	-0.61
SigLIP2	**61.31**	**76.49**	**56.00**	**64.60**	+1.34
Model-matched ViT	60.33	<u>76.43</u>	<u>55.83</u>	<u>64.19</u>	+0.93

Table 9: LongVideoBench visual-only validation accuracy (%) (Wu et al., 2024). The StreamWeave rows use the separate SigLIP2 recent-curation variant; bold and underlined scores mark the best and second best among the reported methods.

Method	Params	#Frames	Accuracy
Training-based methods			
StreamOPD (Wu et al., 2026b)	4B	≤ 32	61.41
Off-the-shelf MLLMs			
GPT-4o (0513) (Hurst et al., 2024)	–	256	**66.70**
Aria (Li et al., 2024b)	25B-A3.5B	256	<u>64.20</u>
Gemini-1.5-Pro (0514) (Team et al., 2024)	–	256	64.00
LLaVA-Video-72B-Qwen2 (Zhang et al., 2024)	72B	128	63.90
Qwen2-VL-7B (Wang et al., 2024)	7B	256	55.60
StreamWeave (*Qwen3-VL-8B*)	8B	≤ 44	45.25
StreamWeave (*Qwen3.5-9B*)	9B	≤ 44	57.29
StreamWeave (*Qwen3-VL-32B*)	32B	≤ 44	55.35
StreamWeave (*Qwen3.6-27B*)	27B	≤ 44	59.24
StreamWeave (*Qwen3.5-397B-A17B*)	397B-A17B-FP8	≤ 44	58.79

C LongVideoBench Transfer with SigLIP2 Recent Curation

LongVideoBench tests offline long-video reasoning rather than a streaming query frontier. These completed runs use the same context layout and history retrieval as the main method but use SigLIP2, rather than each MLLM's released visual encoder, for recent-frame curation. Qwen3.6-27B obtains the strongest result among these five variants at 59.24, followed by Qwen3.5-397B-A17B at 58.79. Results for the main, model-specific-encoder configuration are not yet available on LongVideoBench.

D Detailed OVO-Bench Results

Table 10: Task-level OVO-Bench accuracy (%). SW denotes $+StreamWeave$ with at most 44 frames. The paired references use uniform 44-frame sampling. Bold and underline mark the highest and second-highest score for each task; ties share a mark.

	Qwen3-VL-8B		*Qwen3.5-9B*		*Qwen3-VL-32B*		*Qwen3.6-27B*		*Qwen3.5-397B-A17B*	
Task (n)	44	+SW	44	+SW	44	+SW	44	+SW	44	+SW
Backward										
ASI (148)	72.97	64.86	66.89	64.19	73.65	70.27	<u>75.68</u>	75.00	**77.03**	<u>75.68</u>
EPM (297)	54.88	55.89	61.95	61.95	60.27	59.60	65.66	**68.69**	64.98	<u>66.67</u>
HLD (186)	27.96	55.91	25.27	54.84	48.39	<u>59.14</u>	33.33	53.23	40.86	**73.66**
Real-time										
ACR (109)	65.14	76.15	67.89	73.39	69.72	77.06	71.56	**83.49**	70.64	<u>81.65</u>
ATR (116)	72.41	78.45	72.41	78.45	80.17	<u>81.90</u>	80.17	81.03	79.31	**85.34**
FPD (101)	70.30	70.30	75.25	76.24	75.25	**78.22**	<u>77.23</u>	**78.22**	**78.22**	75.25
OCR (149)	78.52	88.59	81.88	89.26	83.22	87.92	81.88	<u>91.95</u>	80.54	**92.62**
OJR (184)	63.04	73.91	68.48	76.63	65.76	73.91	71.74	<u>79.35</u>	72.28	**80.98**
STU (178)	60.11	58.99	64.04	64.61	59.55	65.73	66.29	<u>67.98</u>	61.24	**70.79**
Forward										
CRR (240)	54.58	57.08	52.92	64.58	52.92	59.17	46.25	**68.33**	60.83	<u>65.42</u>
REC (698)	41.83	33.09	38.25	40.26	40.83	32.81	**48.85**	44.70	<u>48.28</u>	41.83
SSR (629)	67.09	70.91	<u>71.86</u>	62.64	69.00	63.43	54.21	**72.97**	70.59	64.07

E Detailed OVBench Results

The six task-family means in Table 3 are computed from the subtasks below. The official Avg. gives equal weight to each of the 16 subtasks, irrespective of family size.

