MemInStream: Memorizing or Merely Recognizing? Rethinking Streaming Video Memory Evaluation

Zhenqi He[1] Dunyuan Xu[2] Yang Bo[3,†] Kien T.Pham[1] Yanghao Wang[1] Shuhan Li[3] Xintong Guo[3]
Yuyang He[3] Fang Liu[4] HuanKang Guan[3] Jingyan Chen[3] Rui Liu[3] Long Chen[1,*]
[1]HKUST [2]CUHK [3]Huawei HK [4]CityU
zheci@connect.ust.hk, longchen@ust.hk

Abstract

Streaming video understanding aims to continuously process incoming visual observations, making the retention and retrieval of relevant information increasingly important as the video history grows. Although existing benchmarks include memory-related questions, question-answering accuracy alone does not reliably measure memory capabilities. Incorrect predictions may arise from failures in visual understanding rather than forgetting, whereas correct predictions may rely on linguistic priors or recent observations without requiring historical memory. To address these ambiguities, we introduce MEMINSTREAM, a benchmark dedicated to evaluating memory capabilities in streaming video understanding. We develop an evidence-controlled validation protocol that associates each question with temporally localized evidence intervals. The protocol validates evidence solvability and memory necessity by retaining questions that models can answer from the annotated evidence but cannot resolve using recent observations, post-evidence observations, or textual input alone. This helps disentangle memory limitations from understanding failures while reducing memory-independent shortcuts. MEMINSTREAM provides a systematic evaluation of memory capabilities through four major categories, covering Episodic Retrieval, State Tracking & Updating, Multi-Evidence Integration, and Contextual Memory across diverse video scenarios. Overall, MEMINSTREAM provides a comprehensive and reliable benchmark for systematically evaluating memory capabilities in streaming video understanding.

Project Page Code Dataset

1 Introduction

Streaming video understanding aims to continuously perceive and interpret incoming visual information, enabling models to track evolving events and respond to queries in real time [3, 4, 5, 6]. Unlike offline video understanding, where the complete recording is available for analysis, streaming video understanding requires models to process observations sequentially as the video unfolds [7]. As visual history accumulates, models should retain relevant past information under limited memory and computational resources [8, 9]. Such historical information remains essential for interpreting ongoing events and answering subsequent queries, even when the relevant observations

[†]Project Leader.
[*]Corresponding Author.

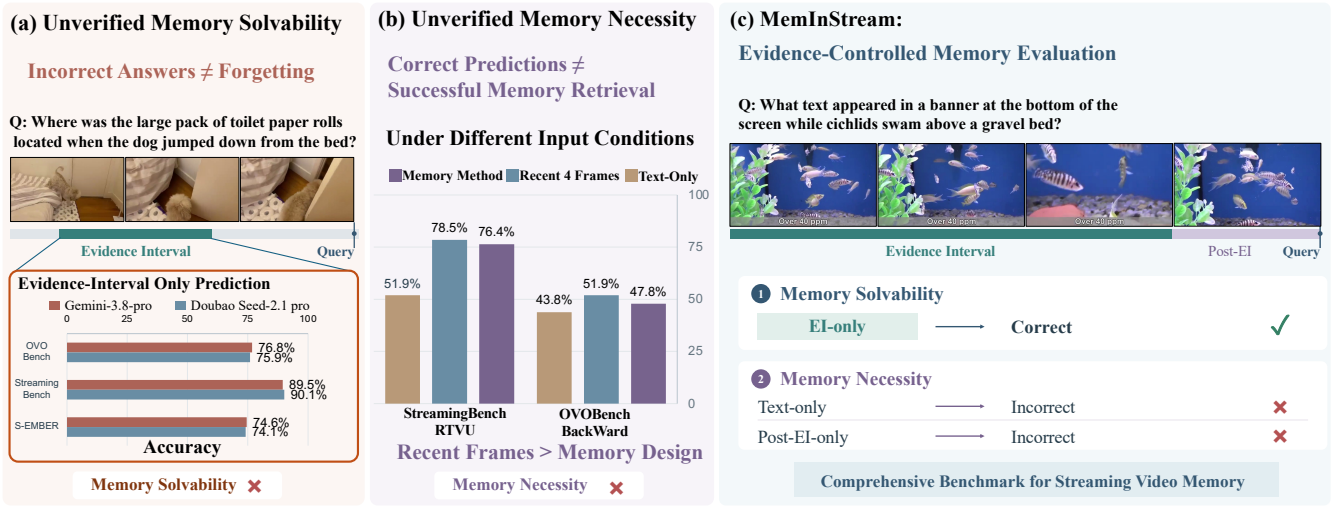


Figure 1: (a) **Unverified Memory Solvability**. Models may fail to answer questions even when the required evidence intervals are directly provided. Results are reported on the backward tracing tasks of OVO-Bench, the real-time video understanding tasks of StreamingBench, and S-EMBER. (b) **Unverified Memory Necessity**. We compare a memory-based method (FluxMem [1]) with SimpleStream [2], which uses only the four most recent frames, and a text-only baseline. Competitive performance with limited or no visual history suggests that correct predictions do not necessarily indicate successful retrieval of historical information. (c) Our evidence-controlled validation protocol ensures both memory solvability and memory necessity, providing a comprehensive benchmark dedicated to evaluating memory capabilities in streaming video understanding.

are no longer available in recent frames [10]. This is particularly important for assistants expected to recall earlier activities and experiences [11]. For example, a user may ask where an object was previously placed or how its state has changed over time. Effective streaming video understanding therefore requires not only understanding incoming observations, but also preserving, updating, and retrieving relevant information from the past, making memory a fundamental capability.

A growing number of benchmarks have been introduced to evaluate streaming video understanding [12, 13, 14]. Early efforts, such as StreamingBench [13], assess real-time perception, omni-source understanding, and contextual understanding. Subsequent benchmarks, including OVBench [15] and OVO-Bench [12], further emphasize temporal awareness, incorporating memory retrieval and backward tracing alongside real-time perception and future-oriented reasoning. More recently, TeleEgo [16] and PhoStream [17] extend streaming evaluation to realistic egocentric and mobile scenarios, covering memory, cross-memory reasoning, and real-time audio-visual understanding. Dedicated memory benchmarks have also emerged: EgoStream [18] investigates episodic memory across different temporal horizons, S-EMBER [14] focuses on temporally grounded episodic retrieval, and EgoProbe [19] evaluates long-range retention and event-specific recall. Despite these advances, performance on memory-related questions does not necessarily distinguish genuine memorization from mere recognition.

