LEARNING WHAT TO KEEP: CAUSAL MEMORY COMPRESSION FOR STREAMING AUDIO-VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming omnimodal systems often need to answer questions about earlier events displayed. Retaining the full history becomes increasingly expensive as the stream grows, forcing the system to decide what information to keep or delete. We introduce Causal Adaptive Memory Compression (CAMC), a method for managing past-information retention. CAMC allocates a limited memory budget across stream segments, audio and video, and individual tokens. It uses two signals: historical novelty, which measures how different new content is from stored information, and temporal change, which measures how different it is from prior observations. A dynamic budget enables CAMC to retain useful information while staying close to an overall retention target. Additionally, we present StreamRecall, a streaming audio-video multiple-choice benchmark that evaluates whether information from earlier events remains accessible when queried later. CAMC achieves the highest average answer accuracy evaluated while retaining a fraction of the original audio-video history.

1 INTRODUCTION

Audio-video systems that process continuous audio-video streams often face questions about prior events, including conversations, actions, or scene changes. Providing accurate responses requires retaining substantial historical information. However, storing the entire audio-video history becomes resource-intensive as the stream advances. Consequently, the system must decide what to retain without expending excessive tokens to store every detail.

As the stream unfolds, only the audio and video observed up to that moment are available for storage decisions. The process is also *query-blind*, meaning the system does not know what a future question will ask. These decisions are irreversible as discarded tokens cannot be recovered later. Thus, the main challenge is to preserve useful information without knowing in advance which details will be relevant.

This choice involves three related decisions: how much information to retain from each stream segment, how to allocate it between audio and video, and which individual tokens to preserve. The system needs to identify content that contributes new information to what is already stored. However, comparing each observation only with the one immediately before it is not enough, as a scene may differ from the previous observation yet resemble content retained earlier in the stream. Two signals help distinguish which information is more useful to keep: temporal change, which measures the difference from the previous observation, and historical novelty, which measures the difference from previously retained information.

This work introduces CAMC (Causal Adaptive Memory Compression), a training-free method that adjusts retention based on stream content. CAMC uses historical novelty and temporal change to determine how much information to retain. These decisions are managed through a cumulative memory budget, enabling some segments to retain more information and others less, while maintaining retention near a specified target.

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.

We also introduce STREAMRECALL, a benchmark that links questions to the exact timing of their supporting evidence, enabling analysis of how well relevant information survives compression and how well the system retrieves it.

Across three streaming audio-video datasets with varying stream lengths and memory requirements, we show that CAMC is an effective strategy for preserving useful historical information under a fixed memory budget and examine how this is driven by information signals and allocation decisions.

2 RELATED WORK

Omni-modal large language models. Multimodal large language models have evolved toward omni-modal systems that jointly process text, vision, and audio within a unified reasoning framework (Shu et al., 2025; Tang et al., 2025; Xu et al., 2025b; Yao et al., 2024; Fu et al., 2024) . Representative models such as Qwen2.5-Omni Xu et al. (2025a) and MiniCPM-o support joint audio-visual understanding through modality-specific encoders connected to a shared language-model backbone. Such models are well suited to temporally evolving audiovisual scenes, but continuously arriving streams also produce long multimodal token sequences. This has motivated growing interest in efficient Omni-LLM inference and memory mechanisms for long and streaming audio-video inputs.

Streaming multimodal understanding and memory. Most video-language models and benchmarks assume offline access to a complete video before answering, whereas streaming systems must process observations incrementally and may receive queries before the stream ends (Lin et al., 2026; Zhang et al., 2025b). StreamingBench formalizes this setting through temporally situated questions over progressively observed audio-video streams, while recent streaming models study how historical context can be maintained without repeatedly processing the complete prefix (Huang et al., 2025; Zhang et al., 2025a; Lando et al., 2025). The challenge is therefore both causal and cumulative: future observations are unavailable, while historical representations continue to grow with stream duration.

Existing approaches address this accumulation through recurrent memory, input-token memory, or KV-cache compression, using selection, consolidation, quantization, or eviction to maintain a controlled active state Yang et al. (2025b); Chen et al. (2026b); Xiao et al. (2026). StreamingTOM Chen et al. (2026a), for example, combines causal pre-LLM token reduction with bounded quantized memory, while OmniMem Sun et al. (2026) performs modality-aware layer-wise allocation and perturbation-aware KV selection for streaming audio-visual LLMs. Related work also retrieves or compresses historical KV segments to control online memory and query cost (Di et al., 2025; Zhang et al.).

CAMC instead studies complementary episodic memory: compressed historical observations remain separately addressable and can later be retrieved for a delayed information request. Historical episodes are formed causally and without access to the future question, with question-conditioned access deferred until after memory construction. Our focus is therefore on deciding what evidence should survive an unknown future query rather than on maintaining a fixed-size active state.

