EOV-BENCH: AN EVIDENCE-ORIENTED BENCHMARK FOR LONG-HORIZON STREAMING VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming video benchmarks are typically organized around predefined tasks, offering limited control over the evidence structure of questions. We introduce EOV-Bench, an evidence-oriented benchmark that constructs questions bottom-up from reusable, multi-scale video annotations. Evidence units are selected and composed before question formulation, controlling question type, temporal span, evidence distribution, and compositional demand—factors that shape difficulty. Reusing the annotation pool makes fuller use of long videos and yields diverse, explicitly grounded questions spanning current perception, historical retention, and cross-temporal reasoning. EOV-Bench contains 4,337 open-ended questions and supports matched retention analysis under a strict causal streaming protocol. Across evaluated strategies, high-fidelity recent frames favor current perception, whereas textual memory better preserves historical information over long time horizons. Ablations further show that long-term retention depends substantially on how observations are described and organized, not only on retrieval. These findings motivate AsyncMemFlow, a training-free framework that decouples memory construction from latency-sensitive answering. An independently configurable pipeline generates detailed descriptions and asynchronously consolidates them into structured memory. At query time, the answerer adaptively uses recent frames or the latest available memory, supplemented with observations not yet incorporated into it. This enables stronger memory construction without blocking online answering. Experiments demonstrate improved historical and joint reasoning while preserving strong current perception, with further gains on OVO-Bench.

1 INTRODUCTION

Four recent frames can be sufficient to match or surpass published streaming-video systems on StreamingBench (Lin et al., 2026) and OVO-Bench (Niu et al., 2025), as shown by SimpleStream (Shen et al., 2026). This result exposes a measurement gap: strong benchmark performance does not necessarily imply that a model preserves or uses long-range video evidence. Reliable evaluation therefore requires identifying the observations needed to answer each question and testing whether models can retain and use them as the stream unfolds. We introduce **EOV-Bench**, which addresses both requirements through evidence-first question construction and strict causal streaming evaluation (Figure 1). By constructing questions from annotated evidence and evaluating them causally, EOV-Bench turns long videos into targeted tests of current perception, historical retention, and reasoning across time.

Question construction is a central source of this measurement gap. Streaming-video benchmarks are commonly organized around predefined tasks, while the evidence structure underlying each question receives less explicit control. Recent benchmarks, including StreamArena, have advanced long-horizon, open-ended evaluation with annotated query times and supporting evidence (Zhang et al., 2026b). EOV-Bench takes a complementary step by making evidence selection and composition primary variables of question construction. Videos are independently described at three temporal scales to capture fine visual details, event-level developments, and broader temporal context. These descriptions form a reusable pool of temporally grounded evidence units, which are selected and composed before questions are formulated. Varying the selected units and the relationships asked

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.

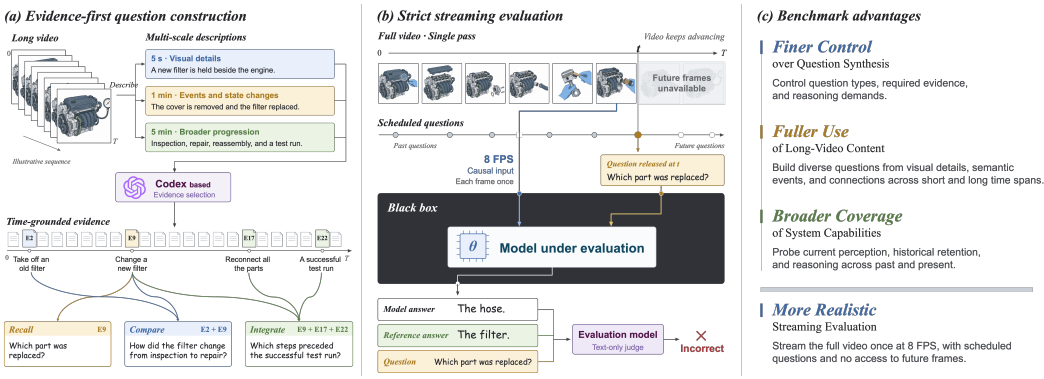

Figure 1: **Towards more reliable online video evaluation.** (a) Evidence-first question construction. (b) Causal streaming evaluation. (c) Benchmark advantages.

about them provides control over question type, temporal span, evidence count and distribution, and compositional demand—factors that shape difficulty. Reusing the same annotation pool produces question families spanning detailed recall, counting, cross-event comparison, and current–history association. This process makes fuller use of information distributed throughout long videos and broadens evaluation beyond any single task category.

To characterize these questions consistently, we introduce an *Evidence Dependency Space* with two complementary axes. Evidence Locality identifies whether support comes from the current observation, earlier history, or both, yielding Current, Historical, and Joint dependencies. Evidence Distribution distinguishes support concentrated within one interval from evidence spanning separated intervals, yielding Localized and Distributed dependencies. EOV-Bench contains 4,337 open-ended questions over approximately 54.3 hours of video across five domains and 24 categories, covering 17 question types. Each question retains explicit evidence intervals and reference answers. Evaluation follows the video clock: timestamped frames arrive sequentially, memory updates do not pause the stream, and questions become visible only at their scheduled times. Historical questions additionally support matched retention analysis, in which their wording, evidence, and answers remain fixed while query delay increases. A shared cohort is evaluated from 30 seconds to eight minutes, with duration-eligible subsets extending the evaluation to 16 and 32 minutes. Explicit grounding, causal delivery, and matched-question analysis together provide a more reliable basis for comparing streaming systems.

Results confirm the diagnostic value of EOV-Bench. With the same answering model, raw-frame inference is strongest on Current questions, whereas Historical and Joint questions favor different memory designs and can reverse model rankings. Historical accuracy also depends on how memory is prepared and exposed. A direct-access variant reusing existing OASIS (Liang et al., 2026) descriptions and node summaries raises its five-delay Historical mean by 7.78 points. Controlled ablations further show that stronger descriptions improve the same answerer, while semantic organization provides additional gains with dense descriptions held fixed. Memory exposure, description quality, and organization therefore matter alongside retrieval.

These findings motivate **AsyncMemFlow**, a training-free framework that decouples memory construction from latency-sensitive answering through independently configurable models and execution paths. A background pipeline converts completed video blocks into a query-independent Memory Map containing dense descriptions, timestamps, event headings, and scene summaries. At query arrival, the answer model uses recent raw frames for Current perception or the latest Map plus retained observations not yet incorporated into it for Historical and Joint reasoning. Because queries neither trigger nor wait for memory construction, stronger reusable memory can be prepared without blocking a compact, responsive answerer.

With a local 8B description model and 8B answerer, AsyncMemFlow outperforms all evaluated baselines at every Historical delay. Replacing only the description model with Qwen-VL-Max improves every delay further and allows even the 2B answerer to surpass all baselines at every Historical delay and on Joint questions. The 8B answerer retains a Current score of 75.19 versus 75.44

for raw frames. A fully local configuration also improves Perception and Backward accuracy on OVO-Bench.

Contributions. Our contributions are threefold: (1) an evidence-first construction paradigm that composes reusable, multi-scale evidence units into explicitly grounded questions with controllable types, temporal scopes, evidence distributions, and reasoning demands; (2) EOV-Bench, comprising 4,337 questions over 54.3 hours of video, an Evidence Dependency Space, and causal matched-retention evaluation, which reveals evidence-dependent ranking reversals and the importance of memory description and organization; and (3) AsyncMemFlow, a training-free framework that decouples memory construction from online answering, improving Historical and Joint reasoning while preserving Current perception and demonstrating gains on OVO-Bench.

2 Related Work

2.1 Memory for Streaming Video Understanding

Streaming video memory spans compressed visual or latent tokens in MA-LMM (He et al., 2024), Flash-VStream (Zhang et al., 2025); KV states in ReKV (Di et al., 2025), StreamMem (Yang et al., 2025), and HERMES (Zhang et al., 2026a); and textual or hybrid representations in ProVideLLM (Chatterjee et al., 2025) and ThinkStream (Liu et al., 2026b). ReKV retrieves stored video KV states, whereas StreamMem and HERMES maintain bounded caches.

OASIS (Liang et al., 2026) maintains a multi-resolution event hierarchy, FOLIO (Fan et al., 2026) uses entity-centered records, and EventMemAgent (Wen et al., 2026) combines event descriptions with visual anchors and change logs. Access ranges from FOLIO's structured matching with conditional semantic expansion to iterative search in M3-Agent (Long et al., 2026) and StreamMind (Zhang et al., 2026b). Memory construction and answering are already separated through OASIS's query-independent memory maintenance, M3-Agent's parallel memorization and control, and StreamMind's asynchronous writer and foreground answering. StreamChat (Xiong et al., 2025) uses independent threads for frame ingestion, memory maintenance, and query processing.

Existing methods improve how video history is stored and accessed, but successful retrieval alone does not ensure sufficient evidence for answering. AsyncMemFlow therefore emphasizes what memory preserves and how it is organized. Preparing detailed descriptions and their temporal context independently of online answering allows stronger memory construction to support compact answer models.

2.2 Online Video Understanding Benchmarks

StreamingBench (Lin et al., 2026), OVBench (Huang et al., 2025), and OVO-Bench (Niu et al., 2025) establish time-indexed evaluation of online video understanding. StreamingBench also analyzes temporal clue positions and single- versus multi-frame support. RIVER (Shi et al., 2026) reuses historical questions across query delays to study retention. SimpleStream (Shen et al., 2026) nevertheless obtains strong results from a few recent frames on StreamingBench's real-time visual understanding subset and OVO-Bench, motivating closer examination of question-level evidence requirements.

StreamArena (Zhang et al., 2026b) evaluates hour-scale, open-ended interactions with annotated supporting segments, including distributed historical evidence. SVCBench (Liu et al., 2026a) probes state maintenance through streaming counting, while OVO-S-Bench (Li et al., 2026) annotates spatial evidence intervals and screens for text-only shortcuts. ProReady-QA (Azad et al., 2026) evaluates response readiness, SPOT-Bench (Chatterjee et al., 2026) measures proactive response timing, and VSAS-Bench (Vasu et al., 2026) incorporates inference latency into streaming evaluation.

Existing benchmarks increasingly annotate supporting evidence, but annotations alone do not control which observations a question requires or how they must be combined. EOV-Bench therefore makes evidence selection and composition explicit when constructing questions. This enables targeted tests of perception, retention, and integration, helping reveal how different evidence demands expose the strengths and limitations of memory designs.

3 EOV-BENCH

EOV-Bench makes the evidence requirements of streaming video questions explicit at construction time. Starting from reusable, temporally grounded annotations, we select and compose supporting observations before formulating questions. This evidence-first process produces questions that probe current perception, historical recall, and integration across past and present observations, with support concentrated within one interval or distributed across multiple intervals. We further evaluate Historical questions at different query delays while keeping their wording, evidence, and reference answers fixed, examining whether the same questions remain answerable as the stream continues.