More specifically, two fundamental ambiguities arise when evaluating streaming memory through question-answering accuracy. **(1) Unverified Memory Solvability.** An incorrect answer does not necessarily indicate a memory failure, as models may struggle to perceive, interpret, or reason over the relevant visual information even when the necessary evidence is directly available. As shown in Fig. 1(a), recent video understanding models fail to answer nearly a quarter of the questions even when provided with the required evidence intervals. Without verifying that models can answer a question when its supporting evidence is explicitly available, errors under streaming conditions cannot be reliably attributed to failures in retaining or retrieving historical information. **(2) Unverified Memory Necessity.** Conversely, a correct answer does not necessarily demonstrate successful memory retrieval, as linguistic priors, answer-option cues, or recent visual observations may suffice without accessing historical evidence. As illustrated in Fig. 1(b), a text-only baseline achieves competitive accuracy on the evaluated benchmarks, while a method using only the four most recent frames [2] even outperforms memory-based method [1], which explicitly leverages historical memory. These results suggest that existing evaluations may permit shortcuts that bypass historical memory.

Table 1: Comparison of streaming video benchmarks in terms of memory coverage and evidence-controlled validation. ✓indicates explicit coverage, while △denotes partial coverage or related validation mechanisms.

Benchmark	Memory Coverage				Evidence-Controlled Validation			
	Episodic Retrieval	State Tracking & Updating	Multi-Evidence Integration	Contextual Memory	Evidence Solvability	Evidence Necessity	Evidence Interval	Types of Videos
StreamingBench [ICASSP2026]	✓	△	△	✓	✗	△[a]	✗	General
OVBench [CVPR2025]	✓	✓	✓	✗	✗	✗	△	General
OVO-Bench [CVPR2025]	✓	△	✓	✗	✗	✗	△	Mixed
TeleEgo [arXiv2025]	✓	✓	✓	✗	✗	✗	✓	Egocentric
PhoStream [ICML2026]	✓	✗	✓	△	△[b]	✗	✗	Mixed
EgoStream [arXiv2026]	✓	✗	✓	✗	✗	△[c]	✓	Egocentric
S-EMBER [arXiv2026]	✓	△	✓	✗	△[d]	△[e]	✓	Egocentric
MemInStream [Ours]	✓	✓	✓	✓	✓	✓	✓	Mixed

Validation notes: △ denotes a related or partial control. [a]Humans remove questions answerable without video; [b]answers are verified using video prefixes; [c]text-only baseline evaluation; [d]ground-truth-guided sampling retains other history; [e]an LLM removes text-solvable questions.

Together, these observations highlight the need to establish not only whether historical evidence is sufficient to answer a question, but also whether it is necessary.

To address these ambiguities, we introduce MEMINSTREAM, a benchmark designed to systematically evaluate memory capabilities in streaming video understanding. At its core, we develop an evidence-controlled validation protocol that assesses whether historical evidence is both sufficient and necessary to answer each question. Specifically, each question is annotated with a query timestamp and evidence intervals containing the required historical information. As illustrated in Fig. 1(c), we first assess *memory solvability* by directly providing the annotated evidence intervals to determine whether models can answer the question when the relevant information is available. We then assess *memory necessity* through text-only, recent-only, and post-evidence conditions, examining whether questions can be answered without the required historical evidence. These complementary checks help exclude questions confounded by visual understanding difficulties or shortcuts that bypass historical memory. Furthermore, to evaluate diverse aspects of streaming memory, MEMINSTREAM comprises 17 fine-grained tasks organized into four categories: *Episodic Retrieval* (recalling historical facts and events), *State Tracking & Updating* (maintaining and updating evolving states), *Multi-Evidence Integration* (integrating information across historical observations), and *Contextual Interactive Memory*. Together, these designs enable a more systematic and targeted evaluation of memory capabilities in streaming video understanding.

Our contributions can be summarized as follows: **(i) A comprehensive streaming video memory benchmark.** We introduce MEMINSTREAM, a benchmark comprising four major memory categories and 17 fine-grained tasks across diverse scenarios, systematically assessing models' ability to retain, update, and retrieve historical information under streaming conditions. **(ii) Evidence-controlled memory evaluation.** We develop an evidence-controlled validation protocol that examines both the solvability and necessity of historical evidence, reducing the confounding effects of visual understanding limitations and memory-independent shortcuts in memory evaluation. **(iii) Systematic evaluation of streaming memory capabilities.** We evaluate representative streaming video models across diverse memory tasks and scenarios, examining their performance under different memory demands and highlighting areas for further improvement.

2 Related Work

2.1 Streaming Video Understanding

Streaming video understanding requires models to process continuously arriving observations and answer queries as events unfold, using only information available at the response time [3, 4]. Unlike offline video understanding, which permits analysis of a complete recording, streaming systems must preserve relevant information before future queries are known. Memory is therefore essential for recalling past events, tracking changing states, and connecting temporally separated observations under bounded computational resources. Existing benchmarks examine

these capabilities from complementary perspectives: StreamingBench [13] evaluates real-time visual, multisource, and contextual understanding. OVBench [15] and OVO-Bench [12] organize tasks around past, current, and future events. TeleEgo [16] extends evaluation to prolonged egocentric experiences and cross-memory reasoning, and PhoStream [17] assesses audiovisual understanding and response timing in mobile scenarios. However, QA accuracy can conflate memory with evidence understanding and memory-independent answer cues. A failure may reflect difficulty interpreting the relevant evidence even when it is available, while a correct answer may rely on recent observations rather than retained history. The competitive performance of SimpleStream using only recent frames underscores the need to verify whether historical memory is required [2].

Existing streaming memory methods differ primarily in how they retain and access historical information. HERMES [20], ReKV [21], and InfiniPot-V [22] reuse, retrieve, or compress KV caches, preserving previously computed attention states for subsequent queries. FluxMem [1], VideoChat-Online [15], and StreamForest [23] maintain compact visual representations through token compression or hierarchical memory banks, balancing historical coverage against storage and computation. Response-G1 [24] instead represents historical context through explicit scene graphs. Complementing these memory representations, training-based approaches such as StreamingVLM [8], AURA [25], JoyAI-VL-Interaction [26], and MOSS-VL-Realtime [27] adapt models to continuous observation and streaming responses through dedicated training. Despite differences in representation and training, these methods share the need to preserve useful information for future queries. MemInStream provides a common evaluation setting with evidence-controlled validation of memory solvability and necessity, distinguishing it from existing benchmarks, as summarized in Table 1

2.2 Video Memory

Video memory benchmarks increasingly examine which historical details remain accessible, for how long, and under what temporal conditions. LongVideoBench [28] evaluates retrieval and reasoning over extended video and subtitle inputs, while S-EMBER [14] grounds episodic recall in localized visual evidence. EgoStream [18] diagnoses recall across cognitive dimensions and temporal horizons using answer validity windows, and EgoProbe [19] tests both long-range retention and event-specific recall under interference from similar experiences. APM-Bench [29] extends evaluation to persistent memory across interrupted sessions, while CapMem [30] examines textual captions as reusable episodic memory. These benchmarks broaden the temporal scope and specificity of memory evaluation, but answering accuracy still needs to be interpreted in relation to evidence availability and the model's ability to understand that evidence.