Multimodal token compression. Token compression reduces redundancy in image, video, and audio representations through pruning, merging, pooling, or importance-based selection (Bolya et al., 2022; Chen et al., 2024; Xing et al., 2024; Yang et al., 2025a; Shen et al., 2026; Tao et al., 2025; Tan et al., 2025; Lin et al., 2025; Lee & Lee, 2025). Recent work extends these ideas to Omni-LLMs (Tao et al., 2026a; Ding et al., 2026; Gong et al.; Su et al., 2026; Deng et al., 2026; Park et al., 2026). OmniZip Tao et al. (2026a) performs training-free audio-guided compression by using audio saliency to retain temporally aligned video tokens, while OmniSIFT Ding et al. (2026) adopts modality-asymmetric compression, first reducing visual redundancy and then using visual anchors to select audio tokens. OmniPack Su et al. (2026) combines modality-specific pre-LLM compression with query-conditioned refinement inside the LLM, while VisionZip Yang et al. (2025a), VidCom2 Liu et al. (2025), and SEATS Xin et al. (2026) explore complementary similarity-, saliency-, and query-guided token reduction strategies.

These approaches provide strong mechanisms for reducing multimodal context at inference time, but streaming historical memory imposes a different constraint: retention decisions must be made

before future observations and the eventual question are available, and discarded evidence cannot later be recovered. CAMC therefore targets causal, query-blind, and irreversible compression of historical audio-video memory.

3 Problem Formulation

We consider question answering over incrementally observed audio-video streams. The system constructs an external episodic memory from past observations and retrieves relevant evidence when a question arrives. Answering is restricted to information available at question time.

Streaming observations. The stream is processed as temporally ordered chunks. After modality-specific encoding,

$$X_t = (V_t, A_t), \tag{1}$$

where $V_t \in \mathbb{R}^{n_t^v \times d}$ and $A_t \in \mathbb{R}^{n_t^a \times d}$ are the visual and audio embeddings, respectively, and $L_t = n_t^v + n_t^a$ is the number of multimodal tokens. Chunks are processing units and need not correspond to semantic events.

At question time, we distinguish historical observations already incorporated into memory from an optional current observation X_{cur}, which is supplied directly to the answering model and contains no future content. The exact history–current boundary is determined by the evaluation protocol.

Causal episodic memory. Let $\mathcal{M}_t$ denote the historical memory after processing chunks $X_1, \ldots, X_t$. A memory policy $\mathcal{U}$ updates it as

$$\mathcal{M}_t = \mathcal{U}(\mathcal{M}_{t-1}, X_t), \qquad \mathcal{M}_0 = \emptyset. \tag{2}$$

Stored representations retain their temporal metadata so that retrieved evidence can be restored in chronological order.

Memory construction is *causal*: the update for X_t cannot access future observations. It is also *query-blind*: the future question is unavailable while historical memory is formed. Thus, compression decisions must be made before knowing either future content or the information request that will later be posed.

Irreversible historical compression. Question-time access to history is restricted to $\mathcal{M}_t$: discarded historical tokens cannot later be revisited or recompressed when the question becomes known. The objective is therefore to preserve answer quality while reducing the historical information retained in memory. Section 4 introduces CAMC's compression policy.

Question-time retrieval and answering. For a question q arriving after t historical chunks, the shared question-time pipeline retrieves relevant memory and predicts

$$\begin{aligned} \mathcal{H}_q &= \text{Retrieve}(q, \mathcal{M}_t), \\ \widehat{y} &= \mathcal{F}(q, \mathcal{H}_q, X_{\text{cur}}), \end{aligned} \tag{3}$$

where $\mathcal{H}_q$ denotes retrieved historical payloads in chronological order and $\mathcal{F}$ is the answering model. When present, X_{cur} is supplied without historical compression; otherwise $X_{\text{cur}} = \emptyset$. Retrieval cannot recover information discarded during historical memory construction.

Across compression comparisons, retrieval and answering are held fixed to isolate historical memory quality. Historical retention should not be interpreted as bounded memory: even at a fixed retained fraction, $\mathcal{M}_t$ may continue growing with stream duration.

4 Causal Adaptive Memory Compression

CAMC performs training-free, content-adaptive historical compression across chunks, modalities, and tokens under a given retention budget. For each incoming chunk, CAMC uses previously observed content to determine how much memory to allocate and which embeddings to preserve. Retained payloads accumulate in $\mathcal{M}_t$, while summaries, running statistics, and budget variables are maintained separately in compression state s_t.

4.1 Content Signals from Past and Present

Historical novelty asks whether an observation differs from what has already been retained; temporal change asks whether it differs from what was just observed. Revisiting a scene can exhibit high temporal change but low historical novelty if its summary resembles retained history. Both signals are embedding-space proxies rather than semantic guarantees.