3.1 A DECOUPLED ABSTRACTION OF ONLINE VIDEO UNDERSTANDING

For a video stream $\mathcal{X} = (X_1, X_2, \ldots)$, only the observed prefix $X_{\leq t}$ is accessible at time t. A model updates its state independently of question arrival and answers a separately arriving question Q_{t_q} through

$$M_t = \mathcal{U}_\theta(M_{t-1}, X_t), \tag{1}$$

$$A_\tau = \mathcal{F}_\theta(Q_{t_q}, C_\tau, M_\tau), \qquad A_\tau \in \mathcal{A} \cup \{\perp\}, \tag{2}$$

where C_τ denotes current observations, M_τ may contain visual tokens, KV states, text, or structured events, and $\perp$ denotes no response. We evaluate *Querying*, which answers using evidence available at $\tau = t_q$. This separates continuous memory maintenance from question arrival. Across different memory representations, EOV-Bench characterizes each question through its supporting observations in the source video, providing a common basis for comparing evidence use. Appendix B.1 describes the broader response framework, including *Tracing*.

3.2 EVIDENCE DEPENDENCY SPACE

For each question, annotated supporting evidence consists of the video observations used to establish its reference answer, all occurring no later than its evaluation time. Two complementary axes describe where this evidence lies and how it is distributed over time.

Evidence Locality. Support is *Current* when it lies within C_τ, *Historical* when it lies entirely in the preceding video history, and *Joint* when it includes both. This distinguishes direct perception, historical answering, and the use of history to interpret current observations. For example, identifying who is hidden inside a sack in the current view requires an identity established earlier, making it a Joint question.

Evidence Distribution. Support is *Localized* when concentrated within one contiguous temporal interval and *Distributed* when spanning multiple separated intervals. These groups distinguish questions supported within one interval from those requiring observations across separated moments, such as comparisons between historical events. Together, the axes characterize evidence requirements independently of a model's memory representation. Further interpretation of these definitions appears in Appendix B.1.

3.3 EVIDENCE-FIRST CONSTRUCTION AND QUESTION SETTING

Question construction begins with selecting the observations needed to establish an answer. By varying the selected evidence and the relation asked about it, the same annotation pool supports different question families. For example, one historical shot supports event recall; two shots support cross-event comparison; and an earlier shot paired with the current view supports current–history association (Figure 2). Evidence selection and composition are thus explicit construction variables.

From reusable evidence to questions. Videos are directly annotated at three scales: 5-second clips capture visual details, 1-minute clips capture events and local state changes, and 5-minute clips capture broader event progression. Each scale is described independently from the video rather than summarized from shorter-scale text, providing complementary observations for evidence selection. A dedicated agent built with Codex for each question type selects supporting intervals according

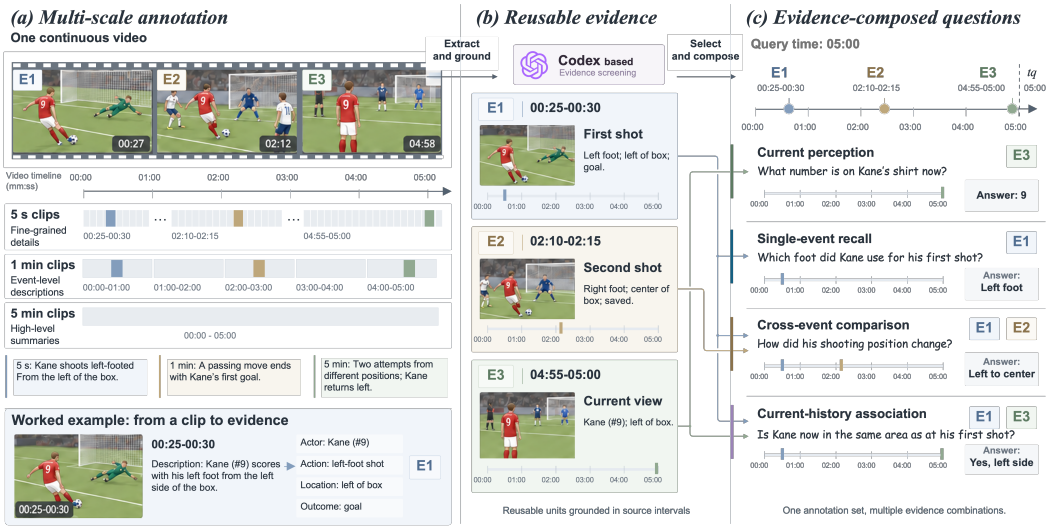

Figure 2: Evidence-first question construction. (a) Multi-scale video annotations capture details, events, and broader context. (b) Reusable evidence units retain source intervals and observations. (c) Selecting and composing these units produces questions with different evidence dependencies and explicit reference answers.

to the intended locality and distribution, checks cross-scale consistency, and constructs questions and reference answers from evidence without unresolved contradictions. Canonical answers and accepted alternative answers are manually reviewed for each question. Each instance records its query time, evidence intervals, canonical answer, and accepted paraphrases. Quality-control scope is discussed in Appendix B.3; annotation details and prompts appear in Appendix F.

Question families. The reported evaluation covers 17 Querying types: five Current, seven Historical, and five Joint types. These families assess complementary uses of the selected evidence.

Current questions assess directly observable text, objects, attributes, spatial relations, and entity counts, providing a perception reference without historical support.

Historical questions assess recalling past events and states, as well as combining observations across events through counting, comparison, and temporal reasoning. For example, comparing the outcomes of a kitchen wok fire and a later barn fire requires recovering both events and establishing their difference.

Joint questions assess using history to interpret the present through reference resolution, state changes, current–history comparison, multi-event context grounding, and state evolution. They test whether retained information can be connected to the ongoing situation.

EOV-Bench comprises 4,337 open-ended questions spanning approximately 54.3 hours of content from 49 videos across five domains and 24 categories. Appendix B.2 explains the complete taxonomy and evaluated scope; Appendix D provides the question types and representative examples.

3.4 Evaluation Protocol

Strict real-time streaming. To evaluate evidence use as video unfolds, main experiments deliver timestamped 8-FPS frames sequentially according to the video clock, without pausing the stream for memory updates. Questions are revealed only at their scheduled video times. Future frames are inaccessible, and past observations cannot be replayed unless retained by the evaluated method.

Matched-question retention. To examine historical retention as the stream continues, we keep each question, its supporting evidence, and its reference answer fixed while varying the query delay. For question q, queries are issued at $t_q = b_q + \delta$, where b_q marks when its support-

ing evidence has been observed and any question-defined observation window has ended, and $\delta \in \{30, 60, 120, 240, 480, 960, 1920\}$ seconds. A shared cohort of 1,217 questions is evaluated from 30 seconds to eight minutes, supporting retention analysis on the same questions. The 16- and 32-minute evaluations use duration-eligible subsets, so comparisons across these longer delays also reflect cohort changes. Appendix B.4 explains eligibility and interpretation; detailed counts appear in Appendix H.

Open-ended answer evaluation. Open-ended questions remove candidate-answer cues. Models are instructed to answer within five words or output `REJECT` when evidence or confidence is insufficient. Exact canonical or accepted answers receive full credit; other responses are scored by a frozen text-only judge using task-specific rubrics and scores in $\{0, 0.5, 1\}$. For comparison, transition, and sequence questions, scoring evaluates the requested relation, direction, or order rather than isolated correct facts. Empty, failed, and rejected responses score zero. The scorer is calibrated using a separate, manually reviewed batch of Raw Frames outputs from Qwen2.5-VL that is excluded from the final result tables. Reported evaluation results are not manually adjusted. Appendix B.4 explains scoring scope; the scoring criteria and aggregation rules appear in Appendix G.

4 AsyncMemFlow

To accommodate the computation required for richer descriptions, AsyncMemFlow builds on the decoupled formulation in Section 3.1 and runs memory construction asynchronously through an independently configured pipeline, while the answerer remains ready to respond to incoming queries. Completed video blocks are described in detail and organized into a query-ready Memory Map. To preserve the advantage of recent raw frames for current perception, the answerer routes Current questions directly to recent visual observations. Historical and Joint questions receive the latest completed Map together with retained frames covering observations not yet incorporated into memory. Queries consume evidence available at arrival without waiting for ongoing description generation or memory organization (Figure 3).

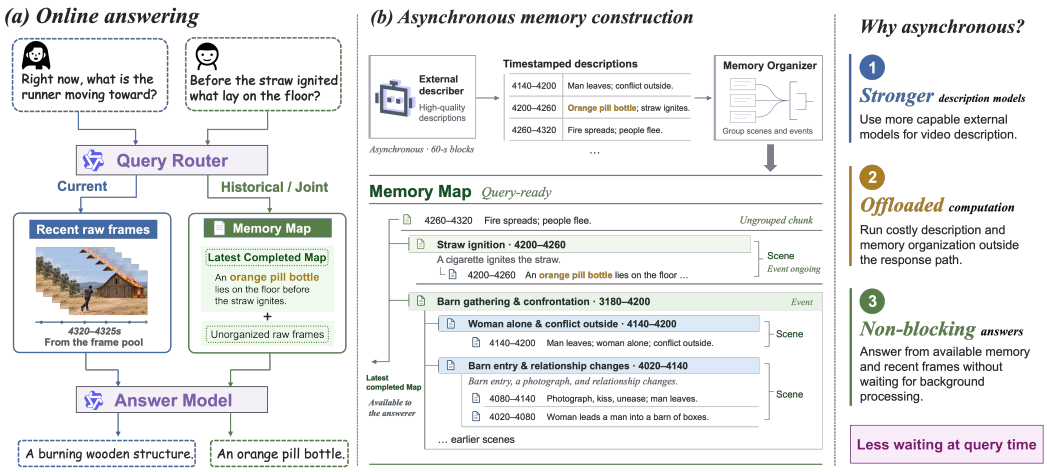

Figure 3: Overview of **AsyncMemFlow**. (a) The router directs each question to the corresponding evidence source. (b) The describer and memory organizer asynchronously prepare a query-ready Memory Map. The actual descriptions are provided in Appendix E. Images in this figure have been processed for visual presentation.

4.1 Asynchronous External Description

A separate description model builds detailed textual history in the background, allowing the online answerer to remain compact and swift. We use Qwen-VL-Max to describe completed one-minute video blocks, recording entities, actions, visible states, and temporal progression.

Blocks are processed in temporal order, and completed descriptions pass to the external memory organizer. Frame ingestion and online answering continue while description and organization execute. Queries use already completed memory, with no requirement to finish an ongoing construction call. This separation permits stronger describers and richer visual inputs to prepare reusable historical evidence before it is needed for answering.

4.2 Asynchronous Memory Organization

Memory Map pairs event context with preserved local detail. An external organizer groups descriptions into scenes and related scenes into larger events, adding summaries and headings while retaining the associated dense descriptions and timestamps. The Map accommodates grouped history alongside recent descriptions that have not yet been assigned to a completed group (Appendix E).

This structure makes both the course of an event and its specific observations available to the answer model. For example, a scene summary records how straw catches fire, while the associated dense description retains the orange pill bottle visible before ignition. The summary provides event context, and the detailed record answers what lay on the floor before the fire.