Video memory methods preserve historical information through compact visual representations or explicit, retrievable records. MovieChat [31] and MA-LMM [32] compress and accumulate visual information in memory banks, extending access to past observations beyond the immediate input context. M3-Agent [33] builds entity-centric episodic and semantic memories for subsequent retrieval and reasoning, while EgoLife [11] connects retrieval from extended egocentric recordings to everyday assistance through EgoRAG. Caption-based approaches studied in CapMem [30] provide an alternative by storing textual descriptions that can be reused across queries. Although these representations differ in storage cost and retained detail, assessing their effectiveness requires questions that both depend on historical evidence and remain answerable when that evidence is available. MemInStream makes these requirements explicit through joint memory solvability and necessity checks, providing evidence-controlled evaluation across four memory categories and 17 tasks.

3 MemInStream Benchmark

3.1 Task Formulation

Streaming video memory evaluation assesses a model's ability to retain and utilize historical information while processing visual observations sequentially. Let $V_{1:t} = (v_1, \ldots, v_t)$ denote the observations received up to time t. The model incrementally updates its memory state and answers a question q at query timestamp t_q:

$$M_t = \mathcal{U}(M_{t-1}, v_t), \qquad \hat{a}_q = \mathcal{A}(M_{t_q}, q, \mathcal{O}_q), \tag{1}$$

where $\mathcal{U}$ and $\mathcal{A}$ denote the memory update and answering functions, respectively, and $\mathcal{O}_q$ is the set of candidate answers. The memory state M_t is architecture-independent and may comprise cached visual tokens, compressed

representations, textual summaries, or external memory. Models must answer using their maintained memory, without unrestricted access to the original video history or future observations.
To identify the historical information required for each question, we annotate K evidence intervals:

$$\mathcal{E}_q = \{[s_i, e_i]\}_{i=1}^{K}, \qquad 0 \le s_i < e_i < t_q, \tag{2}$$

which localize the relevant visual evidence preceding the query. For contextual memory tasks, we additionally annotate timestamped dialogue events $\mathcal{H}_q = \{(t_j, d_j)\}_{j=1}^{J}$, where $t_1 < \cdots < t_J < t_q$, together with a subset $\mathcal{R}_q \subseteq \{1, \ldots, J\}$ identifying the necessary dialogue turns. These dialogue events are incorporated into the model's memory sequentially, rather than provided as a complete transcript at query time. Ground-truth answers, task labels, and evidence annotations are withheld from the streaming model and used only for evaluation.

3.2 Taxonomy of Streaming Video Memory

We organize streaming video memory into four categories and 17 fine-grained tasks, covering historical retrieval, state maintenance, temporal integration, and interaction-dependent memory. Tasks are defined by their primary memory requirements rather than question phrasing, and categories are not necessarily mutually exclusive.
Episodic Retrieval. This category concerns retrieving information from specific past observations without continuous state tracking or integration across independent events. Its six tasks cover previously observed entities (*Entity / Object Recall*), visual properties (*Attribute Recall*), past actions or events (*Action / Event Recall*), locations or spatial arrangements at a particular moment (*Spatial / Location Recall*), event times or durations (*Temporal Fact Recall*), and whether an entity or event was observed within a given period (*Occurrence Verification*).
State Tracking & Updating. This category requires maintaining and revising information as new observations arrive. *Latest-State Retrieval* recovers an entity's most recently observed state; *State Transition Tracking* follows its changes over time; *Identity & State Persistence* maintains entity identities and associated states across separated observations; and *Correction / Revision Memory* updates information corrected or superseded by later evidence.
Multi-Evidence Integration. This category combines historical observations when no single evidence interval suffices. *Temporal Ordering* establishes chronological relationships among events; *Sequential Action Integration* combines actions to infer an overall process or outcome; *Counting* aggregates occurrences across time; *Cross-Time Comparison* identifies similarities or differences between observations; and *Multi-hop Memory Binding* connects information through intermediate entities, events, or relationships.
Contextual Memory. This category preserves information established through prior interactions for use in subsequent visual observations and queries. *Entity–Identity Binding* retains user-defined associations between visual entities and names, identities, or roles. *Conversational Context Memory* extends this requirement to references, constraints, and contextual information established across dialogue turns. Both require combining relevant conversational history with visual evidence, as in recalling Jerry's actions after an earlier identity assignment or resolving a reference to a previously specified object.

4 Benchmark Construction

In this section, we describe the construction of MEMINSTREAM, including candidate collection, question generation, evidence annotation, and quality validation. Our pipeline aims to provide comprehensive coverage of streaming memory capabilities while reducing the influence of understanding limitations and memory-independent shortcuts.

4.1 Overall Pipeline

As illustrated in Fig. 2, we construct MEMINSTREAM from two complementary sources. First, we collect memory-related questions from existing streaming video benchmarks. Second, we retrieve additional videos and generate new questions to improve the coverage of memory capabilities and real-world scenarios. Candidates from both sources are standardized according to the task formulation in Sec. 3.1, with each question assigned a query timestamp and the historical evidence intervals required to answer it. All candidates subsequently undergo the same evidence-controlled validation principles, followed by human verification to obtain the final benchmark.

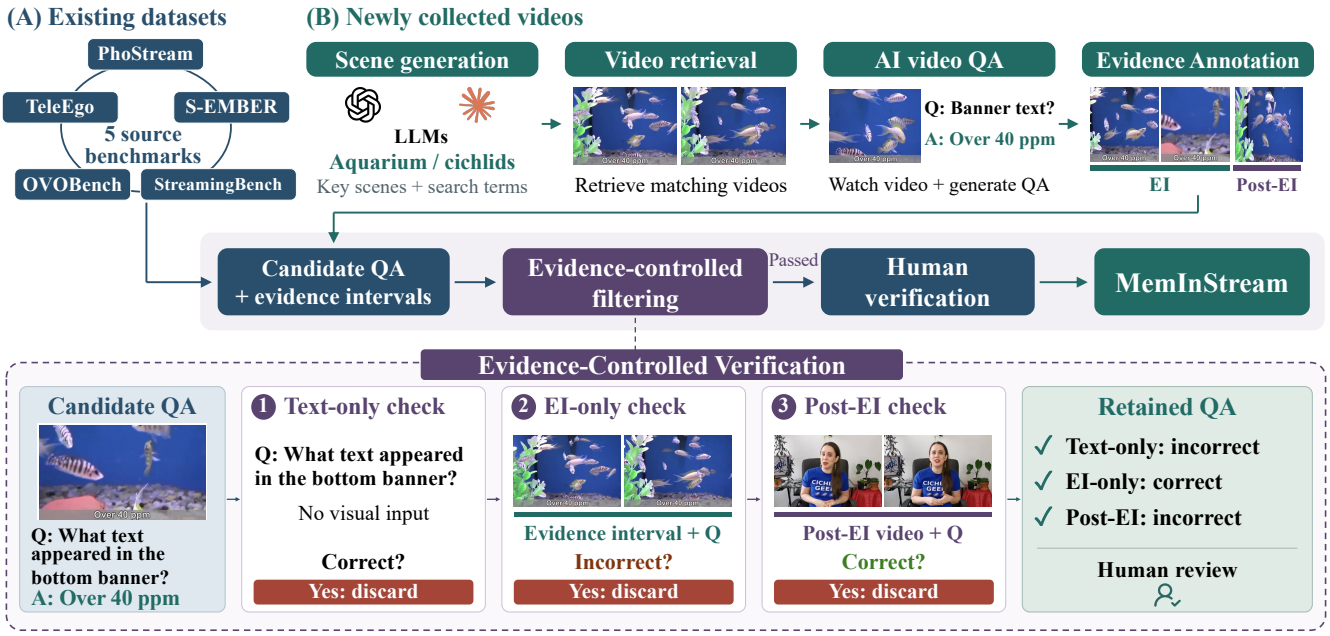


Figure 2: Overview of the MEMINSTREAM construction pipeline.