For modality $m \in \{a, v\}$, we summarize chunk t by the normalized mean embedding

$$k_t^m = \frac{\frac{1}{n_t^m} \sum_i x_{t,i}^m}{\left\| \frac{1}{n_t^m} \sum_i x_{t,i}^m \right\|_2}. \tag{4}$$

Let $\mathcal{K}_{t-1}^m$ denote the append-only set of summaries from previously retained content. Historical novelty and temporal change are

$$\begin{aligned} h_t^m &= \mathrm{clip}\left(1 - \max_{u \in \mathcal{K}_{t-1}^m} \langle k_t^m, u \rangle; 0, 2\right), \\ \delta_t^m &= \mathrm{clip}\big(1 - \langle k_t^m, k_{t-1}^m \rangle; 0, 2\big), \end{aligned} \tag{5}$$

where k_{t-1}^m summarizes the immediately preceding raw chunk. A joint audio-video key k_t^j, obtained from the normalized equal-weight average of the modality summaries, supplies analogous signals. Their weighted combination defines the chunk content score:

$$I_t = \sum_{m \in \{v,a,j\}} w_h^m h_t^m + \sum_{m \in \{v,a,j\}} w_\delta^m \delta_t^m. \tag{6}$$

Weights are renormalized over available terms; their values and initialization details are given in Appendix A.1.

4.2 Adaptive Chunk Allocation under a Cumulative Budget

Content-driven proposal. CAMC interprets I_t relative to previous scores using their running mean μ_{t-1} and sample variance v_{t-1}. When sufficient statistics are available, it standardizes the score as $z_t = (I_t - \mu_{t-1})/(\sqrt{v_{t-1}} + \epsilon)$. A normalized or fallback gate then determines the base retention proposal:

$$\begin{aligned} g_t &= \begin{cases} \sigma(\alpha\, \mathrm{clip}(z_t; -20, 20)), & z_t \text{ available}, \\ \mathrm{clip}(I_t; 0, 1), & \text{otherwise}, \end{cases} \\ r_t^{\mathrm{base}} &= r_{\min} + (r_{\max} - r_{\min}) g_t, \end{aligned} \tag{7}$$

where σ is the sigmoid.

Startup stabilization. Before reliable running statistics are available, CAMC applies a temporary correction that raises below-target proposals toward the internal target $r_\star$ while leaving above-target proposals unchanged. Let r_t^{info} denote the resulting stabilized retention proposal. The correction vanishes as score statistics accumulate; its exact schedule is given in Appendix A.1.

The resulting content proposal is $B_t^{\mathrm{info}} = \mathrm{rnd}(L_t r_t^{\mathrm{info}})$. The nominal allowance $r_\star L_t$, content proposal B_t^{info}, and final budget B_t are distinct: content determines what CAMC would like to spend, while cumulative accounting determines what it can spend while remaining near the target.

Cumulative budget control. CAMC tracks deviation from the long-term target through $E_t = \sum_{s=1}^{t} (B_s - r_\star L_s)$. Underspending creates credit for later chunks, whereas overspending consumes credit or creates debt. The content proposal is adjusted according to this running balance and projected onto the feasible per-chunk budget range to obtain B_t.

The exact credit update, bootstrap handling, projection constraints, and feasibility conditions are detailed in Appendix A.1. These constraints allow temporary over- or underspending while keeping cumulative retention close to the target; they do not impose constant-size memory.

4.3 MODALITY-AWARE ALLOCATION

Chunk allocation determines when to spend memory; modality allocation determines where within the audio-visual chunk to spend it. Using modality score $\omega_t^m = h_t^m + \delta_t^m$, CAMC assigns

$$p_t^a = \frac{\exp(\omega_t^a/\tau)}{\exp(\omega_t^a/\tau) + \exp(\omega_t^v/\tau)}, \qquad p_t^v = 1 - p_t^a. \tag{8}$$

Joint signals affect the total chunk budget but do not enter this split. The desired audio count $\widehat{B}_t^a = \mathrm{rnd}(B_t p_t^a)$ is constrained by modality capacity and minimum-coverage requirements, with video receiving the remainder so that $B_t^a + B_t^v = B_t$.