Table 11: OVBench accuracy by all 16 subtasks (%). Columns follow the official six-task, 16-subtask taxonomy. Avg. is the unweighted mean of the 16 subtask accuracies. Subtasks are shown to one decimal and Avg. to two. Methods without a public task breakdown report only Avg.; colored subscripts in Avg. show differences from the corresponding 44-frame reference. Bold and underlined scores mark the best and second best.

Method	Params	#Frames	FP			THV			PM			SP		STP		TP			
			AA	GSP	MP	AP	SV	OPr.	AR	PR	TR	AL	OPos.	AT	OT	AS	SL	OES	Avg.
Training-based methods																			
VideoChat-Online (Huang et al., 2025)	4B	2 fps	64.1	59.7	16.6	63.1	58.3	62.8	**42.2**	54.4	70.6	**54.1**	24.8	88.7	48.5	73.0	25.9	71.7	54.90
StreamForest (Zeng et al., 2026)	7B	1 fps	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	60.50
Off-the-shelf MLLMs																			
Gemini-1.5-Flash (Team et al., 2024)	–	$\leq$ 64	**71.4**	53.6	21.9	56.5	60.8	40.6	36.7	47.9	62.5	32.3	37.5	87.0	50.0	83.3	22.3	46.9	50.70
LLaVA-OV-7B (Li et al., 2024a)	7B	$\leq$ 64	68.0	62.7	35.9	58.4	50.3	46.5	29.4	60.7	58.0	43.1	14.2	86.5	49.7	70.7	28.1	30.2	49.50
Qwen2-VL-7B (Wang et al., 2024)	7B	$\leq$ 64	60.3	66.1	22.1	54.9	51.5	51.1	37.8	64.4	69.3	35.3	28.5	**97.0**	49.4	65.1	30.8	11.7	49.70
Qwen3-VL-8B	8B	44	57.7	66.7	26.1	53.2	59.8	69.2	26.3	68.0	65.3	32.8	51.4	88.3	54.7	74.4	26.9	61.5	55.14
+StreamWeave	8B	$\leq$ 44	65.4	74.1	22.7	56.7	59.4	63.7	29.1	63.2	64.7	36.9	44.3	88.3	52.5	75.3	27.7	57.8	55.10 (-0.04)
Qwen3.5-9B	9B	44	53.8	70.5	29.4	46.4	62.2	58.2	33.4	73.9	73.6	34.7	49.5	77.0	51.6	79.5	26.8	62.5	55.18
+StreamWeave	9B	$\leq$ 44	67.9	75.4	25.8	57.9	61.2	68.6	28.8	65.4	66.9	45.0	56.0	86.1	48.8	**83.3**	30.9	63.7	58.23 (+3.05)
Qwen3-VL-32B	32B	44	55.1	66.3	27.9	56.7	57.2	64.6	29.7	73.4	77.6	31.3	**62.5**	85.7	48.1	80.0	28.4	58.8	56.45
+StreamWeave	32B	$\leq$ 44	67.9	75.0	23.3	56.2	63.3	67.7	30.3	71.4	67.5	36.6	57.0	94.3	57.8	80.9	30.3	51.4	58.18 (+1.73)
Qwen3.6-27B	27B	44	47.4	76.1	31.9	67.4	**65.3**	**72.0**	35.6	**76.3**	**86.2**	32.5	48.3	90.9	54.7	78.1	29.8	80.3	60.80
+StreamWeave	27B	$\leq$ 44	67.9	78.4	35.9	**69.5**	64.8	68.0	30.0	72.5	75.8	46.9	51.4	89.6	54.0	80.5	35.1	80.0	**62.51** (+1.70)
Qwen3.5-397B-A17B	397B-A17B-FP8	44	53.8	71.9	**36.8**	69.1	62.8	55.4	25.9	75.0	82.2	26.3	48.3	83.5	**59.9**	76.7	28.6	**85.2**	58.84
+StreamWeave	397B-A17B-FP8	$\leq$ 44	69.2	**78.9**	30.1	64.4	64.4	64.3	34.4	72.8	75.8	43.1	55.4	88.3	51.9	82.3	**35.2**	79.4	61.86 (+3.02)

F Reproducibility and Responsible Use

We record model and prompt versions, source identifiers and repository revisions, sampling windows, the semantic similarity threshold, protected positions, recall cap, image transport, generation settings, scorer version, and evaluation-item IDs. Per-question records retain the causal frontier, current and historical references, tool and repair actions, materialization, failures, and image counts. Protocol failures remain in the denominator; credentials are excluded from code, manifests, and result files.