4.2 Question and Evidence Annotation

Existing Benchmark Collection. We collect candidate questions from StreamingBench [13], OVO-Bench [12], TeleEgo [16], PhoStream [17], and S-EMBER [14]. We focus on questions that require recalling or utilizing historical observations and organize them according to our unified memory taxonomy. Rather than directly incorporating these questions, we treat them as candidates for further evidence annotation and validation.

New Video Collection and Question Generation. To complement existing benchmarks, we employ large language models to identify representative scenarios and generate corresponding video search terms for different memory tasks. We then retrieve relevant videos from the web and use video-language models to examine their content and generate candidate question-answer pairs. The newly generated questions target diverse memory requirements, including historical retrieval, state updating, multi-evidence integration, and contextual memory. All generated questions undergo the same subsequent annotation and validation procedures as those collected from existing benchmarks.

Question and Answer Format. To balance objective evaluation with answer selection complexity, we adopt a unified format combining single-answer and multiple-answer questions. Each question contains a set of candidate options, with one or more correct answers. The number of correct options is not disclosed to the evaluated models, which are instructed to select all correct options. A prediction is considered correct only when the selected option set exactly matches the ground truth.

Evidence Annotation. For each candidate, we employ video-language models to examine the video and identify the historical observations necessary to answer the question. Specifically, we annotate the query timestamp t_q and a set of evidence intervals $\mathcal{E}_q$ containing the required visual information. Depending on the memory task, the evidence may consist of a single historical episode or multiple temporally separated observations. For questions requiring multi-evidence integration, all relevant intervals must be identified to ensure that the annotated evidence jointly supports the answer. For contextual memory questions, we additionally construct time-ordered dialogue histories and identify the historical turns necessary to resolve each query. These questions are designed to require both relevant dialogue context and visual evidence, such that neither source alone is sufficient to determine the answer. All annotations are subsequently examined during evidence-controlled validation and human verification.

4.3 Evidence-Controlled Validation

To disentangle memory from visual understanding and memory-independent shortcuts, we introduce an evidence-controlled validation protocol using two strong video understanding models, Gemini-3.8-Flash and Seed-2.1-Pro, to assess two complementary properties: *memory solvability* and *memory necessity*.

Memory Solvability. We assess whether a question can be answered when the required historical evidence is directly available. In the *evidence-only* setting, both models receive the annotated evidence intervals, question, and answer options, without needing to retain information across the full video stream. We retain only questions correctly answered by both models, reducing the influence of visual understanding and reasoning difficulties.

Memory Necessity. We assess whether a question can be answered without the required historical evidence under two conditions. The *text-only* setting provides only the question and answer options to identify linguistic or option-level shortcuts. The *post-EI* setting provides the video segment from the end of the final evidence interval to the query timestamp, identifying answers recoverable from subsequent observations. We retain only questions that both models answer incorrectly under each condition, reducing the likelihood that questions can be solved without historical memory.

Contextual Memory Validation. For contextual memory questions, we additionally verify their joint dependence on historical dialogue and visual evidence using the same two validation models. We consider three conditions: (1) *Relevant dialogue + EI*, where both models must answer correctly given the necessary dialogue turns and evidence intervals; (2) *Irrelevant dialogue + EI*, where both models must answer incorrectly when the necessary dialogue turns are replaced with irrelevant historical turns; and (3) *Relevant dialogue + Post-EI*, where both models must answer incorrectly when the evidence intervals are replaced with post-EI observations. We retain only questions satisfying all three conditions, reducing the likelihood that either historical dialogue or visual evidence can be bypassed.

Automatic Filtering and Human Verification. During automatic validation, video frames are uniformly sampled at 4 fps. We employ Gemini-3.8-Flash and Seed-2.1-Pro as validation models, retaining only questions that satisfy the agreement-based solvability and necessity criteria described above. For questions with multiple correct options, correctness requires an exact match between the predicted and ground-truth option sets. Following automatic filtering, human reviewers examine the retained questions against the original videos to verify answer correctness, temporal alignment of query timestamps and evidence intervals, and the relevance and completeness of the annotated evidence. Questions with ambiguous answers or inconsistent annotations are excluded.

4.4 Benchmark Statistics

Fig. 3 provides an overview of the taxonomy and temporal characteristics of MEMINSTREAM. The benchmark contains 3,799 questions spanning four major memory categories and 17 fine-grained tasks, as shown in Fig. 3(a). These questions are associated with 8,217 annotated evidence intervals, averaging 2.16 intervals per question. Each interval lasts 9.98 seconds on average, with a median duration of 6 seconds. Moreover, 56.7% of questions involve multiple evidence intervals, highlighting the need to retain and integrate information from several historical segments.

To characterize the temporal demands of memory retrieval, we measure the distances from both the latest and earliest evidence to the query:

$$d_q^{\text{latest}} = t_q - \max_{[s,e]\in\mathcal{E}_q} e, \qquad d_q^{\text{earliest}} = t_q - \min_{[s,e]\in\mathcal{E}_q} s.$$

As illustrated in Fig. 3(c–d), the median distances are 36 and 96 seconds, respectively. Notably, 34.6% of questions have their latest evidence at least 60 seconds before the query, while 66.0% require evidence beginning at least 60 seconds earlier, highlighting substantial temporal dependencies. Moreover, 56.7% of questions involve multiple annotated evidence intervals, requiring models to retain and potentially integrate information across different moments. Together, these statistics demonstrate that MEMINSTREAM evaluates not only the retrieval of isolated historical observations but also the maintenance and integration of information over extended video streams.