4.4 TOKEN SELECTION AND STATE UPDATE

Within each modality, CAMC ranks tokens using local distinctiveness, historical novelty, and temporal change. For normalized token $\hat{x}_{t,i}^m = x_{t,i}^m / \max(\|x_{t,i}^m\|_2, \epsilon)$, let $c(s) = \mathrm{clip}(1-s; 0, 2)$. Then,

$$\begin{aligned} d_{t,i}^{m,\mathrm{loc}} &= c\big(\langle \hat{x}_{t,i}^m, k_t^m \rangle\big), \qquad h_{t,i}^{m,\mathrm{tok}} = c\left(\max_{u \in \mathcal{K}_{t-1}^m} \langle \hat{x}_{t,i}^m, u \rangle\right), \\ \delta_{t,i}^{m,\mathrm{tok}} &= c\big(\langle \hat{x}_{t,i}^m, k_{t-1}^m \rangle\big), \qquad u_{t,i}^m = \lambda_{\mathrm{loc}} d_{t,i}^{m,\mathrm{loc}} + \lambda_{\mathrm{hist}} h_{t,i}^{m,\mathrm{tok}} + \lambda_{\mathrm{temp}} \delta_{t,i}^{m,\mathrm{tok}}. \end{aligned} \tag{9}$$

The nonnegative utility weights sum to one.

CAMC retains the B_t^m highest-scoring tokens in each modality, $\mathcal{I}_t^m = \mathrm{TopK}(\{u_{t,i}^m\}_{i=1}^{n_t^m}, B_t^m)$.

Selected embeddings and positional metadata remain unchanged. After restoring their original sequence order as $\widetilde{X}_t = (\widetilde{V}_t, \widetilde{A}_t)$, CAMC appends them to memory as $\mathcal{M}_t \leftarrow \mathrm{Append}(\mathcal{M}_{t-1}, \widetilde{X}_t)$.

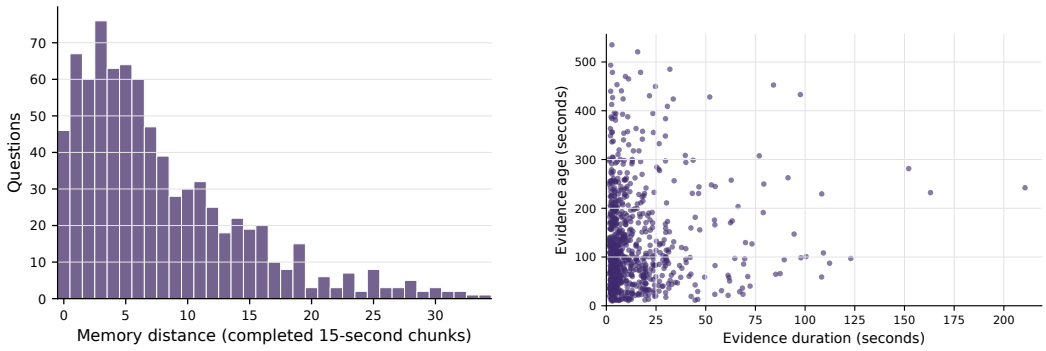
The compression state s_t appends summaries of the selected tokens, updates the raw-chunk references and running statistics, and advances the budget ledger. Historical keys and payloads remain append-only; they are not recomputed when a later question arrives. The resulting memory is then available to the shared retrieval and answering pipeline of Section 3.

5 STREAMRECALL: DELAYED HISTORICAL EVIDENCE EVALUATION

Motivation and construction. For studying historical memory, it is useful to know *when* the evidence required by a question occurred. This allows us to understand how long relevant information must survive before it is queried, and to distinguish error due to chunk-level historical compression from error due to memory recall. Without having exact timestamps for the event needed to answer the question, these failure modes are difficult to analyze from answer accuracy alone.

Therefore, we construct STREAMRECALL, an evaluation set with multiple-choice questions about omnimodal events in video streaming. STREAMRECALL is based on videos from the LongVALE dataset Geng et al. (2025), which provides temporally grounded captions for events over long videos. Overall, we provide 750 omnimodal questions across 424 videos. Each question is associated with a target event containing its supporting evidence. These annotations enable the controlled evidence-access analysis in Sec. 6.4, where historical compression is held fixed while query-time access to the annotated evidence is varied.

Data generation pipeline. Each item is generated from a temporally grounded LongVALE event and passes through three stages: automatic MCQA generation, model-based debiasing and evidence verification, and human validation. GPT-5.6 Terra generates questions and distractors from the target event and its temporal context, after which Gemini 3-Flash removes items that are guessable without the video or unreliable when the supporting segment is provided. Each remaining question is independently reviewed by two human annotators. This pipeline reduces 2,009 generated candidates to 800 automatically verified items and finally to 750 validated questions over 424 videos. The complete generation, filtering, and annotation protocol is provided in Appendix A.3.



(a) Distance in completed 15-s memory chunks.

(b) Evidence duration versus evidence age.

Figure 1: Characterization of delayed historical evidence in STREAMRECALL. **Left:** relevant evidence spans a broad range of distances under the 15-s chunking used by our streaming evaluation. **Right:** temporally distant evidence occurs across a range of supporting-event durations, including compact events.