Organization proceeds in the background, independently of questions. At query time, the latest completed Map is supplied in full as answer context. The model can jointly examine observations from different events without issuing retrieval requests or selecting memory branches. Preparing this structure in advance supports both detailed historical recall and cross-event integration within a single answer-generation call.

4.3 Evidence-Adaptive Querying

Online answering combines immediate visual access with prepared semantic memory. At question arrival, the answer model first acts as a text-only router. CURRENT questions receive the latest five causal frames. HISTORICAL/JOINT questions receive the latest completed Memory Map together with sampled retained frames not yet incorporated into it. This visual supplement follows the uncovered portion of the retained frame pool, giving the answerer access to observations whose description or organization is still pending.

All evidence is frozen at question arrival: a query neither triggers nor waits for memory construction. The answerer uses the available history and visual evidence while the background pipeline continues preparing memory for later questions. Sampling and routing settings appear in Appendix H.1; execution and resource scope are discussed in Appendix B.5.

5 Experiments

Experimental Setup. We follow the strictly causal protocol in Section 3.4, without pausing the video clock for memory updates. AsyncMemFlow uses Qwen3-VL-Instruct (Bai et al., 2025a) answer models at 2B, 4B, and 8B scales, with Qwen-VL-Max as the description model unless otherwise specified. EOV-Bench scores are reported on a 0–100 scale under the shared partial-credit protocol. Scoring and implementation details appear in Appendices G and H.

5.1 Main Results

Our architecture enables stronger historical reasoning with compact answer models. Our architecture delivers the strongest memory retention among the evaluated methods, achieving the highest Historical accuracy across all seven delays while maintaining low response latency. With Qwen-VL-Max constructing memory, even AsyncMemFlow's 2B answer model exceeds every listed baseline. These results demonstrate the value of preparing informative historical context for a compact answer model. These gains are achieved while maintaining low response latency, with isolated mean response times of 0.923–1.496 seconds. Timing scope is detailed in Appendix B.5.

Different evidence requirements favor different strategies. The capability profiles in Table 1 explain why evidence-adaptive querying matters. In comparisons using the same answer model, the evaluated memory methods do not improve aggregate Current performance over direct raw-frame

Table 1: Main results on EOV-Bench (0–100 scores). Bold/underlined scores rank first/second in each column. Historical columns summarize five delays from 30 seconds to eight minutes; Table 2 reports the full seven-delay evaluation.

	Current						Historical (5-delay mean)								Joint					
	Localized						Localized			Distributed					Localized			Distributed		
Method	EAR	ECR	OCR	OR	SRR	**Avg.**	ER	ESR	LCR	MER	EC	CEC	TRR	**Avg.**	CRR	SCR	CHC	MCG	SER	**Avg.**
Raw Frames (Qwen3-VL-8B)	**83.02**	**74.60**	**81.10**	**70.32**	**68.15**	**75.44**	1.32	2.20	1.72	7.59	0.83	1.11	22.75	5.36	10.22	4.60	5.17	11.76	6.45	7.64
Raw Frames (LLaVA-OneVision)	73.92	56.51	69.23	55.36	56.60	62.32	5.38	10.50	2.50	7.04	12.50	3.52	7.50	6.99	14.52	8.64	8.19	8.82	0.00	8.03
ReKV (Di et al., 2025)	39.03	37.01	29.34	29.66	41.92	35.39	10.62	25.37	4.53	20.56	28.33	14.07	15.68	17.02	14.84	5.99	10.18	5.88	0.00	7.38*
HERMES (Zhang et al., 2026a) (Q)	67.26	70.10	69.83	61.07	64.22	66.50	20.29	36.17	17.81	32.59	46.67	23.15	58.52	33.60	39.25	10.66	17.24	23.53	16.13	21.36
HERMES (Zhang et al., 2026a) (L)	65.54	58.58	57.01	49.30	50.91	56.27	10.74	22.23	8.28	19.81	31.67	8.33	17.63	16.95	24.73	6.80	10.78	26.47	0.00	13.76
StreamForest (Zeng et al., 2025)	63.98	60.77	57.33	47.44	51.12	56.13	18.75	31.13	15.31	21.11	51.67	14.26	24.62	25.26	27.96	7.72	7.33	11.76	6.45	12.24
Flash-VStream (Zhang et al., 2025)	61.97	55.14	44.33	45.00	64.44	54.18	5.58	13.44	2.34	22.59	24.17	6.48	28.47	14.73	7.53	0.74	4.31	11.76	0.00	4.87
VST (Guan et al., 2026)	67.42	69.53	69.42	52.41	51.39	62.03*	21.06	29.57	20.59	37.00	46.67	12.09	28.38	27.91*	34.51	7.80	8.26	30.00	5.36	17.19*
Dispider (Qian et al., 2025)	46.56	53.12	16.58	37.85	37.06	38.23*	15.45	30.15	22.19	21.11	40.83	7.78	18.81	22.33	19.23	11.80	19.47	0.00	9.68	12.04*
OASIS (Liang et al., 2026)	74.51	69.71	76.88	69.91	65.42	71.29	17.97	28.34	24.38	36.11	33.33	21.67	40.38	28.88	38.71	**43.38**	41.38	35.29	20.97	35.95
AsyncMemFlow (2B)†	75.33	69.09	78.47	62.56	54.40	67.97	34.51	42.91	40.78	47.22	**65.83**	26.67	33.22	41.59	49.46	42.83	48.70	32.35	**48.38**	44.34
AsyncMemFlow (4B)†	79.92	67.75	79.37	61.84	62.32	70.24	**37.54**	48.16	38.91	56.48	53.33	**44.26**	54.92	47.66	**61.29**	42.46	46.55	**41.17**	41.94	**46.68**
AsyncMemFlow (8B)†	**83.02**	**74.60**	**81.10**	69.41	67.83	75.19	37.50	**50.98**	**45.16**	**61.11**	55.83	36.67	**62.71**	**49.99**	58.60	37.32	**49.13**	38.23	43.54	45.36

Note: Avg. macro-averages the 5/7/5 Current/Historical/Joint subtypes. Each subtype pools evaluated questions across videos, rather than averaging video scores. Historical cells average 30-s, 1-min, 2-min, 4-min, and 8-min scores on the shared 1,217-question cohort; Historical Avg. averages these five delay-specific macro scores. LC, CCR, and Tracing are not evaluated. *Incomplete evaluation coverage. †Qwen-VL-Max supplies descriptions; 2B/4B/8B denotes the answer model. (Q) and (L) denote the Qwen3-VL-8B and LLaVA-OneVision backbones, respectively.

inference. AsyncMemFlow preserves this visual access, scoring 75.19 with the 8B answer model, close to the raw-frame score of 75.44. Historical and Joint questions change the ranking: HERMES–Qwen3-VL-8B exceeds OASIS on Historical but falls below it on Joint. EOV-Bench exposes these distinct demands for current perception, historical recall, and current–history integration, supporting evidence-specific design choices.

Table 2: Historical scores across seven query delays and isolated mean response time (s). Bold/underlined values rank first/second (higher scores, lower times).

Method	**30s**	**1min**	**2min**	**4min**	**8min**	**16min**	**32min**	**Avg. response time (s) ↓**
Raw Frames (Qwen3-VL-8B)	8.46	5.79	4.42	4.39	3.75	3.83	3.71	**0.384**
ReKV	16.91	17.27	17.04	18.40	15.50	15.35	13.82	0.879
HERMES (LLaVA-OneVision)	24.98	20.19	15.34	13.85	10.41	13.06	11.22	0.608
HERMES (Qwen3-VL-8B)	42.24	36.79	31.58	30.35	27.04	26.62	22.17	0.563
StreamForest	29.70	23.95	25.69	23.46	23.52	24.94	24.52	1.134
Flash-VStream	16.52	16.65	14.26	13.00	13.20	12.72	12.94	1.186
VST*	36.83	29.09	26.84	25.32	21.47	23.37	16.16	1.586
Dispider	23.35	21.58	22.89	21.08	22.74	21.88	20.79	1.382
OASIS	33.04	31.21	27.56	26.47	26.13	24.43	23.71	6.947
OASIS (our version)†	41.65	39.21	37.20	33.86	31.36	34.17	30.77	7.122
AsyncMemFlow (Qwen3-VL-2B)	45.95	41.28	40.09	41.02	39.61	38.47	37.64	0.923
AsyncMemFlow (Qwen3-VL-4B)	52.79	**49.49**	45.78	44.08	46.15	46.66	47.13	1.243
AsyncMemFlow (Qwen3-VL-8B)	**55.70**	48.95	**47.14**	**47.57**	**50.61**	**48.74**	**51.59**	1.496

Note: Scores macro-average seven Historical subtypes, each pooling evaluated questions across videos, including REJECT responses. The first five delays share 1,217 questions; 16/32-minute results use duration-eligible subsets of 1,018/706 questions. Cross-delay comparisons at these longer delays are subject to cohort changes. *Incomplete evaluation coverage. † Direct access to flattened OASIS descriptions and node summaries; see Appendix H.2.

Generalization to OVO-Bench. On OVO-Bench, a fully local AsyncMemFlow configuration achieves 79.94% Perception and 60.17% Backward accuracy. The OASIS source table (Liang et al., 2026) contains apparent inconsistencies between its subtask scores and reported averages. Even with the weaker local 8B description model, AsyncMemFlow exceeds both the reported OASIS Qwen3-VL-8B averages (by 1.80 and 2.96 points, respectively) and averages recomputed from its listed subtask scores. Neither description generation nor memory organization uses external APIs. Full results and the numerical comparison appear in Appendix C.

5.2 Memory Analysis and Ablations

We examine three factors that determine the usefulness of historical memory: how existing evidence is exposed to the answer model, the quality of the descriptions, and the organization of memory. Comparison scope is discussed in Appendix B.6.

Memory access strategies substantially affect historical performance. Our OASIS variant reuses its generated descriptions and node summaries, in a flattened context supplied directly to the answer model without retrieval. This changes how existing OASIS memory is presented for answering. The variant improves Historical scores by 5.23–9.74 points across all seven delays (Table 2); its five-delay mean on the fixed cohort rises from 28.88 to 36.66. These gains show that the input organization and access policy can substantially affect the utility of already generated historical evidence. Appendix H.2 distinguishes this configuration from OASIS's published ablations.

Higher-quality descriptions improve historical recall. Holding the 8B answer model and Map setting fixed, Qwen-VL-Max descriptions outperform those from a separate local 8B description model at every delay (Table 3). Gains range from 3.69 to 10.01 points with Map organization and from 8.41 to 16.35 points without it. Better historical descriptions therefore benefit the same answer model under both input organizations, supporting independent investment in the description model.

Table 3: Ablation of the description model, Memory Map organization, and answer model capacity on Historical questions, macro-averaged over seven subtypes. All answer models are Qwen3-VL (Bai et al., 2025a) models.