5 Experiment

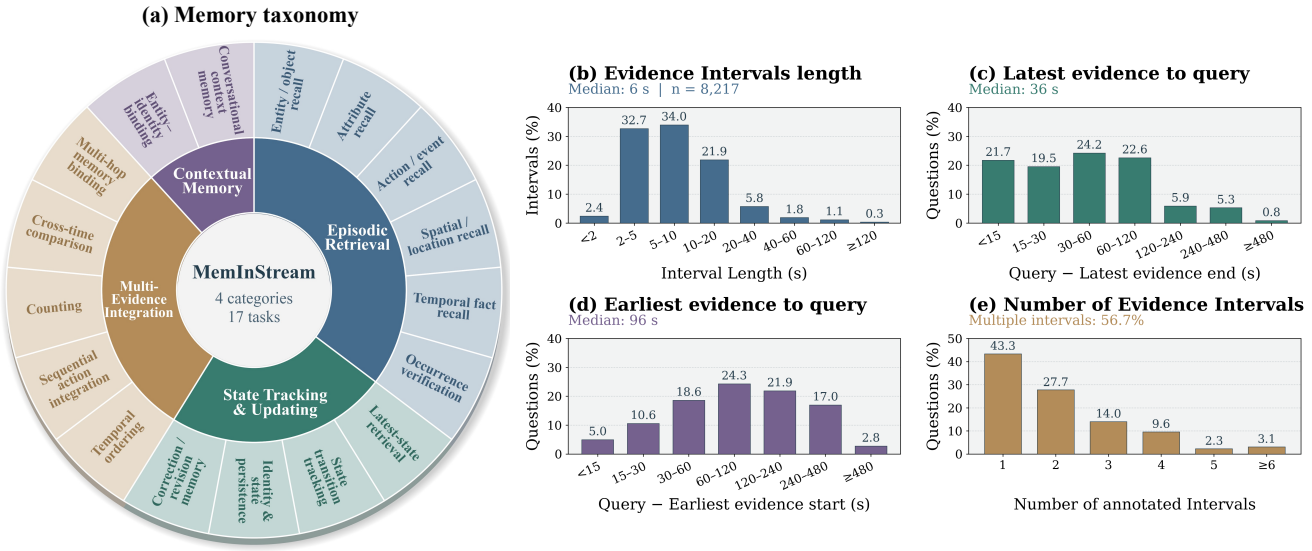


Figure 3: (a) Hierarchical taxonomy comprising four major memory categories and 17 fine-grained tasks. (b) Distribution of evidence interval lengths. (c–d) Temporal distances from the latest and earliest evidence intervals to the query, respectively. (e) Distribution of the number of evidence intervals per question.

5.1 Evaluation Protocol

Answer Format and Metrics. MemInStream adopts a unified evaluation format combining single-answer and multiple-answer questions, where models must select all correct options without knowing the number of correct answers. Compared with conventional single-choice QA, this format requires models to identify the complete set of valid answers rather than selecting a single plausible option, reducing reliance on guessing. Meanwhile, the structured answer space enables objective and reproducible evaluation, avoiding the ambiguity of assessing open-ended responses. We adopt strict exact-match accuracy, where a prediction is correct only if the predicted answer set exactly matches the ground-truth set:

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\hat{\mathcal{A}}_i = \mathcal{A}_i\right), \tag{3}$$

where $\hat{\mathcal{A}}_i$ and $\mathcal{A}_i$ denote the predicted and ground-truth answer sets, respectively. We report overall accuracy and macro-averaged accuracy across memory tasks to account for differences in task sizes.

5.2 Implementation Details

Offline and proprietary models. We evaluate three open-source offline MLLMs, Qwen2.5-VL-7B [34], Qwen3-VL-8B [35], and InternVL3.5-8B [36], together with two proprietary models, Gemini-3.8-Flash and Doubao Seed-2.1-pro. Following the recent-frame protocol of SimpleStream [2], we evaluate these models with recent-frame budgets of $N \in \{4, 30\}$ at 1 fps, using frames sampled from the N seconds immediately preceding the query timestamp. Each model receives the available sampled frames in chronological order, together with the question and answer options, without maintaining persistent visual memory across queries.

Online models. We divide open-source online MLLMs into training-free and training-based methods. The training-free group includes FluxMem [1], Response-G1 [24], and HERMES [20]. FluxMem and Response-G1 are evaluated with both Qwen2.5-VL-7B and Qwen3-VL-8B backbones, while HERMES uses Qwen2.5-VL-7B. The training-based group comprises TimeChat-online [6], Flash-VStream [5], StreamForest [23], AURA [25], JoyAI-VL-Interaction [26], and MOSS-VL-Realtime [27]. Online models process video observations chronologically and answer at the query timestamp using the context retained by their respective mechanisms.

5.3 Main Results

Overall Performance. Table 2 presents the evaluation results across four memory categories and 17 fine-grained tasks. Performance varies substantially across models and evaluation settings, highlighting the challenges of streaming video memory. Among online models, StreamForest achieves the highest average accuracy across all four memory categories, with particularly strong performance on State Tracking & Updating (37.06%) and Multi-Evidence Integration (37.34%), outperforming even the evaluated proprietary models in these categories. Nevertheless, performance remains relatively limited across many tasks, suggesting that reliably retaining and utilizing historical information remains challenging for current streaming video models.

Challenges Across Memory Tasks. Different memory capabilities exhibit distinct performance characteristics. In particular, Multi-hop Memory Binding remains challenging even for strong models, with Gemini-3.8-Flash (30 frames) and StreamForest achieving only 13.07% and 23.86%, respectively. A similar difficulty emerges in Conversational Context Memory, where several online models perform substantially worse than on Entity–Identity Binding. For example, StreamForest achieves 52.54% on Entity–Identity Binding but only 18.28% on Conversational Context Memory. These results suggest that integrating temporally separated evidence and maintaining dialogue-dependent context present additional challenges beyond retrieving individual historical facts.

Impact of Visual Input and Model Design. Increasing the number of input frames generally improves episodic retrieval for offline models, with Gemini-3.8-Flash improving from 26.38% to 44.96% and Doubao Seed-2.1-Pro from 20.93% to 38.63% when increasing from 4 to 30 frames. However, such improvements are not consistent across all memory categories, indicating that additional visual observations alone do not guarantee better memory performance. Moreover, online methods exhibit considerable performance differences across backbone models. For example, replacing Qwen2.5-VL with Qwen3-VL substantially improves both FluxMem and Response-G1 across all four categories. These observations highlight the importance of both visual history coverage and underlying model capabilities when evaluating streaming memory methods.

5.4 Analysis

Evidence Coverage and Memory Performance. Fig. 4 analyzes the relationship between recent-frame accuracy and the coverage of annotated evidence intervals. Across both frame budgets, all five models achieve substantially higher accuracy when the required evidence is fully covered by the input window. For instance, Gemini-3.8-Flash achieves 90.6% and 86.2% accuracy under full overlap with 4 and 30 frames, respectively, compared with only 21.8% and 19.8% without overlap. A similar trend is observed for Seed-2.1-pro, whose accuracy increases from 19.2% to 76.3% under the 30-frame setting. These results highlight the strong dependence of recent-frame performance on evidence availability. However, correct predictions without evidence overlap may arise from alternative visual cues, linguistic priors, or chance, rather than genuine historical memory. Conversely, even complete evidence coverage does not guarantee correct predictions, highlighting that access to the required evidence alone is insufficient and that effectively identifying and retrieving relevant information is equally important.