Query placement. Each question is assigned a frozen query boundary after its annotated target event. We define *evidence age* as the elapsed time between the end of the supporting event and this query boundary. At evaluation time, only content observed before the boundary is available, and target-event annotations are never exposed to the compressor. Thus, historical memory must be formed causally and query-blindly; the annotations are used only for dataset analysis and controlled diagnostics.

Delayed historical evidence. Figure 1 shows that supporting evidence occurs across a broad range of historical distances under our 15-s chunking and spans diverse event durations, including compact events. STREAMRECALL therefore tests whether useful audio-visual evidence can survive substantial intervening stream content before it is queried.

Its temporal grounding complements broader streaming benchmarks by enabling direct analysis of both long-term retention and query-time access to retained historical evidence.

6 EXPERIMENTS AND RESULTS

6.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate on three complementary streaming audio-video QA settings. **StreamingBench** Lin et al. (2026) provides 1,000 multiple-choice questions over 200 videos; we use its *Omni-source Understanding* split, which tests joint audio-video understanding across emotion, scene, source, and alignment tasks. **LVOmniBench** Tao et al. (2026b) contains 1,014 questions over 275 substantially longer videos (10–90 min; 34.5 min on average), testing compression over extended audiovisual histories. Finally, **STREAMRECALL** (Sec. 5) contains 750 questions over 424 videos and specifically stresses delayed access to temporally grounded historical evidence.

Models and baselines. Primary experiments use Qwen2.5-Omni-7B Xu et al. (2025a). We compare CAMC with random retention, OmniZip Tao et al. (2026a), OmniSIFT Ding et al. (2026), OmniPack Su et al. (2026), VisionZip Yang et al. (2025a), SEATS Xin et al. (2026), and VidCom2 Liu et al. (2025), together with the uncompressed model when a protocol-compatible reference is available. For methods with query-dependent or otherwise noncausal stages, we use the corresponding pre-question variant and mark it explicitly in the tables.

Streaming evaluation protocol. All methods follow the query-blind, irreversible memory setting of Sec. 3. Historical memory is constructed before the question is available; the current query interval remains uncompressed, and question-conditioned retrieval operates only over the already

Table 1: Comparison of answer accuracy across the three streaming audio-video evaluation settings at the tight ($\sim$ 25%) and main ($\sim$ 43%) historical-retention regimes. Comparisons should be interpreted within each benchmark protocol. † denotes a result whose achieved retention differs materially from the nominal operating point.

Method	Benchmark Performance			Average
	StreamingBench	LVOmni	StreamRecall	Avg. Score
Qwen2.5-Omni-7B (Raw)	52.41	–	55.07	–
Tight retention regime ($\approx 25\%$)				
Random	54.41	34.85	45.87	45.04
OmniZip	57.52	33.27†	45.20†	45.33
OmniSIFT	54.71	**37.02**	48.13	46.62
OmniPack-Pre-Causal	55.01	35.64	48.93	46.53
VisionZip-OM	56.91	36.72	49.47	47.70
SEATS-Pre	55.01	33.17	**50.93**	46.37
VidCom2	–	32.87†	–	–
CAMC (ours)	**58.02**	36.72	50.80	**48.51**
Main retention regime ($\approx 43\%$)				
Random	54.71	35.24	51.07	47.01
OmniZip	53.91	33.76†	51.87	46.51
OmniSIFT	53.51	**37.91**	48.80	46.74
OmniPack-Pre-Causal	53.91	36.72	52.53	47.72
VisionZip-OM	53.41	36.13	52.53	47.36
SEATS-Pre	54.51	35.64	52.53	47.56
VidCom2	55.21	36.72	52.80	48.24
CAMC (ours)	**55.41**	**37.91**	**53.20**	**48.84**

Scores are multiple-choice accuracy (%); tight and main target approximately 25% and 43% historical audio-video retention. †Achieved retention differs materially from the target.

retained history. Within each benchmark, retrieval and answering are held fixed across compression methods.

Retention regimes and metric. Physical historical retention is $R = N^{\text{hist}}_{\text{stored}}/N^{\text{hist}}_{\text{raw}}$, measured over encoded audio-video tokens before retrieval-time expansion and excluding text and the uncompressed current interval. We evaluate a *tight* ($\sim 25\%$) and *main* ($\sim 43\%$) retention regime and report achieved retention for every method; materially unmatched operating points are marked with †. Our primary metric is multiple-choice answer accuracy.

6.2 Results and Analysis

Table 1 compares CAMC with competing compression methods at the tight ($\sim 25\%$) and main ($\sim 43\%$) retention regimes. Across the three evaluation settings, CAMC achieves the highest average accuracy at both regimes and ranks first or within 0.3 percentage points of the best method on every benchmark.