Answer model	Description model	Map	30s	1min	2min	4min	8min	16min	32min
2B	Qwen-VL-Max	✓	45.95	41.28	40.09	41.02	39.61	38.47	37.64
2B	Qwen-VL-Max	×	45.07	40.24	38.20	36.38	35.60	36.33	36.62
4B	Qwen-VL-Max	✓	52.79	49.49	45.78	44.08	46.15	46.66	47.13
4B	Qwen-VL-Max	×	49.83	43.22	38.37	36.10	35.07	34.87	32.12
8B	Qwen-VL-Max	✓	55.70	48.95	47.14	47.57	50.61	48.74	51.59
8B	Qwen-VL-Max	×	52.94	47.02	41.03	41.06	42.32	42.26	43.52
8B	Qwen3-VL-8B-Instruct*	✓	49.87	45.26	42.04	41.46	41.02	42.10	41.58
8B	Qwen3-VL-8B-Instruct*	×	44.53	35.77	30.02	27.91	26.28	27.09	27.17

* A separate Qwen3-VL-8B-Instruct instance on the same server generates descriptions asynchronously with thinking disabled; it can even share a single A100 GPU with the answer model. Map-on/off use identical questions at each delay. The first five delays share one cohort; 16/32-minute evaluations use subsets (Section 3.4).

Memory Map effectively mitigates memory decay. Map-off retains the dense descriptions as a flat sequence; Map-on adds event headings, scene summaries, and hierarchical grouping. With the description model and answer model fixed, Map improves aggregate Historical scores at every delay in all four configurations (Table 3). It also reduces the score decline from 30 seconds to eight minutes on the same 1,217 questions in every configuration. For the 4B answer model with VL-Max descriptions, this decline shrinks from 14.76 to 6.64 points. With the local 8B description model, Map raises the eight-minute score from 26.28 to 41.02. Organization helps preserve the usefulness of stored descriptions as more video accumulates.

Table 4: Effect of Memory Map organization on Historical performance across seven query delays.

Answer model	Description model	30s	1min	2min	4min	8min	16min	32min
2B	Qwen-VL-Max	+0.88	+1.04	+1.89	+4.64	+4.01	+2.14	+1.02
4B	Qwen-VL-Max	+2.96	+6.27	+7.41	+7.98	+11.08	+11.79	+15.01
8B	Qwen-VL-Max	+2.76	+1.93	+6.11	+6.51	+8.29	+6.48	+8.07
8B	Qwen3-VL-8B-Instruct	+5.34	+9.49	+12.02	+13.55	+14.74	+15.01	+14.41

6 Conclusion

We introduced EOV-Bench by jointly addressing question construction and evaluation for long-horizon streaming video understanding. Bottom-up construction from reusable evidence annotations makes question dependencies explicit, while strictly causal evaluation and matched retention analysis test whether systems preserve and use the required evidence as video unfolds. EOV-Bench reveals that streaming video understanding should be examined through distinct evidence capabilities: method rankings reverse across Current, Historical, and Joint questions, and existing strategies struggle to balance these capabilities. Guided by these findings, AsyncMemFlow separates asynchronous high-quality memory construction from answering that remains ready for incoming questions. Routing preserves direct visual access for Current questions, while a query-ready Memory Map organizes historical evidence for Historical and Joint reasoning. Our results demonstrate that this design alleviates capability trade-offs and mitigates historical performance decay as video accumulates.

AI USE STATEMENT

Generative AI tools were used in benchmark construction, video description and memory generation, question answering, and model-based evaluation, as detailed in the main text and appendices. During manuscript preparation, generative AI tools also assisted with language editing, translation, LaTeX and table formatting, and code development. The authors reviewed AI-assisted materials against the source videos, annotations, and experimental outputs, and take responsibility for the final content, results, and conclusions of this work.

ETHICS STATEMENT

Streaming-video memory can retain sensitive information and propagate errors from generated descriptions. Real-world deployment should respect the privacy and usage restrictions of source videos and account for potential inaccuracies in stored memories and answers.

REPRODUCIBILITY STATEMENT

The paper and appendices document benchmark construction, causal evaluation protocols, model configurations, key prompt templates, and scoring procedures, together with supplementary results and ablations. We will publicly release the EOV-Bench dataset, benchmark construction and evaluation code, and the AsyncMemFlow implementation to support reproduction and further research.

REFERENCES

Shehreen Azad, Vibhav Vineet, and Yogesh S Rawat. StreamReady: Learning what to answer and when in long streaming videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40494–40504, 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, 2025a. 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, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Dibyadip Chatterjee, Edoardo Remelli, Yale Song, Bugra Tekin, Abhay Mittal, Bharat Bhatnagar, Necati Cihan Camgoz, Shreyas Hampali, Eric Sauser, Shugao Ma, Angela Yao, and Fadime Sener. Streaming videollms for real-time procedural video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22586–22598, October 2025.

Dibyadip Chatterjee, Zhanzhong Pang, Fadime Sener, Yale Song, and Angela Yao. Don't pause! every prediction matters in a streaming video, 2026. URL https://arxiv.org/abs/2604.24317.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. VideoLLM-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18407–18418, June 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, ZhongTao, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 42115–42127, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/67a9b444cbcd647572c88194619f72d5-Paper-Conference.pdf.

Haoyang Fan, Dhruv Parikh, Anvitha Ramachandran, Sameh Gobriel, Nilesh Jain, Rajgopal Kannan, and Viktor Prasanna. Folio: Focused semantic memory for streaming video understanding, 2026. URL https://arxiv.org/abs/2607.13298.

Yiran Guan, Liang Yin, Dingkang Liang, Jianzhong Ju, Zhenbo Luo, Jian Luan, Yuliang Liu, and Xiang Bai. Video streaming thinking: Videollms can watch and think simultaneously. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 80–99, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-37255-0.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13504–13514, 2024. doi: 10.1109/CVPR52733.2024.01282.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3328–3338, June 2025.

Yifei Li, Pengyiang Liu, Yuhang Zang, Zhongyue Shi, Qi Fu, Hongye Hao, and Jiwen Lu. Ovo-s-bench: A hierarchical benchmark for streaming spatial intelligence in multimodal llms, 2026. URL https://arxiv.org/abs/2606.03890.

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 (CVPR)*, pp. 2821–2831, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liang_OASIS_On-Demand_Hierarchical_Event_Memory_for_Streaming_Video_Reasoning_CVPR_2026_paper.html.

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, 2026. doi: 10.1109/ICASSP55912.2026.11463959.

Pengyiang Liu, Zhongyue Shi, Hongye Hao, Qi Fu, Xueting Bi, Siwei Zhang, Xiaoyang Hu, Zitian Wang, Linjiang Huang, and Si Liu. Svcbench: A streaming video counting benchmark for spatial-temporal state maintenance. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 522–538, Cham, 2026a. Springer Nature Switzerland. ISBN 978-3-032-37044-0.

Zikang Liu, Longteng Guo, Handong Li, Ru Zhen, Xingjian He, Ruyi Ji, Xiaoming Ren, Yanhao Zhang, Haonan Lu, and Jing Liu. Thinking in streaming video. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 21–38, Cham, 2026b. Springer Nature Switzerland. ISBN 978-3-032-37447-9.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 146197–146246,

2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ec370e882e8a02be955679ddbd2731b8-Paper-Conference.pdf.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18902–18913, 2025. doi: 10.1109/CVPR52734.2025.01761.

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

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

Yansong Shi, Qingsong Zhao, Tianxiang Jiang, Xiangyu Zeng, Yi Wang, and Limin Wang. River: A real-time interaction benchmark for video llms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 9447–9471, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/1022661f3f43406065641f16ce25eafa-Paper-Conference.pdf.

Pavan Kumar Anasosalu Vasu, Cem Koc, Fartash Faghri, Chun-Liang Li, Bo Feng, Zhengfeng Lai, Meng Cao, Oncel Tuzel, and Hadi Pouransari. Vsas-bench: Real-time evaluation of visual streaming assistant models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9654–9663, June 2026.

Siwei Wen, Zhangcheng Wang, Xingjian Zhang, Lei Huang, and Wenjun Wu. Eventmemagent: Hierarchical event-centric memory for online video understanding with adaptive tool use, 2026. URL https://arxiv.org/abs/2602.15329.

Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 69332–69351, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ad2fa437f7c23e4e9875599c6065d18a-Paper-Conference.pdf.

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, 2025. URL https://arxiv.org/abs/2508.15717.

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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 75804–75835. Curran Associates, Inc., 2025. doi: 10.52202/085713-2547. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6dd91fec726dbed8915a1fbadd91d1d2-Paper-Conference.pdf.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams, 2024. URL https://arxiv.org/abs/2406.08085.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-VStream: Efficient real-time understanding for long video streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21059–21069, October 2025. 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 Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8411–8430, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.381. URL https://aclanthology.org/2026.acl-long.381/.

Xichen Zhang, Guankai Li, Yinghao Zhu, Shijian Wang, Sitong Wu, Shaozuo Yu, Meng Chu, Yuan Lu, and Jiaya Jia. Streamarena: Toward continuous, interactive, and long-horizon agentic streaming video understanding, 2026b. URL https://arxiv.org/abs/2608.05703.

APPENDIX

A LIMITATIONS

This work focuses on long-horizon memory in streaming video. EOV-Bench evaluates Current, Historical, and Joint questions under the Querying mechanism. Although the broader formulation includes Tracing, which retains an instruction until its triggering condition becomes observable, neither Current nor Joint Tracing is evaluated. Extending evaluation to these response-triggering capabilities remains future work.

Our findings on memory access are limited by the answer-model scales and implementations evaluated. The limited benefit of retrieval may partly reflect smaller models' difficulty in executing retrieval effectively, rather than an inherent limitation of retrieval-based memory. Poorer retrieval execution observed with a 4B answer model is consistent with this explanation, but further evaluation with larger models is needed; such experiments were beyond our available computational resources.

Limited availability of implementations also restricted our memory-access interventions to OASIS, a representative structured retrieval-based memory system. Whether the same findings hold across other retrieval architectures remains to be established.

Finally, for computational economy, evaluation on OVO-Bench was restricted to the Qwen3-VL-8B configuration. Although this configuration outperforms the compared methods, cross-benchmark validation across additional model scales remains future work.

B SUPPLEMENTARY EXPLANATIONS AND DISCUSSION

This appendix expands on EOV-Bench in Section 3 and AsyncMemFlow in Section 4. It clarifies the evidence definitions, evaluated question types, quality control, and score interpretation, together with the execution and resource scope of asynchronous memory construction and the interpretation of experimental comparisons.

B.1 RESPONSE MECHANISMS AND EVIDENCE DEFINITIONS

The formulation in Section 3.1 separates memory maintenance from question arrival and accommodates two response mechanisms. *Querying* answers from evidence available at $\tau = t_q$. *Tracing* retains an instruction until its condition becomes observable at $\tau \geq t_q$, with $\perp$ representing no response. These times index video time, not inference latency. The present experiments evaluate Querying; Tracing describes the broader response-triggering framework.