Temporal Distance and Memory Performance. Fig. 4(c-d) examines accuracy as a function of the temporal distance from the latest evidence end and the earliest evidence start to the query. The methods exhibit different patterns across temporal ranges. JoyAI achieves 44.0% accuracy when the latest evidence is within 10 seconds of the query, but its accuracy drops from 36.8% at 10–30 seconds to 17.9% at 30–60 seconds. A similar decline appears with increasing distance to the earliest evidence, with accuracy remaining around 19–20% beyond 60 seconds. In contrast, StreamForest achieves the highest accuracy across all latest-evidence distance bins beyond 10 seconds, reaching 44.1% at distances of at least 300 seconds, compared with 32.2% for Response-G1 and 33.9% for MOSS. FluxMem falls to 5.1% in this range. However, these differences must be interpreted alongside the evaluated input configurations: JoyAI observes approximately 30 seconds of recent video, while FluxMem receives at most 256 seconds, leaving the annotated evidence outside their input windows at longer distances. Moreover, the distance bins contain different questions and dataset compositions. These results therefore reflect the combined effects of evidence availability, input budgets, and question difficulty, rather than an isolated measure of memory decay.

Table 2: Comparison on MEMINSTREAM. Task scores are exact-match accuracy (%); each Avg. is the macro-average over tasks within that category. Frames/FPS reports input frame counts for proprietary and open-source offline models, and sampling rates (fps) for online models. Offline models use only the N most recent sampled frames before the query. †Response-G1 dynamically adjusts the fps between 0.25 and 1. ‡Up to 60 recent frames per query. *TimeChat-Online uses 4–316 input frames with adaptive sampling at 0.2, 0.5, or 1 fps. §Up to 768 input frames. $^{-}$/$^{+}$JoyAI-VL-Interaction without/with memory, respectively.

			MemInStream																			
			Episodic Retrieval							State Tracking & Updating					Multi-Evidence Integration						Contextual Memory		
Method	Size	Frames FPS	EOR	AR	AER	SLR	TFR	OV	Avg.	LSR	STT	ISP	CRM	Avg.	TO	SAI	CNT	CTC	MMB	Avg.	EIB	CCM	Avg.
Proprietary Models																							
Gemini-3.8-Flash	–	4	34.5	21.7	33.3	24.6	20.0	24.1	26.4	24.2	24.8	20.6	25.4	23.8	21.5	28.1	15.6	28.4	11.9	21.1	37.3	36.6	36.9
Gemini-3.8-Flash	–	30	52.5	50.5	44.5	37.9	30.0	54.4	45.0	53.1	23.5	20.6	34.5	32.9	25.2	35.2	27.9	35.1	13.1	27.3	33.9	35.5	34.7
Doubao Seed-2.1-pro	–	4	35.4	21.4	28.4	16.4	16.3	7.6	20.9	24.2	27.6	25.6	15.5	23.2	17.8	30.5	8.9	24.6	8.5	18.1	37.3	26.9	32.1
Doubao Seed-2.1-pro	–	30	50.7	48.5	39.6	32.5	26.3	34.2	38.6	47.9	31.0	22.5	26.1	31.9	20.7	33.3	25.3	29.9	16.5	25.1	39.0	33.3	36.2
Open-Source Offline MLLMs																							
Qwen2.5-VL [34]	7B	4	7.2	6.4	7.3	4.6	3.7	1.3	5.1	3.1	6.2	7.5	4.9	5.4	7.0	9.5	2.3	6.2	8.5	6.7	23.7	6.5	15.1
Qwen2.5-VL [34]	7B	30	9.4	10.7	5.7	4.6	2.2	10.1	7.1	4.6	5.5	6.9	4.9	5.5	4.1	11.4	1.3	6.2	6.3	5.8	22.0	7.5	14.8
Qwen3-VL [35]	8B	4	20.2	11.0	13.3	5.7	3.3	7.6	10.2	9.3	16.6	5.6	7.8	9.8	13.7	17.6	4.7	16.6	4.6	11.4	32.2	8.6	20.4
Qwen3-VL [35]	8B	30	25.6	29.1	16.4	11.4	9.4	12.7	17.4	18.0	11.0	5.6	11.3	11.5	10.0	19.1	7.6	16.6	2.8	11.2	37.3	2.2	19.7
InternVL3.5 [36]	8B	4	27.8	14.1	15.1	12.5	5.5	3.8	13.1	22.7	28.3	13.8	20.4	21.3	18.5	19.1	8.6	23.2	13.1	16.5	49.2	18.3	33.7
InternVL3.5 [36]	8B	30	29.2	34.8	16.2	12.1	6.5	19.0	19.6	31.4	19.3	13.8	18.3	20.7	17.4	20.5	11.2	20.9	8.5	15.7	39.0	16.1	27.6
Open-Source Online MLLMs(Training-Free)																							
FluxMem (Qwen2.5) [1]	7B	1	11.7	16.1	6.5	5.7	2.9	13.9	9.5	2.6	2.8	3.1	6.3	3.7	1.9	6.7	1.0	5.7	3.4	3.7	17.0	4.3	10.6
FluxMem (Qwen3) [1]	8B	1	35.9	51.2	19.8	20.4	17.8	24.1	28.2	21.7	20.7	18.1	40.1	25.2	24.1	25.2	10.9	24.6	8.0	18.6	47.5	4.3	25.9
Response-G1 (Qwen2.5) [24]	7B	0.25–1†	6.7	7.4	5.0	3.6	4.1	6.3	5.5	2.1	3.5	1.3	8.5	3.8	4.8	4.8	2.6	4.3	2.8	3.9	6.8	0.0	3.4
Response-G1 (Qwen3) [24]	8B	0.25–1†	36.8	65.2	22.4	24.3	18.6	27.9	32.5	23.7	24.8	28.1	54.9	32.9	28.9	25.2	13.5	26.1	6.8	20.1	47.5	8.6	28.0
HERMES (Qwen2.5) [20]	7B	0.5	10.8	16.1	5.7	2.1	2.0	10.1	7.8	1.6	3.5	3.1	14.8	5.7	3.0	5.2	0.8	4.7	3.4	3.4	17.0	3.2	10.1
Open-Source Online MLLMs (Training-Based)																							
Flash-VStream (Qwen2)	7B	1	30.9	31.1	28.4	20.4	30.2	17.7	26.5	34.0	21.4	29.4	59.2	36.0	39.6	30.5	21.4	36.0	16.5	28.8	50.9	19.4	35.1
StreamForest (Qwen2)	7B	1	52.5	67.6	38.5	37.1	27.4	29.1	42.0	33.5	29.7	40.0	45.1	37.1	48.5	39.5	33.1	41.7	23.9	37.3	52.5	18.3	35.4
TimeChat-Online	7B	0.2–1*	17.9	36.8	7.6	8.6	12.2	19.0	17.0	7.2	14.5	6.3	54.2	20.5	20.0	21.0	18.0	16.1	5.7	16.1	37.3	0.0	18.6
AURA [25]	8B	2‡	40.8	44.2	27.1	24.6	25.7	20.3	30.4	38.1	22.1	26.3	24.7	27.8	22.2	25.2	22.7	27.0	15.9	22.6	42.4	18.3	30.3
JoyAI-VL-Interaction^{-} [26]	8B	2‡	42.2	46.8	27.6	22.9	27.4	24.1	31.8	37.1	29.7	20.0	27.5	28.6	18.9	30.0	20.8	28.0	13.6	22.3	44.1	9.7	26.9
JoyAI-VL-Interaction^{+} [26]	8B	2‡	53.4	65.6	32.0	32.1	31.2	31.6	41.0	43.3	35.9	28.8	67.6	43.9	39.6	37.1	30.7	37.4	18.2	32.6	50.8	8.6	29.7
MOSS-VL-Realtime [27]	11B	2‡	26.9	31.1	14.8	11.8	11.6	12.7	18.2	18.6	17.9	10.6	9.9	14.2	9.6	9.5	14.6	16.6	4.0	10.9	18.6	9.7	14.2
MOSS-VL-Realtime [27]	11B	1§	31.8	45.8	16.2	21.1	14.3	22.8	25.3	19.1	26.2	26.3	37.3	27.2	29.6	32.4	29.7	30.3	10.8	26.6	42.4	6.5	24.4