StreamingBench. CAMC achieves the highest accuracy at both operating points. Under tight retention, it reaches 58.02%, outperforming the strongest competing method, OmniZip, by 0.50 percentage points. At the main operating point, CAMC obtains 55.41%, exceeding VidCom2 at 55.21%. Thus, CAMC remains the top-performing method across substantially different historical-memory budgets rather than benefiting only at a single retention level. A task-level breakdown is provided in Appendix A.5.

Table 2: Component analysis of CAMC on Qwen2.5-Omni-7B. StreamingBench uses the tight ($\sim$ 25%) regime, while LVOmniBench and STREAMRECALL use the main ($\sim$ 43%) regime. Values are accuracy (%); parentheses report the change relative to full CAMC. All comparisons use closely matched historical-retention budgets.

Variant	StreamingBench	LVOmniBench	STREAMRECALL
Full CAMC	**58.02**	**37.91**	**53.20**
w/o historical novelty	57.92 (-0.10)	35.83 (-2.08)	52.80 (-0.40)
w/o temporal change	57.52 (-0.50)	36.13 (-1.78)	52.67 (-0.53)
Uniform modality retention	58.32 (+0.30)	37.31 (-0.60)	53.07 (-0.13)

LVOmniBench. CAMC remains strong in the longer-horizon LVOmniBench setting. At tight retention, CAMC reaches 36.72%, within 0.30 percentage points of OmniSIFT (37.02%) and matching VisionZip-OM. At the main operating point, CAMC reaches 37.91%, tying OmniSIFT for the highest accuracy. These results show that CAMC remains competitive even when compression is applied over much longer audiovisual histories.

STREAMRECALL. CAMC remains competitive under aggressive compression and becomes the strongest compressed method at the main operating point. At tight retention, CAMC obtains 50.80%, only 0.13 percentage points below SEATS-Pre (50.93%). At the main regime, CAMC reaches 53.20%, outperforming all compressed baselines, including VidCom2 at 52.80%. The uncompressed model reaches 55.07%, leaving CAMC only 1.87 percentage points below Raw while retaining less than half of the historical audiovisual tokens.

Overall, CAMC delivers a strong accuracy–memory trade-off across three streaming settings with substantially different temporal scales and task distributions. Its strongest gains are not confined to a single benchmark or retention regime, supporting the use of causal, content-adaptive historical memory formation across diverse streaming conditions.

6.3 COMPONENT ANALYSIS

Table 2 isolates the contributions of CAMC's historical-novelty and temporal-change signals, together with its adaptive audio–video allocation. Removing temporal change reduces accuracy by 0.50 points on StreamingBench, 1.78 points on LVOmniBench, and 0.53 points on STREAMRECALL. Removing historical novelty also hurts all three benchmarks, with the largest individual drop on LVOmniBench, from 37.91% to 35.83% (-2.08 points). The two signals therefore provide complementary information for causal memory formation: temporal change has the larger removal effect on StreamingBench and STREAMRECALL, while historical novelty is particularly important on LVOmniBench.

The effect of adaptive modality allocation is more dataset-dependent. Replacing CAMC's content-dependent audio–video split with uniform modality retention decreases accuracy by 0.60 points on LVOmniBench and 0.13 points on STREAMRECALL, while StreamingBench slightly favors the uniform split (58.32% versus 58.02%). Thus, adaptive modality allocation improves two of the three evaluations but is not uniformly beneficial across datasets.

6.4 DIAGNOSING HISTORICAL EVIDENCE ACCESS

The results above evaluate how effectively different compression strategies preserve historical audiovisual information under constrained memory. However, preserving useful evidence does not guarantee that the model can reliably identify and use it when a question arrives much later. We therefore study a complementary question: to what extent is performance limited by query-time access to relevant historical evidence rather than by information loss during compression?

Gold-evidence access oracle. We construct a diagnostic oracle using the annotated target-event interval in STREAMRECALL. Historical compression remains unchanged, causal, and query-blind; annotations are revealed only when the question arrives. At query time, the oracle restores completed historical chunks overlapping the annotated evidence interval in temporal order, while the current

(a) Gold-evidence access diagnostic on STREAMRECALL at the main ($\sim$ 43%) historical-retention regime. Compression is unchanged; only query-time evidence access is modified.

Method	Normal	Gold Evidence	Gain
Raw	55.07	83.20	+28.13
Random	51.07	76.53	+25.46
OmniZip	51.87	78.27	+26.40
OmniSIFT	48.80	76.00	+27.20
OmniPack-Pre-Causal	52.53	78.00	+25.47
VisionZip-OM	52.53	80.93	+28.40
SEATS-Pre	52.53	81.20	+28.67
VidCom2	52.80	77.20	+24.40
CAMC (ours)	53.20	79.60	+26.40

(b) Accuracy (%) by evidence age on STREAMRECALL under the main historical-retention regime. All methods operate at closely matched retention (43%).