Interpreting evidence locality and distribution. Evidence attributes describe the annotated source-video support used to establish a reference answer, without assuming that this support is unique or minimal. Locality is defined independently of whether a model successfully preserves the evidence. For Tracing, locality is determined at response time. Localized and Distributed describe temporal continuity of support, not a distinction between retrieval and reasoning. Together with the response mechanism, these attributes distinguish evidence requirements from question wording and model-specific memory representations.

B.2 COMPLETE TAXONOMY AND EVALUATED SCOPE

The complete taxonomy defines 19 Querying types: five Current, seven Historical, and seven Joint types. Historical Localized types cover event recall, entity-state recall, and causal reasoning; Historical Distributed types cover multi-event recall, event counting, cross-event comparison, and temporal relation reasoning. Joint Localized types cover cross-temporal reference resolution, state-change reasoning, and current–history comparison; Joint Distributed types cover multi-event context grounding, state evolution, longitudinal comparison, and cross-event causal reasoning.

Longitudinal Comparison (LC) and Cross-Event Causal Reasoning (CCR) remain in the taxonomy but are excluded from evaluation because too few high-quality instances were available. The reported evaluation therefore covers 17 Querying types: five Current, seven Historical, and five Joint (CRR, SCR, CHC, MCG, and SER).

The broader framework also includes Current Tracing for event occurrence, action boundaries, and state-change triggering, as well as Joint Tracing extensions. Neither Current nor Joint Tracing contributes to the reported evaluation, which focuses on long-horizon memory. Table 6 gives the complete taxonomy and representative examples, including these unevaluated extensions.

B.3 QUESTION QUALITY CONTROL

Quality control includes automated cross-scale consistency screening of candidate evidence and question-by-question human review of canonical answers and accepted alternative answers. During construction, evidence with unresolved contradictions is excluded rather than used to establish a reference answer. Appendix F describes the annotation process and consistency screening.

Agreement among raw annotation records does not by itself establish video-grounded correctness. The consistency screening and human review described here should be distinguished from a measured final-question error rate or a human–judge agreement study, neither of which is reported.

B.4 INTERPRETING RETENTION AND ANSWER SCORES

For a Historical question q, the causal boundary b_q marks when its supporting evidence has been observed and any question-defined observation window has ended. Its wording, evidence, and reference answer remain fixed across valid delays. The 30-second, 1-, 2-, 4-, and 8-minute evaluations use exactly the same 1,217 question IDs. At 16 and 32 minutes, eligibility requires $b_q + \delta$ not to ex-

ceed the source video's duration, leaving 1,018 and 706 questions, respectively. Selection depends on temporal validity, not model responses or scores.

The first five delays support fixed-cohort retention analysis. The longer delays support comparisons between methods on the eligible set at each delay; differences across these delays also reflect cohort composition. Temporal ineligibility is separate from incomplete baseline execution. Detailed counts appear in Table 9, and subgroup aggregation is specified in Appendix G.

Scope of answer scoring. The text-only judge receives the question, model answer, canonical answer, and accepted alternatives, but no video or evidence annotations. It evaluates agreement with the supplied references, not their independent correctness against the video. Scores include partial credit, rather than counting only fully correct answers. Overall scores macro-average 5 Current, 7 Historical, or 5 Joint subtypes; each subtype pools evaluated questions across videos. Empty, failed, and rejected responses contribute zero. Any score restricted to non-rejected answers must identify that denominator and report the rejection rate. Human review and corrections on a separate batch of Raw Frames outputs from Qwen2.5-VL (Bai et al., 2025b) are used to debug and calibrate the scorer. That batch is excluded from the final result tables, and reported evaluation results are not manually adjusted. The scoring configuration, task-specific criteria, and aggregation rules appear in Appendix G.

B.5 AsyncMemFlow: Execution and Resource Scope

Asynchronous execution. AsyncMemFlow answers from the latest completed memory without requiring an ongoing description or organization call to finish. This removes an update-completion dependency from the response procedure; it does not imply zero resource contention or constant latency. Background processing must keep pace with incoming video blocks to avoid a growing backlog. More capable description models and richer visual inputs can increase memory-construction cost, which is moved off the query-response path rather than eliminated.

Construction execution time versus response latency. The reported 11.25 seconds is the execution time of the measured OASIS (Liang et al., 2026) memory-construction operation for a 32-frame chunk in our configuration. It is not a measured query wait, and it is not assigned to every root merge or every incoming question. Whether construction delays a query depends on scheduling and shared resources. This measurement alone does not characterize OASIS as an architecturally synchronous system.

The response times in Table 2 were measured in isolation, separately from accuracy evaluation and construction execution time. They are not added to the 11.25-second construction measurement and do not establish end-to-end latency while background description, organization, and answering run concurrently.

Retained context and recent coverage. The reported results concern the tested finite video lengths and resource configurations. Full Memory Map input can grow with the retained textual history; the 120-second frame-pool bound is not a bound on textual context or response cost. The visual supplement uses only uncovered observations still retained at query time. The evaluation therefore does not establish gap-free coverage under arbitrarily long construction stalls or bounded response cost over unbounded history. Sampling and routing settings appear in Appendix H.1.

B.6 Interpreting Experimental Comparisons

Complete systems and paired comparisons. The main comparisons evaluate complete systems. In particular, 2B, 4B, and 8B denote AsyncMemFlow's answer models, with Qwen-VL-Max supplying descriptions unless specified otherwise. These results do not isolate the independent effect of asynchronous scheduling or compare systems at a fixed total model-capacity budget. All variants in Table 3 generate descriptions asynchronously. Description-source comparisons fix the answer model and Map setting; Map comparisons fix the answer model and description model; comparisons of answer model capacity fix the description model and organization. The OASIS (Liang et al., 2026) direct-access variant changes both the input presentation and memory-access procedure, as detailed in Appendix H.2.

Clarification with the OASIS authors. We directly consulted the OASIS authors to clarify the settings in their published ablations (Section 4.5, Table 5) (Liang et al., 2026). They confirmed that Flatten Memory retains retrieval tooling while placing all generated memory in the initial context. The detailed distinction between these ablations and our direct-access configuration appears in Appendix H.2.

What the Map ablation measures. Map-off already retains dense descriptions. Map-on adds scene summaries, event headings, and hierarchical grouping, so its ablation evaluates the complete semantic organization module. It does not isolate hierarchy from added semantic content or hold input length fixed. The displayed comparison in Appendix E illustrates these differences. The answer model still locates and reasons over evidence within the supplied context even when explicit retrieval calls are absent.

Retention comparisons. Map-on and Map-off use identical question IDs within each delay. The decrease from 30 seconds to eight minutes is measured on the same 1,217 questions; comparisons between these columns therefore track performance as later observations accumulate. At 16 and 32 minutes, eligibility changes the cohort, so those columns support within-delay comparisons rather than a fixed-cohort forgetting claim. Appendix B.4 and Appendix G give the eligibility and aggregation rules.

Coverage and timing. Incomplete baseline execution is marked in the result tables. Comparisons involving such rows concern their reported coverage; this is distinct from duration-based question eligibility. The isolated response-time measurements and construction-call execution times are separate quantities and are not added together. Their scope is specified in Appendix B.5.

C Full Results on OVO-Bench

We further evaluate AsyncMemFlow on the Perception and Backward tasks of OVO-Bench (Niu et al., 2025). This configuration uses a Qwen3-VL-8B (Bai et al., 2025a) answer model and a separate Qwen3-VL-8B instance as the description model for asynchronous description generation on the same GPU, with no external APIs for description generation or memory organization. Table 5 compares our results with baseline scores reported in OASIS (Liang et al., 2026). AsyncMemFlow achieves 79.94% Perception accuracy and 60.17% Backward accuracy, exceeding the reported OASIS Qwen3-VL-8B results by 1.80 and 2.96 percentage points, respectively.

Baseline scores retain their source-reported settings and are not our reproductions. We observe apparent aggregation inconsistencies in the OASIS source table: averaging its listed Qwen3-VL-8B subtask scores yields approximately 78.06% Perception and 56.47% Backward, rather than the reported 78.14% and 57.21%. We retain the published values in Table 5 for faithful attribution. Even using the local 8B description model instead of Qwen-VL-Max, AsyncMemFlow exceeds both the published averages and those recomputed from the listed subtask scores. These results evaluate the complete AsyncMemFlow pipeline in a fully local configuration on an additional benchmark; the interpretation of system comparisons is discussed in Appendix B.6.

Table 5: Results on OVO-Bench (Niu et al., 2025). Accuracy (%) on Perception and Backward tasks. Baseline scores are taken from OASIS (Liang et al., 2026); AsyncMemFlow scores are from our evaluation. Bold marks the best result among the listed methods in each column.

	Perception							Backward			
Method	OCR	ACR	ATR	STU	FPD	OJR	**Avg.**	EPM	ASI	HLD	**Avg.**
LLaVA-Video-7B	69.13	58.72	68.83	49.44	**74.26**	59.78	62.32	56.23	53.47	51.53	53.74
LLaVA-OneVision-7B	66.44	57.80	73.28	53.37	71.29	61.96	62.04	56.24	55.41	21.51	43.71
Qwen2-VL-7B	40.40	50.46	63.76	47.19	66.34	53.55	50.98	48.71	35.45	45.03	43.73
Qwen2-VL-72B	65.77	60.55	69.83	51.69	69.31	54.35	61.92	52.53	60.81	**57.53**	56.95
Flash-VStream-7B (Zhang et al., 2024)	24.16	29.36	28.45	33.71	25.74	28.80	28.37	39.06	37.16	5.91	27.38
VideoLLM-online-8B (Chen et al., 2024)	8.05	2.83	12.04	14.04	45.54	21.10	22.22	18.80	12.18	17.73	16.24
StreamChat (Xiong et al., 2025)	51.68	47.71	58.62	41.01	61.06	47.83	50.12	50.17	54.05	37.63	47.39
Dispider (Qian et al., 2025)	57.72	49.54	62.07	44.94	61.39	51.63	54.55	48.48	55.41	45.43	36.06
Qwen2.5-VL-7B (Bai et al., 2025b)	76.51	57.80	68.10	46.63	66.34	56.52	60.93	49.83	58.11	44.62	50.24
Qwen3-VL-8B (Bai et al., 2025a)	83.89	58.72	70.69	56.18	69.31	64.13	66.79	51.18	60.81	43.55	51.19
OASIS (Qwen2.5-VL-7B) (Liang et al., 2026)	85.23	72.48	66.38	52.25	67.33	64.67	67.26	51.85	58.78	48.92	52.61
OASIS (Qwen3-VL-8B) (Liang et al., 2026)	91.95	**80.73**	81.03	67.42	67.33	79.89	78.14	**61.95**	60.14	47.31	57.21
AsyncMemFlow (Qwen3-VL-8B)	**93.24**	76.15	**81.90**	**76.97**	69.31	**82.07**	**79.94**	57.58	**67.57**	55.38	**60.17**

Note: Our answer model and description model are separate Qwen3-VL-8B instances running on the same GPU. Description generation runs asynchronously; neither description generation nor memory organization uses external APIs. Baselines retain their source-reported configurations and are not our reproductions.