The experiments are vision-only. Applications in surveillance, wearable devices, or persistent assistants should define consent, retention, access control, and human-review policies. An auditable evidence trace helps diagnose model behavior but does not remove privacy or deployment obligations.

G Qualitative Cases across Streaming Subtasks

Figures 3–10 show one correctly answered Qwen3.5-9B StreamWeave example for every OVO-Bench subtask and every StreamingBench Real-time task. Each row follows the observed video in time: purple frames were returned by history retrieval, orange frames belong to the uniformly sampled look-back, green frames to the recent glance, and gray frames were observed but not selected for that question. These labels come from the saved per-question evidence trace, not a schematic allocation. No frame lies beyond the question frontier. The OVO-Bench HLD example correctly returns "Unable to answer," as required by its annotation.

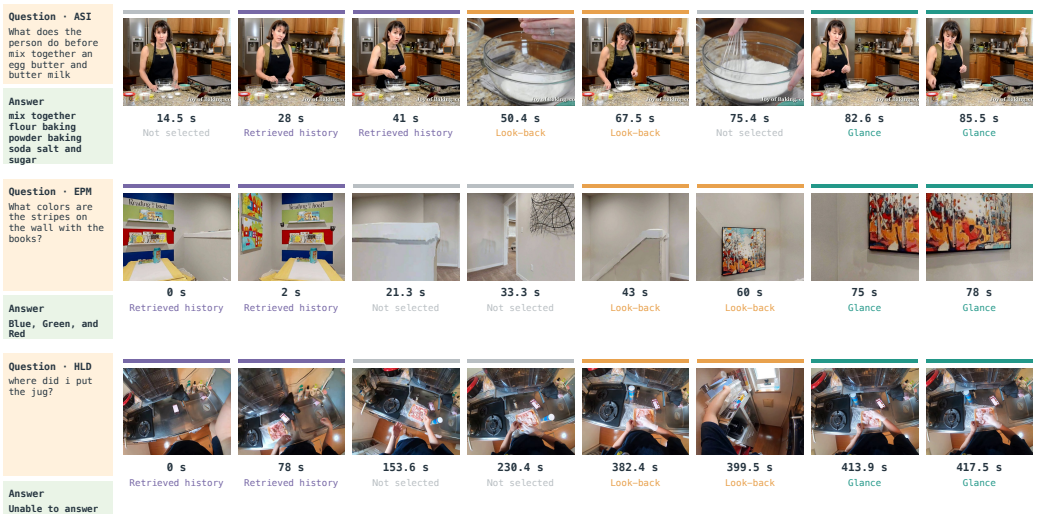


Figure 3: Successful OVO-Bench Backward cases: ASI, EPM, and HLD, top to bottom. The EPM query recovers a brief book-wall detail from the start of the observed video.

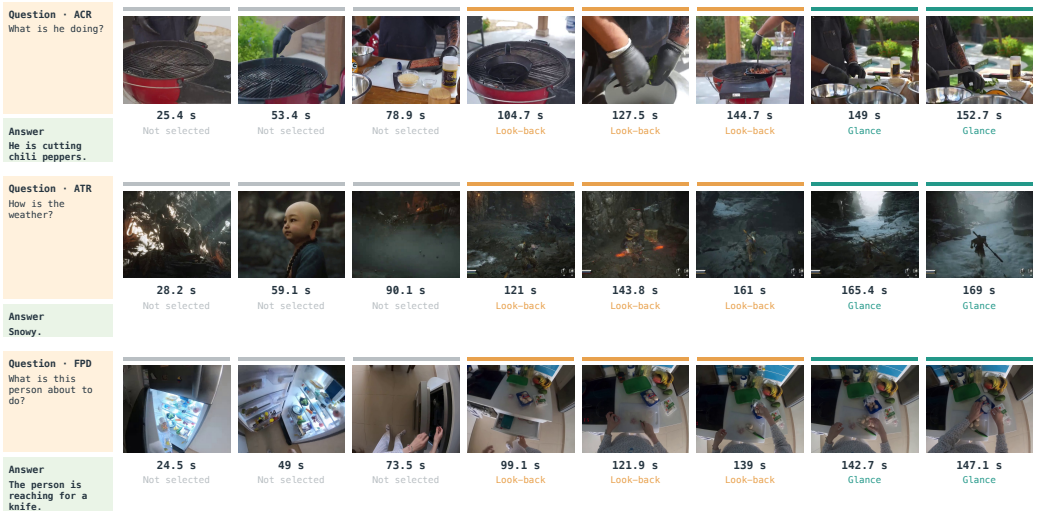


Figure 4: Successful OVO-Bench Real-time cases: ACR, ATR, and FPD.