Method	10–30 s	30–60 s	60–120 s	$>$ 120 s
Random	65.38	51.28	50.43	47.99
OmniZip	66.67	55.56	50.86	47.68
OmniSIFT	62.82	50.43	46.55	46.44
OmniPack-Pre-Causal	66.67	57.26	51.29	**48.30**
VisionZip-OM	**67.95**	57.26	**52.16**	47.37
SEATS-Pre	**67.95**	**58.97**	50.43	47.99
VidCom2	**67.95**	57.26	**52.16**	47.99
CAMC (ours)	**67.95**	**58.97**	**52.16**	**48.30**

Table 3: Historical-evidence analysis on STREAMRECALL. **Left:** privileged access to the annotated evidence interval reveals substantial headroom across all compression methods. **Right:** compressed-model accuracy decreases as the relevant evidence becomes more temporally distant, while CAMC remains tied for the best result in every evidence-age interval.

interval remains uncompressed. No future content or historical recompression is used. The oracle therefore provides privileged evidence access rather than oracle-assisted compression.

Table 3a reveals substantial gains from privileged evidence access across every evaluated compression method. CAMC improves from 53.20% to 79.60%, a gain of 26.40 percentage points, while competing compressors exhibit similarly large improvements. This indicates substantial performance headroom even after the historical memory has already been formed.

More importantly, the same effect appears in the absence of compression. The uncompressed model improves from 55.07% to 83.20%, a gain of 28.13 points, when historical context is restricted to the annotated evidence. Because Raw does not discard historical audiovisual tokens, this improvement cannot be explained solely by useful information being removed during compression. Instead, the result indicates that locating and accessing relevant evidence within long historical context is itself a major source of error.

The access gap grows with evidence age. Performance degrades as supporting evidence becomes more temporally distant. For CAMC, the gap between normal and gold-evidence access grows from 6.41 percentage points at 10–30 s to 33.75 points beyond 120 s. The same degradation appears across competing compressors, while CAMC remains tied for the highest compressed-model accuracy in every evidence-age interval (Table 3).

Notably, gold-evidence performance remains high even for distant events, indicating that older evidence can remain useful once it is successfully surfaced. The widening access gap therefore reveals substantial headroom in query-time use of retained historical memory, beyond the information lost during compression itself. We provide the full age-stratified analysis and gold-evidence trajectories in Appendix A.6.

7 CONCLUSION

In this work, we present CAMC, a training-free compression method that uses historical novelty and temporal change to adjust how much information is retained across stream segments, audio and video, and individual tokens under a given memory budget. Across three streaming audio-video benchmarks, CAMC achieves the highest average accuracy at both tested retention levels and performs within 0.3 percentage points of the best result on each benchmark. Our ablations demonstrate that historical novelty and temporal change offer complementary signals for determining which information to retain. Additionally, we introduce STREAMRECALL, a benchmark that links questions to the time at which their supporting evidence appears. This framework allows us to examine whether information from earlier events remains accessible when required at later stages. Results indicate that long-term performance depends on preserving relevant information during compression and successfully retrieving that information required for a given question.

AI USE STATEMENT

In this work, we used generative AI tools at several stages of the research workflow. GPT-5.6 Terra was used to generate candidate multiple-choice questions and distractors for STREAMRECALL, and Gemini 3-Flash was used for model-based debiasing and evidence-grounded verification as described in the dataset construction protocol. Generative AI tools, including ChatGPT and Codex, were also used to assist with brainstorming and refining research ideas, experimental and ablation design, implementation and debugging. They additionally assisted with literature discovery and summarization, manuscript organization, drafting, editing, and language refinement.

All AI-assisted research outputs were reviewed and validated by the authors. AI-assisted code was inspected and tested before being used for reported experiments, and experimental claims were checked against the corresponding run outputs, configurations, and evaluation records. Literature suggested or summarized by generative AI was checked against the original sources before being cited or used to support claims. Automatically generated STREAMRECALL questions underwent independent model-based filtering followed by human validation before inclusion in the final benchmark. The authors reviewed and revised all AI-assisted manuscript text and verified the methodological descriptions, equations, reported results, and conclusions. We take responsibility for the final content of this work, including text, claims, code, data, and other artifacts produced with the aid of generative AI.

REFERENCES

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. Streamingtom: Streaming token compression for efficient video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24675–24685, 2026a.

Yilong Chen, Xiang Bai, Zhibin Wang, Chengyu Bai, Yuhan Dai, and Ming Lu. Streamkv: Streaming video question-answering with segment-based kv cache retrieval and compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3120–3128, 2026b.

Yuchen Deng, Zidang Cai, Hai-Tao Zheng, Jie Wang, Feidiao Yang, and Yuxing Han. Omnirefine: Alignment-aware cooperative compression for efficient omnimodal large language models. *arXiv preprint arXiv:2605.12056*, 2026.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 42115–42127, 2025.