D QUESTION TYPES AND EXAMPLES

Table 6 presents the question types and representative examples. The complete taxonomy and evaluated scope are explained in Appendix B.2.

Table 6: Question setting and representative examples. Shared classification labels are merged. Joint Tracing examples illustrate the framework and are not part of the reported benchmark.

Mechanism	Evidence locality	Evidence distribution	Question Instances	Example
Querying	Current	Localized	Optical Character Recognition (**OCR**)	Right now, what station name is visible above the train car? [1]
			Entity-Attribute Recognition (**EAR**)	Right now, what color is the wire being soldered? [2]
			Object Recognition (**OR**)	Right now, what is the coach holding? [3]
			Spatial Relation Recognition (**SRR**)	Right now, where is the tall red thermos relative to the white foam blocks on the table? [4]
			Entity Counting (**ECR**)	Right now, how many white foam blocks make up the completed T-shaped structure? [4]
	Historical	Localized	Event Recall (**ER**)	How had the two men ended their argument outside the cafe? [5]
			Entity-State Recall (**ESR**)	Before the bicycle wheel was set aside, in what orientation had the lecturer been holding it during the demonstration? [4]
			Localized Causal Reasoning (**LCR**)	What caused the items on the desk to scatter? [6]
		Distributed	Multi-Event Recall (**MER**)	What vehicle did the patterned-shirt man drive behind before reaching the gas station, and what did he do to the man in the light shirt once he was there? [7]
			Event Counting (**EC**)	Before reaching the Kingston train area, how many distinct pursuit methods had the man used? [1]
			Cross-Event Comparison (**CEC**)	How did the outcome of the kitchen wok fire differ from the outcome of the later barn fire? [7]
			Temporal Relation Reasoning (**TRR**)	The lecturer handled the book __ he lifted the bicycle wheel for the demonstration. [4]
	Joint	Localized	Cross-Temporal Reference Resolution (**CRR**)	Who is hidden inside the sack being carried to the train right now? [1]
			State-Change Reasoning (**SCR**)	How has the freight car's state changed right now compared with the preceding five minutes? [1]
			Current-History Comparison (**CHC**)	How is the engine component being cleaned right now compared with the preceding five minutes? [8]
		Distributed	Multi-Event Context Grounding (**MCG**)	What earlier cleaning actions produced the clean seat and dashboard shown together right now? [9]
			State-Evolution Reasoning (**SER**)	Across the preceding five minutes and right now, how has the pursuit transport evolved? [1]
			Longitudinal Comparison (**LC**)	Compared with the two earlier phone calls, how does the man's current phone-call setting and posture differ? [7]
			Cross-Event Causal Reasoning (**CCR**)	What earlier chain of events produced the burning wreckage and embers visible now? [7]
Tracing	Current	Localized	Event Occurrence Triggering (**EOT**)	Notify me when the welding arc ignites. [10]
			Action Boundary Triggering (**ABT**)	Notify me when the puller begins extracting the shaft. [10]
			State-Change Triggering (**SCT**)	Notify me when the scoreboard changes from 0–2 to 0–3. [11]
	Joint	Localized	Historical Reference Triggering (**HRT**)	Notify me when the black-and-white photo of a smiling man and woman that was found in the white container is shown again. [7]
		Distributed	Multi-Evidence Reference Triggering (**MERT**)	Notify me when the station wagon from the earlier rural trip is seen together with the garbage truck used during the striped-shirted worker's earlier collection route. [7]
			Accumulated-Condition Triggering (**ACT**)	Notify me when all four tour-material milestones have been observed: a tour notebook, a pinned route map, printed flyers with envelope preparation, and a man unfolding and reading tour material. [7]
			Pattern-Completion Triggering (**PCT**)	Notify me when this fire-response pattern completes: straw has ignited, flames have spread over a wooden structure, people have fled, and a firefighter arrives. [7]

Sources: [1] *The General*; [2] *How to Solder—Beginner's Guide*; [3] *Japan vs France, VNL 2024*; [4] *How to Speak*; [5] *Award-Winning Indian Indie Film* (film_tv_02_001); [6] *The Big Trees*; [7] *Mount Joy*; [8] *Forklift Complete Engine Rebuild*; [9] *8 Hour Disaster Car Detail*; [10] *Hydraulic Pump Rebuild—But Nothing Goes to Plan*; [11] *Fulham vs Crystal Palace*.

E MEMORY MAP: ORGANIZATION AND SHARED DESCRIPTIONS

We compare Map-off and Map-on text memory for *Mount Joy* (video `film_tv/02/002`) at query time 4325 s. Both versions contain the same five complete dense descriptions spanning 4020–4320 s in newest-first order. The ongoing 4320–4325 s interval has no completed minute-level description at this snapshot.

For readability, the five shared descriptions are printed once in Section E.3 and referenced below as D1–D5. These are presentation labels only: both memory inputs contain the complete description text. The wording, clip-relative timestamps, and uncertainty of the descriptions are preserved. The original Chinese event heading and scene summaries in Map-on are translated into English for presentation.

E.1 MAP-OFF: FLAT DENSE DESCRIPTIONS

D1	**D2**	**D3**	**D4**	**D5**
4260–4320 s	4200–4260 s	4140–4200 s	4080–4140 s	4020–4080 s

E.2 MAP-ON: ORGANIZED MEMORY

Source interval	**Organization and associated descriptions**
4260–4320 s	**Ungrouped minute: D1.**
4200–4260 s	**Completed scene; larger event ongoing.** The woman lights a cigarette and places the burning cigarette on dry straw, which begins to burn; an orange glow subsequently appears in an upstairs window of the barn. *Dense description: D2.*
Completed event (3180–4200 s). Following and confrontation, relationship dynamics, and the barn gathering.	
4140–4200 s	**Scene within the completed event.** The woman is alone inside the barn; several men have a brief confrontation outside; the segment ends with close-ups of jeans being unbuttoned and an object being taken from a pocket. *Dense description: D3.*
4020–4140 s	**Scene within the completed event.** The red-haired woman leads the man in green into a barn filled with cardboard boxes and shows him a black-and-white photograph of two people; after they kiss, the woman becomes uneasy and the man leaves. *Dense descriptions: D4, D5.*

Older siblings and ancestors are omitted. The 3180–4200 s event retains its full range, although only its 4020–4200 s children are shown. This is an excerpt of text memory; accompanying visual input and the question prompt are not displayed.

Scope of the Map comparison. Map-off and Map-on share the underlying dense descriptions, but Map-on also introduces event headings, scene summaries, and hierarchical grouping. The ablation therefore evaluates the complete semantic consolidation and organization module, not an information- or length-matched change of formatting alone. Its gains should not be attributed exclusively to hierarchy, independently of the added summaries and their semantic content. The displayed excerpt does not establish a fixed input-token budget.

E.3 Shared Dense Descriptions

Each record below appears in both Map-off and Map-on. The identifiers match the structural comparison above.

D1 4260–4320 s

At 0s, a solitary wooden barn is visible at night, one window faintly lit; by 7s, the frame cuts to near-blackness with blurred limbs moving. From 12s–20s, a woman turns her head, then a bearded man lies on leafy ground, blinking slowly into awareness. At 21s–26s, his face fills frame, eyes wide, mouth slightly open; another person's face is blurred behind him. At 29s, a man in yellow shirt and glasses appears startled beside a tank-top-clad figure; they assist or restrain someone. By 34s, figures flee into darkness; trees and distant lights appear. At 38s, the barn erupts in massive flames across its upper structure, fire climbing vertically, embers falling. The blaze intensifies through 55s, obscuring details. At 58s–60s, a backlit figure runs leftward, away from the fully engulfed building.

D2 4200–4260 s

The video begins in near-darkness with a blurry view of fabric, then reveals an orange pill bottle lying among dry straw. A hand enters frame, lighting a cigarette with a metallic lighter; the woman's face is partially visible, eyes lowered. She places the lit cigarette onto the straw, which quickly ignites into a bright, flickering flame. After ˜10 seconds of burning straw, the scene cuts to a static wide shot of a wooden barn against a dusky sky; clouds drift slowly as light fades. Near the end, a small window on the barn's upper gable briefly emits an orange glow before going dark again. No dialogue or identifiable speech occurs; only mechanical sounds of the lighter and crackling fire are discernible.

D3 4140–4200 s

A woman with short brown hair and bangs, wearing a black tank top and necklace, looks upward with a faint smile in a dim interior (0:00–0:04). A man in a green t-shirt walks past stacks of labeled cardboard boxes, turns to face her, and they exchange a silent glance before he walks away (0:05–0:19). She remains standing, holding a red plastic cup, glancing down (0:20–0:25). Cut to exterior night: the man walks toward a house, meets another man, and they briefly grapple or gesture aggressively (0:26–0:33); two others join them. After a black screen (0:34–0:36), the woman is again shown indoors, still holding the cup, then suddenly leans forward and exits frame (0:37–0:43). Blurry motion follows, then tight close-ups show hands unbuttoning denim jeans and pulling something from the front pocket (0:51–0:56).

D4 4080–4140 s

A woman holds up a black-and-white photo of two men and shows it to a man beside her; they lean in together, examining it with focused expressions. She speaks, he responds, and their proximity increases until they kiss intimately for several seconds. Immediately after breaking the kiss, the woman's expression shifts to alarm or distress as she speaks again; the man reacts with furrowed brow and looks away. He then stands and walks toward the background where cardboard boxes are stacked, while she remains seated, lowering her head with a troubled look. The lighting remains low and warm throughout, with shallow depth of field keeping background elements blurred.

D5 4020–4080 s

At 0:00, a woman with red hair and heavy eye makeup looks off-camera with a slight smile in a nighttime outdoor scene with blurred multicolored lights. By 0:03, she stands, holding a red cup, and moves past a seated man in a green shirt near a drained concrete skate bowl; others are present, one crouching, another holding a skateboard. At 0:08, the man rises and follows her as she walks toward a large wooden structure. From 0:13–0:22, they walk side-by-side across a grassy/dirt area toward the barn's double doors. At 0:23, they enter a dim interior filled with stacked cardboard boxes; she gestures rightward while speaking, still holding the cup, and he follows closely. Between 0:27–0:33, they navigate narrow aisles between boxes; camera pans slightly to keep them centered. From 0:34 onward, they pause intermittently—she turns to him, smiles, talks, and gestures again; he listens, nods, and at 0:51 bends down near a pile of boxes while she watches. She continues walking forward, glancing back, until the clip ends at 0:59 with her looking toward the camera, smiling, holding a small white object (possibly a photo or card) in her left hand.

F BENCHMARK CONSTRUCTION DETAILS

F.1 MULTI-SCALE VIDEO DESCRIPTION

Descriptions are generated at three temporal scales: 5-second clips capture fine-grained observations, 1-minute clips capture event sequences and local state changes, and 5-minute clips capture broader event progression and scene transitions. Each scale is described directly from its corresponding video clip; the longer-scale descriptions are not obtained by concatenating or summarizing shorter-scale textual outputs. The resulting records provide complementary source-grounded descriptions for evidence selection and question construction.