Task abbreviations. EOR: Entity / Object Recall; AR: Attribute Recall; AER: Action / Event Recall; SLR: Spatial / Location Recall; TFR: Temporal Fact Recall; OV: Occurrence Verification; LSR: Latest-State Retrieval; STT: State Transition Tracking; ISP: Identity & State Persistence; CRM: Correction / Revision Memory; TO: Temporal Ordering; SAI: Sequential Action Integration; CNT: Counting; CTC: Cross-Time Comparison; MMB: Multi-hop Memory Binding; EIB: Entity–Identity Binding; CCM: Conversational Context Memory.

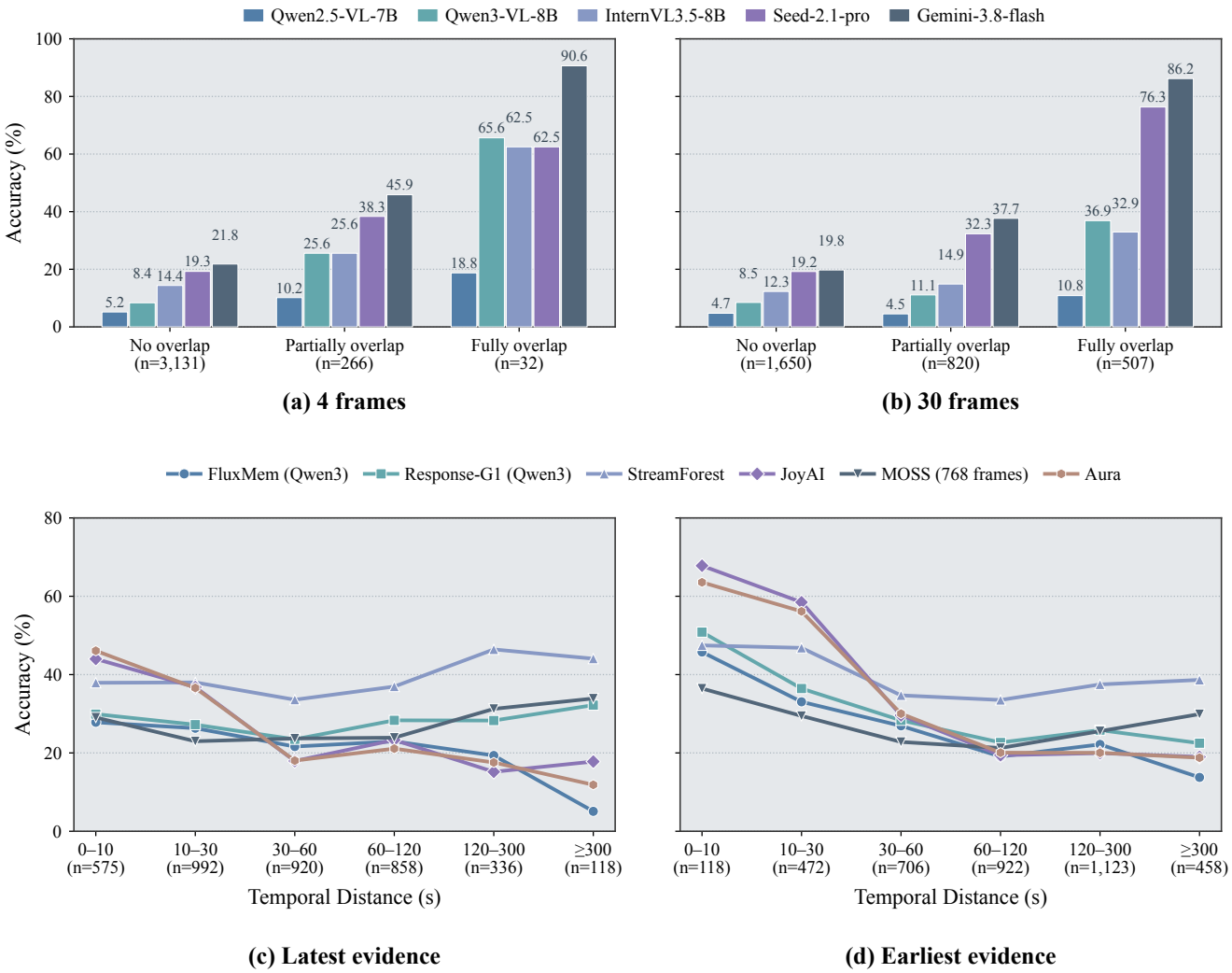


Figure 4: Accuracy under different evidence overlap levels and temporal distances. (a, b) Accuracy for 4-frame and 30-frame inputs, respectively. No overlap, partial overlap, and full overlap indicate that the input window covers none, some, or all of the annotated evidence intervals, respectively. (c, d) Accuracy versus temporal distance to the query, measured from the latest evidence end and the earliest evidence start, respectively.

6 Conclusion

We introduce MEMINSTREAM, a comprehensive benchmark for evaluating memory capabilities in streaming video understanding. To distinguish genuine memory performance from visual understanding limitations and memory-independent shortcuts, we develop an evidence-controlled validation protocol that examines both memory solvability and necessity. The resulting benchmark comprises 3,799 questions, spanning four memory categories and 17 fine-grained tasks. Extensive evaluations reveal substantial challenges for existing streaming video models, particularly in integrating historical evidence and maintaining contextual information. Further analysis shows that access to relevant historical evidence plays an important role in answering memory-related questions. By providing diverse memory tasks with evidence-controlled validation, MEMINSTREAM offers a systematic foundation for evaluating and advancing memory capabilities in streaming video understanding.

References

[1] Yiweng Xie, Bo He, Junke Wang, Xiangyu Zheng, Ziyi Ye, and Zuxuan Wu. FluxMem: Adaptive Hierarchical Memory for Streaming Video Understanding. In *CVPR*, 2026.

[2] Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. A Simple Baseline for Streaming Video Understanding. *arXiv preprint arXiv:2604.02317*, 2026.

[3] 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 *CVPR*, 2024.

[4] 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 *CVPR*, 2025.