Yue Ding, Yiyan Ji, Jungang Li, Xuyang Liu, Xinlong Chen, Junfei Wu, Bozhou Li, Bohan Zeng, Yang Shi, Yushuo Guan, et al. Omnisift: Modality-asymmetric token compression for efficient omni-modal large language models. *arXiv preprint arXiv:2602.04804*, 2026.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18959–18969. IEEE, 2025.

Chao Gong, Depeng Wang, Zhipeng Wei, Ya Guo, Huijia Zhu, and Jingjing Chen. Echoingpixels: Aliasing-resistant joint token reduction for audio-visual llms. In *Forty-third International Conference on Machine Learning*.

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

Giuseppe Lando, Rosario Forte, Giovanni Maria Farinella, and Antonino Furnari. How far can off-the-shelf multimodal large language models go in online episodic memory question answering? In *International Conference on Image Analysis and Processing*, pp. 499–511. Springer, 2025.

Taehan Lee and Hyukjun Lee. Token pruning in audio transformers: Optimizing performance and decoding patch importance. *arXiv preprint arXiv:2504.01690*, 2025.

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

Yueqian Lin, Yuzhe Fu, Jingyang Zhang, Yudong Liu, Jianyi Zhang, Jingwei Sun, Hai Helen Li, and Yiran Chen. Speechprune: Context-aware token pruning for speech information retrieval. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

Xuyang Liu, Yiyu Wang, Junpeng Ma, and Linfeng Zhang. Video compression commander: Plug-and-play inference acceleration for video large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1910–1924, 2025.

Yeo Jeong Park, Hyemi Jang, Minseo Choi, Jongsun Lee, Jooyoung Choi, and Yongkweon Jeon. Omnidrop: Layer-wise token pruning for omni-modal llms via query-guidance. *arXiv preprint arXiv:2605.14458*, 2026.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, et al. Fastvid: Dynamic density pruning for fast video large language models. *Advances in Neural Information Processing Systems*, 38:123553–123581, 2026.

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4305–4314. IEEE, 2025.

Wanshun Su, Yang Shi, Feihu Liu, Ziwen Yu, Yan Min, Zhuoran Zhang, Qixun Wang, Haotian Wang, Shixuan Liu, Yuanxing Zhang, et al. Omnipack: Unified token compression for efficient omni-modal large language models. *arXiv preprint arXiv:2608.03812*, 2026.

Guangzhi Sun, Yixuan Li, Yudong Yang, and Chao Zhang. Omnimem: Perturbation-aware memory compression for streaming audio-visual llms. *arXiv preprint arXiv:2606.07577*, 2026.

Xudong Tan, Peng Ye, Chongjun Tu, Jianjian Cao, Yaoxin Yang, Lin Zhang, Dongzhan Zhou, and Tao Chen. Tokencarve: Information-preserving visual token compression in multimodal large language models. *arXiv preprint arXiv:2503.10501*, 2025.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18992–19001. IEEE, 2025.

Keda Tao, Kele Shao, Bohan Yu, Weiqiang Wang, Jian Liu, and Huan Wang. Omnizip: Audio-guided dynamic token compression for fast omnimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17682–17692, 2026a.

Keda Tao, Yuhua Zheng, Jia Xu, Wenjie Du, Kele Shao, Hesong Wang, Xueyi Chen, Xin Jin, Junhan Zhu, Bohan Yu, et al. Lvomnibench: Pioneering long audio-video understanding evaluation for omnimodal llms. *arXiv preprint arXiv:2603.19217*, 2026b.

Junbin Xiao, Jiajun Chen, Tianxiang Sun, Xun Yang, and Angela Yao. Mukv: Multi-grained kv cache compression for long streaming video question-answering. *arXiv preprint arXiv:2605.22269*, 2026.

Zijie Xin, Jie Yang, Ruixiang Zhao, Tianyi Wang, Fengyun Rao, Jing Lyu, and Xirong Li. Stage-adaptive token selection for efficient omni-modal llms. *arXiv preprint arXiv:2605.20035*, 2026.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025a.

Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. Streammem: Query-agnostic kv cache memory for streaming video understanding. *arXiv preprint arXiv:2508.15717*, 2025b.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Gengyuan Zhang, Mingcong Ding, Tong Liu, Yao Zhang, and Volker Tresp. Memory helps, but confabulation misleads: Understanding streaming events in videos with mllms. *arXiv preprint arXiv:2502.15457*, 2025a.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21059–21069. IEEE, 2025b.

Haowei Zhang, Shudong Yang, Jinlan Fu, See-Kiong Ng, and Xipeng Qiu. Hermes: Kv cache as hierarchical memory for efficient streaming video understanding, 2026. *URL https://arxiv.org/abs/2601.14724*.