All three prompts request JSON outputs with source identifiers, a temporal interval, scale-specific fields, and dense narration. The 5-second schema explicitly separates entities, objects, actions, relations, visible text, and state changes. The 1-minute schema emphasizes major events, local temporal relations, and character/object state changes. The 5-minute schema emphasizes the storyline, important events, longer-term states, relationship changes, and scene transitions. Uncertainty fields are provided at every scale. These are requested annotation fields, rather than guarantees that every generated statement is correct.

Automated consistency screening before question construction. Multi-scale descriptions are candidate evidence records rather than automatically accepted references. During question construction, Codex checks the relevant records for consistency across scales. Evidence with unresolved contradictions, including conflicting entity identities or relationship roles, is excluded from question construction. This separates the generation of raw descriptions from their acceptance as support for a question and reference answer. The screening principle is to reject disputed evidence rather than construct a question whose answer depends on choosing between inconsistent descriptions.

Human review. Canonical answers and accepted alternative answers are manually reviewed for each question. Appendix B.3 discusses the scope of quality control, including its distinction from quantitative error-rate and human–judge agreement studies.

F.2 FIVE-SECOND DESCRIPTION

Prompt.

```
You are a strict video event annotator. You are given one 5-second
↪  video clip.

Task: describe only information that is clearly visible or audible in
↪  this 5-second window. The output will be used for the
↪  current-window perception layer of a video-understanding benchmark.

Hard requirements:
1. Output exactly one valid JSON object. Do not output Markdown,
↪  explanations, or analysis.
2. Write every natural-language value in English. Do not use Chinese.
3. Before producing the JSON, carefully inspect the clip frame by frame
↪  and include all clearly visible details that fit the schema. Do not
↪  omit identifiable objects, text, layout, actions, or audio merely
↪  to be brief.
4. Do not infer psychology, motives, plot, prior or future events, or
↪  anything not supported by the clip.
5. Put uncertain or unreadable content in uncertainty; never force a
↪  guess.
6. List actions, state changes, and dense_narration in temporal order
↪  when applicable.
7. visual_description should summarize the core scene, people/objects,
↪  and visible change in 1-2 sentences.
8. audio_description should record only clearly audible speech or
↪  non-speech sound. If audio is absent or unclear, state that
↪  conservatively.
```

```
9. dense_narration should be detailed, grounded English prose of about
↪  80-150 words, suitable for a blind viewer.
10. Keep arrays bounded: characters <= 5, objects <= 8, actions <= 6,
↪  relations <= 6, visible_text <= 8, state_changes <= 4, uncertainty
↪  <= 4. Use empty arrays when there is no clear content.
11. Keep schema field names exactly as given. Use short factual phrases
↪  outside dense_narration.

Output JSON schema:
{
  "video_id": "",
  "segment_id": "",
  "time_range": {"start": 0, "end": 5},
  "scale": "5s",
  "scene": {"place": "", "environment": "", "camera_view": ""},
  "characters": [{"id": "char_1", "visual_description": "",
  ↪  "role_if_obvious": "", "position": "", "visibility": "clear |
  ↪  partial | unclear"}],
  "objects": [{"id": "obj_1", "name": "", "visual_description": "",
  ↪  "position": "", "visibility": "clear | partial | unclear"}],
  "visual_description": "",
  "actions": [{"time_order": 1, "subject_id": "", "action": "",
  ↪  "object_id": "", "description": "", "confidence": "high | medium
  ↪  | low"}],
  "relations": [{"type": "spatial | interaction | possession |
  ↪  attention", "subject_id": "", "object_or_character_id": "",
  ↪  "description": ""}],
  "visible_text": [{"text": "", "location": "", "confidence": "high |
  ↪  medium | low"}],
  "audio_description": {
    "speech_presence": "none | speech | unclear",
    "speaker_type": "lecturer | interviewer | interviewee | commentator
    ↪  | participant | multiple | unknown | none",
    "spoken_content_summary": "",
    "key_terms_or_phrases": [],
    "non_speech_audio": [],
    "audio_uncertainty": [{"item": "", "reason": ""}]
  },
  "state_changes": [{"entity_id": "", "before": "", "after": "",
  ↪  "description": ""}],
  "uncertainty": [{"item": "", "reason": ""}],
  "short_caption": "",
  "dense_narration": ""
}
```

F.3 One-minute Description

Prompt (English translation). The prompt emphasizes event-level continuity rather than second-by-second enumeration. Numerical timestamps are restricted to the clip-level `time_range`; relative ordering is expressed in words.

```
You are an online video recent-event annotator. You are given a
↪  1-minute video clip.

Task: Model the event sequence, local context, and clearly audible
↪  information within this minute for recent-memory evaluation. Do not
↪  enumerate 5-second details second by second or write a long-term
↪  storyline summary.

Hard requirements:
1. Output only valid JSON, without Markdown or explanations.
2. Event-level summarization is allowed, but must be grounded in
↪  visible or audible content.
3. Record uncertain content in uncertainty.
```

```
4. Order major_events, state changes, and local temporal relations
↪  chronologically where possible.
5. visual_description should summarize the core visuals, scenes,
↪  people/objects, camera changes, and visible state changes in 2-5
↪  sentences. Include only clearly visible content.
6. audio_description should record only clearly audible information. Do
↪  not invent transcripts. Record unclear speech, uncertain speaker
↪  identity, or unconfirmed content in uncertainty.
7. dense_narration should provide a grounded, information-dense
↪  description of the continuous visuals and key audio within this
↪  minute, without psychological speculation.
8. Treat the input as a complete 1-minute clip. Do not output specific
↪  within-clip seconds, timestamps, or second-level intervals in any
↪  field; do not use expressions such as "at X seconds" or "from X to
↪  Y seconds".
9. time_range should contain only the input clip's start and end. Do
↪  not output numerical times outside this field. Use non-numerical
↪  ordering expressions such as "subsequently", "afterward", "during",
↪  and "finally".
10. All natural-language content in the final JSON must be English,
↪  including events, state changes, relations, descriptions,
↪  uncertainty, audio fields, and dense_narration. Do not output
↪  Chinese or mixed-language text. Use exactly the English schema
↪  field names.

Output JSON schema:
{
  "video_id": "",
  "chunk_1min_id": "",
  "time_range": {"start": 0, "end": 60},
  "scale": "1min",
  "major_events": [],
  "character_state_changes": [],
  "object_state_changes": [],
  "local_temporal_relations": [],
  "causal_hints": [],
  "visual_description": "",
  "audio_description": {
    "speech_presence": "none | speech | unclear",
    "speaker_type": "lecturer | interviewer | interviewee | commentator
    ↪  | participant | multiple | unknown | none",
    "spoken_content_summary": "",
    "key_terms_or_phrases": [],
    "non_speech_audio": [],
    "audio_uncertainty": [{"item": "", "reason": ""}]
  },
  "uncertainty": [],
  "dense_narration": ""
}
```

F.4 FIVE-MINUTE DESCRIPTION

Prompt. The mixed-language phrase in the original prompt is rendered as "scene-level narrative" below.

```
You are an online video long-range context annotator. You are given a
↪  5-minute video clip.

Task: Model the long-term storyline, scene-level narrative, and
↪  important changes for global context evaluation. Do NOT enumerate
↪  5s-level fine-grained details.

Hard requirements:
1. Output ONLY valid JSON. No Markdown, no explanation.
```

```
2. Long-term event summarization is allowed, but MUST be grounded in
↪  what is visible/audible in the video.
3. Write uncertain content into the uncertainty field.
4. main_storyline, important_events, scene_transitions should be
↪  ordered chronologically when possible.
5. visual_description should summarize the core visible scene,
↪  people/objects, camera/view changes, and visible state changes in
↪  2-5 sentences. Only write clearly visible content.
6. audio_description should capture only clearly audible information.
↪  Do NOT invent a transcript. If speech, speaker identity, or content
↪  is unclear, write it into uncertainty.
7. dense_narration is a compressed, grounded description of the
↪  5-minute visual content and important audible content. Do NOT write
↪  invisible motivations or psychological assertions.
8. ALL text output MUST be in English.

Output JSON schema:
{
  "video_id": "",
  "chunk_5min_id": "",
  "time_range": {"start": 0, "end": 300},
  "scale": "5min",
  "main_storyline": [],
  "main_characters": [],
  "important_events": [],
  "long_term_states": [],
  "relationship_changes": [],
  "scene_transitions": [],
  "visual_description": "",
  "audio_description": {
    "speech_presence": "none | speech | unclear",
    "speaker_type": "lecturer | interviewer | interviewee | commentator
    ↪  | participant | multiple | unknown | none",
    "spoken_content_summary": "",
    "key_terms_or_phrases": [],
    "non_speech_audio": [],
    "audio_uncertainty": [{"item": "", "reason": ""}]
  },
  "uncertainty": [],
  "dense_narration": ""
}
```

G Evaluation Details

G.1 Answer Format and Scoring Configuration

EOV-Bench uses open-ended short answers without candidate options. The shared answer-format requirements request the shortest sufficient answer phrase: a blank is filled with the answer alone, a count is expressed as an integer, and OCR responses contain the recognized text. Answers to change and sequence questions must retain the required direction and order. Explanations, question restatements, and multiple candidate answers are discouraged. The base configuration requests `REJECT` when evidence is insufficient, rather than guessing from question wording or prior knowledge.

Text-only judge. The supplied scoring configuration uses Qwen3.5-9B in BF16 with thinking disabled, temperature 0, and a maximum of 32 output tokens. The judge receives the question instance category, answer type, question, model answer, canonical answer, and all accepted alternatives. It receives no video, frames, or video descriptions. Current, Historical, and Joint answers use separate task-specific scoring prompts. Each automated judgment is a single JSON object containing a score in $\{0, 0.5, 1\}$, without an explanation. A valid output has the following form:

```
{"score": 0.5}
```

Scorer calibration and final evaluation. We debug and calibrate the scorer using human review and corrections on a separate batch of Raw Frames outputs from Qwen2.5-VL (Bai et al., 2025b). This batch is used for scorer development and is excluded from the final result tables. Human intervention is confined to this calibration stage; the reported evaluation results are not manually adjusted. This process is separate from the question-by-question review of canonical and accepted alternative answers during benchmark construction.

Scoring procedure. Empty answers, errors, and `REJECT` receive 0. A normalized exact match to the canonical answer or an accepted alternative receives 1. Remaining answers are submitted to the judge. If its output is malformed, one retry adds a format constraint; a second failure is reported as an error rather than accepted as a valid score. Cached judgments distinguish the question, reference answers, and model response, with separate caches for Current, Historical, and Joint questions. Comparisons require a consistent scoring-prompt and cache version.

G.2 Task-specific Scoring Criteria

Current. Full credit requires a correct or semantically equivalent answer. Partial credit requires substantive correct information with limited imprecision or omission. Entity counts must be exact; numerically close counts do not receive partial credit. For spatial relations, an accepted partial relation may receive 0.5, whereas an opposite direction or a different relation dimension receives 0. For OCR, minor spelling errors or limited omissions may receive 0.5, but incorrect essential text does not. Object recognition permits partial credit for an informative superordinate category, not a different object in the same broad class. Attribute recognition permits partial credit for a nearby shade, a coarser description, or a partially correct set of attributes.