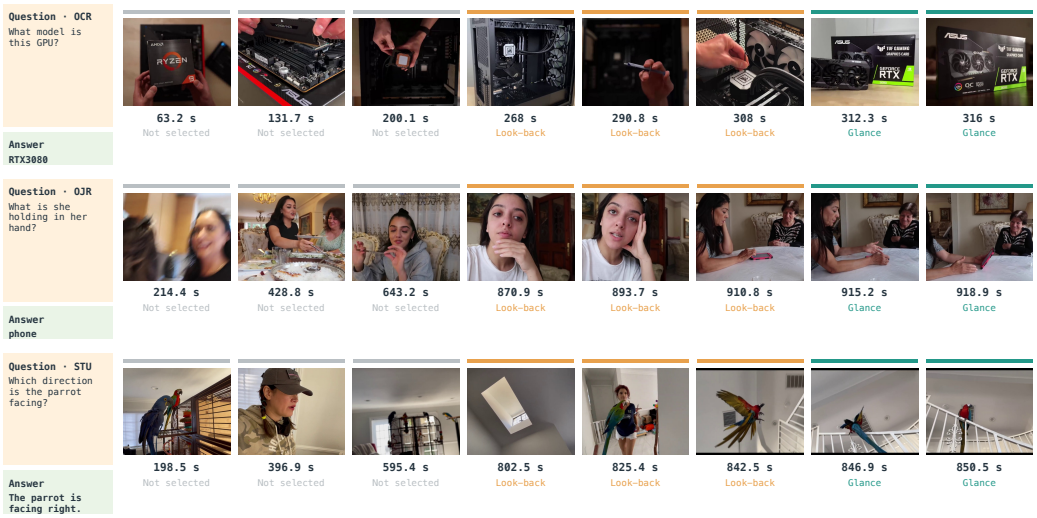


Figure 5: Successful OVO-Bench Real-time cases: OCR, OJR, and STU.

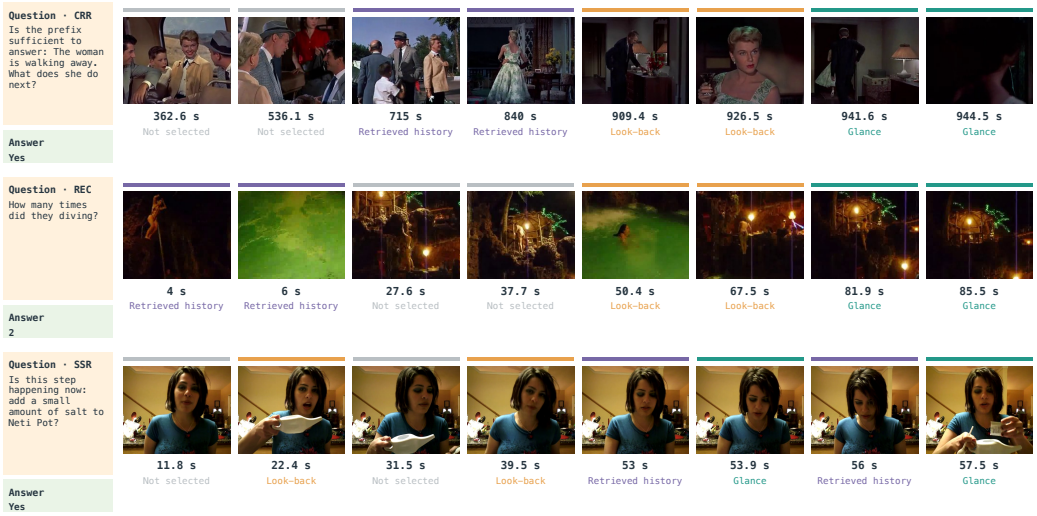


Figure 6: Successful OVO-Bench Forward cases: CRR, REC, and SSR. CRR and SSR are yes/no judgments; REC asks for a cumulative count.