[5] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-VStream: Efficient Real-Time Understanding for Long Video Streams. In *ICCV*, 2025.

[6] Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, Lingpeng Kong, Qi Liu, Yuanxing Zhang, and Xu Sun. TimeChat-Online: 80% Visual Tokens are Naturally Redundant in Streaming Videos. In *ACM MM*, 2025.

[7] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming Dense Video Captioning. In *CVPR*, 2024.

[8] Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. StreamingVLM: Real-Time Understanding for Infinite Video Streams. In *ICLR*, 2026.

[9] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming Long Video Understanding with Large Language Models. In *NeurIPS*, 2024.

[10] Xiangyu Zeng, Yuandong Yang, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Qingyi Si, Dingyu Yao, Changlian Ma, Haoran Chen, Xinyu Chen, Yansong Shi, Junhao Zhou, Yifei Li, Jun Zhang, Chuanyu Qin, Chenxu Yang, Xinlei Yu, Kun Ouyang, Yuchen Shao, Qianshan Wei, Changhai Zhou, Jun Gao, Jiaqi Wang, and Limin Wang. OneStreamer: Unifying Perception, Memory, and Proactive Response in Streaming Video Interaction. *arXiv preprint arXiv:2610.01762*, 2026.

[11] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, and Ziwei Liu. EgoLife: Towards Egocentric Life Assistant. In *CVPR*, 2025.

[12] 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 *CVPR*, 2025.

[13] 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.

[14] Xiaodong Wang, Xuanyi Zhao, Pedro Rodriguez, Devendra Singh Sachan, Barlas Oguz, Seungwhan Moon, Shang-Wen Li, Gargi Ghosh, Xin Dong, and Wen-Tau Yih. S-EMBER: A Large-Scale Benchmark for Streaming Egocentric Memory Retrieval. *arXiv preprint arXiv:2607.02689*, 2026.

[15] 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 *CVPR*, 2025.

[16] Jiaqi Yan, Ruilong Ren, Jingren Liu, Shuning Xu, Ling Wang, Yiheng Wang, Xinlin Zhong, Yun Wang, Long Zhang, Xiangyu Chen, Changzhi Sun, Jixiang Luo, Dell Zhang, Hao Sun, Chi Zhang, and Xuelong Li. TeleEgo: Benchmarking Egocentric AI Assistants in the Wild. *arXiv preprint arXiv:2510.23981*, 2025.

[17] Xudong Lu, Huankang Guan, Yang Bo, Jinpeng Chen, Xintong Guo, Shuhan Li, Fang Liu, Peiwen Sun, Xueying Li, Wei Zhang, Xue Yang, Rui Liu, and Hongsheng Li. PhoStream: Benchmarking Real-World Streaming for Omnimodal Assistants in Mobile Scenarios. In *ICML*, 2026.

[18] Rosario Forte, Giuseppe Lando, and Antonino Furnari. EgoStream: A Diagnostic Benchmark for Streaming Episodic Memory in Egocentric Vision. In *NeurIPS*, 2026.

[19] Haokai Pang, Yuqian Fu, Yifei Wang, Yujiao Shen, Tianwen Qian, Mohamed Elhoseiny, and Hongbo Fu. EgoProbe: Benchmarking Long-Range and Exact Memory in Streaming Egocentric Videos. *arXiv preprint arXiv:2610.17128*, 2026.

[20] Haowei Zhang, Shudong Yang, Jinlan Fu, See-Kiong Ng, and Xipeng Qiu. HERMES: KV Cache as Hierarchical Memory for Efficient Streaming Video Understanding. In *ACL*, 2026.

[21] Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming Video Question-Answering with In-context Video KV-Cache Retrieval. In *ICLR*, 2025.

[22] Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. InfiniPot-V: Memory-Constrained KV Cache Compression for Streaming Video Understanding. In *NeurIPS*, 2025.

[23] 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 *NeurIPS*, 2025.

[24] Ke Ma, Jiaqi Tang, Bin Guo, Xueting Han, Ruonan Xu, Qingfeng He, Ziheng Wang, Xu Wang, Qifeng Chen, Zhiwen Yu, and Yunhao Liu. Response-G1: Explicit Scene Graph Modeling for Proactive Streaming Video Understanding. In *ACL*, 2026.

[25] Xudong Lu, Yang Bo, Jinpeng Chen, Shuhan Li, Xintong Guo, Huankang Guan, Fang Liu, Dunyuan Xu, Peiwen Sun, Heyang Sun, Rui Liu, and Hongsheng Li. AURA: Always-On Understanding and Real-Time Assistance via Video Streams. *arXiv preprint arXiv:2604.04184*, 2026.

[26] Dingyu Yao, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Xiangyu Zeng, Yifei Li, Haowen Hou, Zheming Liang, Congcong Wang, Kaiwen Tuo, Jun Zhang, Yuhan Zhu, Yuhang Cao, Shenglong Ye, Shuai Xie, Shuhuan Gu, Haoyang Huang, Qingyi Si, Nan Duan, and Jiaqi Wang. JoyAI-VL-Interaction: Real-Time Vision-Language Interaction Intelligence. *arXiv preprint arXiv:2606.14777*, 2026.

[27] Pengyu Wang, Chenkun Tan, Shaojun Zhou, Qirui Zhou, Yanxin Chen, Xingyang He, Huazheng Zeng, Jijun Cheng, Chenghao Wang, Xiaomeng Qian, Pengfei Wang, Zhan Huang, Shanqing Gao, Wei Huang, Longjun Cao, Wu Ran, Jie Liu, Changtai Zhu, Hongkai Wang, Yixian Tian, Chenghao Liu, Zhen Ye, Xinghao Wang, Botian Jiang, Guoguo Feng, Zhaoye Fei, Ruixiao Li, Mingshu Chen, Yang Gao, Qinyuan Cheng, Shimin Li, and Xipeng Qiu. MOSS-VL Technical Report. *arXiv preprint arXiv:2608.15045*, 2026.

[28] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. In *NeurIPS*, 2024.

[29] Jianguo Huang, Jinming Liu, Qiyao Wang, Liang Xu, Jianhang Li, Zhimian Wen, Mingda Li, Shule Lu, Zhicheng Wang, Yuhan Guo, Xin Jin, and Wenjun Zeng. APM-Bench: Benchmarking Cross-session Persistent Memory for Egocentric Streaming Video Assistants. *arXiv preprint arXiv:2609.37559*, 2026.

[30] Dingli Liang, Yiqiao Xie, Yukai Huang, Zhaokai Wang, Weitong Cai, Guangwen Feng, Jifei Song, Zhensong Zhang, and Hang Zhang. CapMem: A Benchmark for Caption-Based Episodic Memory in Egocentric Video. In *EMNLP (Industry Track)*, 2026.

[31] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From Dense Token to Sparse Memory for Long Video Understanding. In *CVPR*, 2024.

[32] 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 *CVPR*, 2024.

[33] 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 *ICLR*, 2026.

[34] 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. *arXiv preprint arXiv:2502.13923*, 2025.

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

[36] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025.