Historical. The reference answers must support the historical fact actually expressed by the model. Neither common-sense plausibility nor lexical overlap suffices. Complete recovery of the requested facts, relations, comparisons, or order receives 1; a substantive but meaningfully incomplete answer receives 0.5. Incorrect events or relations, broad guesses, and refusals receive 0.

Joint. Scoring tests the requested cross-temporal relation rather than isolated facts. For state changes and comparisons, full credit requires the relevant earlier and later states and the correct direction. Both endpoints must be present for partial credit: one may be slightly imprecise if the other is correct and the direction is preserved. A single endpoint, a reversed change, or a substantively incorrect endpoint receives 0. This two-endpoint requirement is not imposed on reference-resolution questions that ask for a single entity. For multi-stage processes, an equivalent ordered account re-

ceives 1; retaining the core process while omitting a secondary stage may receive 0.5; an isolated stage receives 0.

G.3 Score Aggregation

For an evaluated question set of size N, its question-level micro-average is

$$\text{Score} = \frac{100}{N} \sum_{i=1}^{N} s_i, \qquad s_i \in \{0, 0.5, 1\}. \tag{3}$$

This normalized score includes partial credit; it is not the percentage of answers receiving 1. Each subtype pools all its evaluated questions across videos, rather than averaging per-video scores. Overall scores in Tables 1–3 use

$$\text{Avg.} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{Score}_c, \tag{4}$$

where Score_c is the micro-average within subtype c, and $\mathcal{C}$ contains the 5 Current, 7 Historical, or 5 Joint evaluated subtypes. Subtypes receive equal weight despite unequal question counts; unevaluated types are excluded. Historical macro scores are computed separately at each delay; Table 1 averages the first five delays. Empty, failed, and rejected responses remain in their subtype denominators and contribute zero. Any score restricted to non-rejected answers must state that denominator and report the rejection rate separately.

Delay-dependent cohorts and Map gains. Let $\mathcal{Q}_0$ denote the 1,217-question cohort used at $\delta \in \{30, 60, 120, 240, 480\}$ seconds. For the two longer delays,

$$\mathcal{Q}_\delta = \{q \in \mathcal{Q}_0 : b_q + \delta \le T_{v(q)}\}, \qquad \delta \in \{960, 1920\}, \tag{5}$$

where $T_{v(q)}$ is the duration of the source video. For the first five delays, $\mathcal{Q}_\delta = \mathcal{Q}_0$. Selection uses temporal validity only, not model answers, rejection, or correctness. Map-on and Map-off use identical question IDs at each delay. Rejections and evaluated failures remain in the specified scoring denominator and contribute zero.

Table 4 reports differences between the corresponding Historical macro scores in Table 3, holding the answer model, description model, and query delay fixed:

$$\Delta_{\text{Map}}(\delta) = \text{Avg.}_{\text{on}}(\delta) - \text{Avg.}_{\text{off}}(\delta). \tag{6}$$

Each macro score weights the seven Historical subtypes equally; reported gains subtract the displayed two-decimal scores. Fixed-cohort degradation is therefore analyzed only from 30 seconds to eight minutes. The longer-delay columns permit within-delay comparisons on their eligible subsets, not a pure cross-delay forgetting comparison against the full initial cohort.

H Implementation and Evaluation Configurations

This appendix specifies AsyncMemFlow's sampling and routing, the OASIS comparison, and shared evaluation settings. Baseline configurations are summarized in Table 8.

H.1 AsyncMemFlow Sampling and Routing

The Frame Processor samples at 1 FPS into a bounded 120-second frame pool. Each completed minute forms an immutable 60-frame block B_k, from which 30 chronological frames are sampled at 0.5 FPS for description:

$$D_k = \mathcal{G}_\phi\big(\mathrm{Sample}_{0.5}(B_k)\big). \tag{7}$$

We use Qwen-VL-Max as $\mathcal{G}_\phi$ unless otherwise specified. Blocks are described in temporal order, and completed descriptions pass to the external memory organizer.

The Query Processor uses the answer model itself as a text-only router. Current questions receive the latest five causal 1-FPS frames. Historical/Joint questions receive the latest completed Memory Map and the retained uncovered raw-frame suffix sampled at 0.2 FPS. The router achieves overall accuracies of 99.49%, 98.03%, and 90.46% with Qwen3-VL-8B, 4B, and 2B (Bai et al., 2025a), respectively. All evidence is frozen at question arrival. Execution, timing, and resource scope are discussed in Appendix B.5.

H.2 Direct Access to OASIS Memory

Our configuration. The *OASIS (our version)* row in Table 2 reuses the descriptions and node summaries already generated by OASIS (Liang et al., 2026), including root summaries. These records are assembled into a flattened textual context following AsyncMemFlow's direct-context presentation and supplied with recent visual observations. Only memory available before question arrival is used. Query-dependent retrieval is disabled, and the answer model generates its answer directly from this context. The description source remains OASIS; this variant does not replace its video descriptions with Qwen-VL-Max outputs.

Flattening here concerns how existing descriptions and node content are exposed to the answer model. The input includes both detailed descriptions and higher-level node summaries. It should therefore not be characterized as concatenating only leaf descriptions, or as answering from only a bounded set of root summaries. The comparison changes the answer context and memory-access procedure together.

Relation to the published ablations. Table 7 distinguishes our configuration from the three ablations in OASIS, Section 4.5 and Table 5 (Liang et al., 2026). Those published ablations use OVO-Bench (Niu et al., 2025); our comparison in Table 2 uses EOV-Bench.

Results and interpretation. Direct access raises Historical scores by 8.61, 8.00, 9.64, 7.39, 5.23, 9.74, and 7.06 points at the seven respective delays in Table 2. On the fixed cohort through eight minutes, the five-delay mean increases from 28.88 to 36.66. This establishes a consistent benefit of the tested input and access configuration on EOV-Bench. It does not identify retrieval omissions as the sole cause, isolate a formatting-only effect, or imply that retrieval is universally harmful. The published OVO-Bench ablations and our EOV-Bench comparison differ in both configuration and evaluation setting.

H.3 Baseline and Shared Evaluation Settings

Table 8 identifies released checkpoints and answer model substitutions separately. The designation *OASIS (our version)* refers to the direct-access configuration above, not an answer model substitution.

Shared inference settings. Models are instructed to answer within five words or output `REJECT` when evidence or confidence is insufficient. Unless otherwise specified, answer generation is deterministic and limited to 64 new tokens; OASIS uses 128. Memory is reset between videos. Benchmark reference answers and evidence annotations are excluded from inference. Experiments use

Table 7: Distinguishing the published OASIS (Liang et al., 2026) ablations from our direct-access configuration.

Variant	Memory construction and available text	Query-time access
Flatten Memory	Removes the root-count limit; generated nodes remain roots, without budget-triggered hierarchical merging.	All generated memory is already in the initial context; retrieval tooling remains available.
w/o RAG	Retains the Event Forest and its root summaries.	Answers from recent windows and root summaries, with the retrieval stage removed.
Naïve RAG	Retains structured event memory.	Retrieves with the original question instead of a generated retrieval intent.
Our version	Reuses OASIS descriptions and node summaries, including root summaries, as a flattened textual context.	Supplies that context directly with recent visual observations; query-dependent retrieval is disabled.

The clarification obtained from the OASIS authors is documented in Appendix B.6. Tool availability does not establish that every query invokes retrieval. The w/o RAG row follows the published definition of removing the retrieval stage.

NVIDIA A100 80GB GPUs. Accuracy evaluation and isolated response-latency measurement are conducted separately.

Table 8: Implementation details and inference configurations of the baseline methods.

Method	Model checkpoints	FPS	Memory and inference configuration
Raw Frames	Qwen3-VL-8B-Instruct (Bai et al., 2025a) / LLaVA-OneVision-Qwen2-7B	1	Latest 5 frames; no persistent historical memory.
ReKV (Di et al., 2025)	LLaVA-OneVision-Qwen2-7B	1	Local KV budget: 15,000; retrieval top-k: 64; retrieval chunk size: 1; encoding chunk size: 8 frames.
HERMES (Zhang et al., 2026a)	LLaVA-OneVision-Qwen2-7B	0.5	KV budget: 6,000; encoding chunk size: 16 frames; QA write-back disabled for Historical and Joint queries.
HERMES–Qwen3-VL[a] (Zhang et al., 2026a)	Qwen3-VL-8B-Instruct	0.5	KV budget: 6,000; encoding chunk size: 16 frames; maximum image pixels: 262,144; QA write-back disabled.
StreamForest (Zeng et al., 2025)	StreamForest-Qwen2-7B	1	Short-term window sizes: [1, 18], with token allocations [729, 128]; long-term budget: 5,120 tokens, with 64 tokens per frame; event-splitting window: 8.
Flash-VStream–Qwen3-VL[a] (Zhang et al., 2025)	Qwen3-VL-8B-Instruct	1	Incremental temporal compression and spatial-detail selection; temporal/spatial length parameters: 120/60, corresponding to 60/30 two-frame units.
VST (Guan et al., 2026)	VST-7B (Qwen2.5-VL-based)	1	Visual chunk budget: 8,192 tokens; FIFO textual memory: 12,000 characters; separate, identical frozen model instances for memory generation and answering.
Dispider (Qian et al., 2025)	Qwen2-1.5B Perception–Decision + Qwen2-7B Reaction	1	16-frame clips; cached completed-clip representations; native decision, retrieval, and reaction components; `SILENT` retained for non-triggered responses.
OASIS (Liang et al., 2026)	Qwen3-VL-8B-Instruct + Qwen3-Embedding-0.6B	1	Current-window capacity: 16 frames; short-term buffer: 32 frames; event-node capacity: 16 frames; 256 tokens per frame; maximum 4 root nodes; up to 2 retrieved events.
OASIS (our version)[b] (Liang et al., 2026)	Same as OASIS	1	Flattened descriptions and node summaries, including root summaries, with recent visual context; retrieval disabled; direct answer generation.

[a] Training-free answer model substitutions, not the methods' original released checkpoints. StreamForest, VST, and Dispider retain their released trained weights.

[b] Reuses OASIS-generated descriptions and node summaries available before each query; see Appendix H.2 for the input configuration and comparison with the published ablations.

Historical question eligibility. The eligible counts below are determined independently of model responses. They are per-delay counts of question IDs, not numbers of successful answers or sums across repeated evaluations. The first five delays share exactly the same IDs; later subsets remove questions whose scheduled query would exceed video duration. Temporal ineligibility is distinct from incomplete baseline execution, which is marked separately in the result tables.

Table 9: Historical question eligibility by query delay. The first row applies separately to each of its five delays.

Delay	Localized	Distributed	Total	Localized share
30 s, 1, 2, 4, 8 min (each)	849	368	1,217	69.76%
16 min	721	297	1,018	70.83%
32 min	518	188	706	73.37%