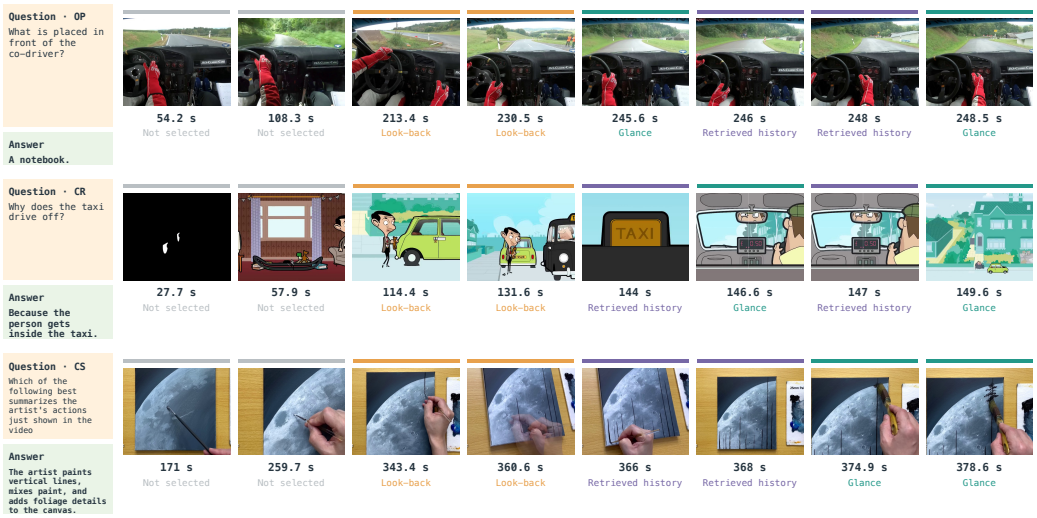


Figure 7: Successful StreamingBench Real-time cases: OP, CR, and CS.

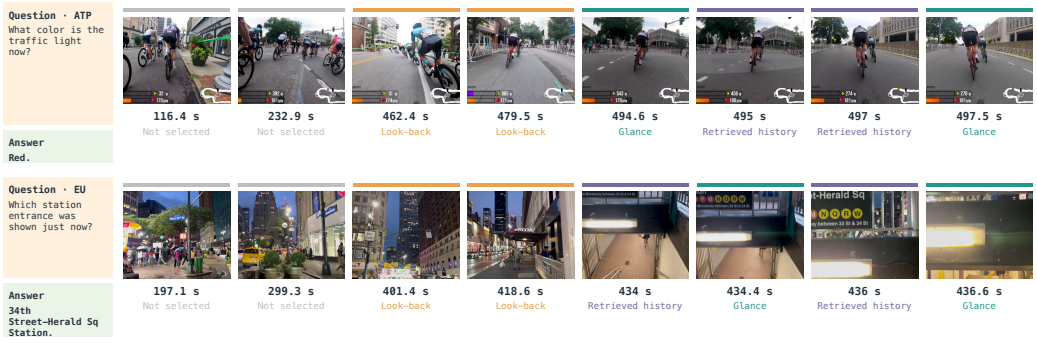


Figure 8: Successful StreamingBench Real-time cases: ATP and EU.

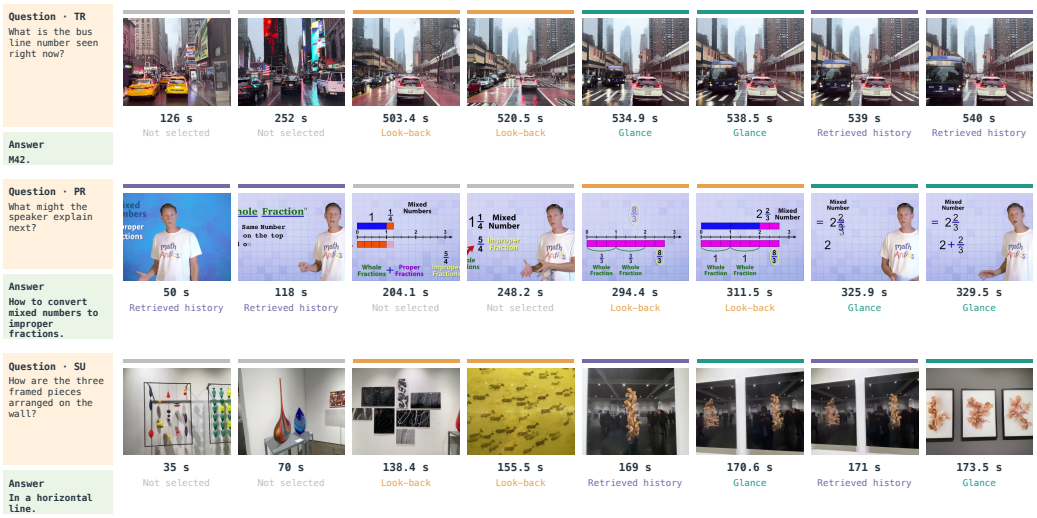


Figure 9: Successful StreamingBench Real-time cases: TR, PR, and SU.



Figure 10: Successful StreamingBench Real-time cases: ACP and CT.