# TOWARDS A DYNAMIC AND FIXED-BUDGET MEMORY BANK FOR STREAMING VIDEO UNDERSTANDING

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

## ABSTRACT

Currently, streaming video understanding is still a daunting task for existing *Multimodal Large Language Models* (MLLMs). Its difficulties not only lie in handling the ever-increasing video frames, but also in the unpredictability of future video content and input instructions. In this paper, we study this task from the perspective of constructing a dynamic but fixed-budget memory bank, and propose a novel and training-free approach termed ***CausalMem***. CausalMem targets at constructing a dynamic visual memory update mechanism, thereby maximizing the amount of information in streaming video within a limited memory space, much like *human brain*. In practice, CausalMem estimates the redundancy of visual tokens and updates the memory bank via carefully constructed *online semantic basis*, which models the principal semantics of video streams. To validate CausalMem, we apply it to three representative MLLMs, namely LLaVA-OneVision, Qwen2.5-VL and Qwen3-VL respectively, and conduct extensive experiments on both streaming and offline[1] video understanding benchmarks. The experimental results not only show the great advantages than existing methods using streaming or offline settings, *e.g.*, $+3.2\%$ and $+3.0\%$ average accuracy gains respectively, but also witness the superior semantic preservation for streaming videos, *e.g.*, using $12k$ token budgets to memorize hour-long streaming videos, which achieves more than **20**$\times$ visual token compression ratio and only occupies about **82 MB** storage. **Our code** is given in the supplementary materials.

## 1 INTRODUCTION

After achieving great successes in a variety of image-language tasks Antol et al. (2015); Chen et al. (2015); Mishra et al. (2019), recent *Multimodal Large Language Models* (MLLMs) Tong et al. (2025); Liu et al. (2023); Li et al. (2023) also yield notable progress in the video domain Zhang et al. (2024); Li et al. (2024); Bai et al. (2025b). To overcome the substantial computation caused by the excessive video tokens, numerous works Ye et al. (2025); Ju et al. (2026); Shao et al. (2025); Fan et al. (2026) have been recently devoted to the research of efficient video understanding, employing techniques such as *token pruning* Yang et al. (2025a); Ye et al. (2025); Ju et al. (2026) and *KV Cache compression* Yang et al. (2025b); Di et al. (2025); Chen et al. (2026a). In terms of video content compression, current research primarily focuses on offline video processing, which typically requires globally semantical compression strategies based on the entire video content. Moreover, for streaming video understanding, the fixed-budget compression still remains underexplored.

In particular, achieving efficient streaming video understanding is an daunting challenge for existing MLLMs Li et al. (2024); Tong et al. (2025); Bai et al. (2025a), of which difficulties mainly lie in two aspects. Firstly, MLLMs need to process the video of potentially unbounded length in the streaming scenario. For instance, LLaVA-OneVision Li et al. (2024) needs to cope with more than $700k$ visual tokens for a two-hour video (1 FPS). Moreover, even the advanced video compression methods Chen et al. (2025); Ju et al. (2026) will encounter a rapid accumulation of visual tokens over time, placing a heavy burden on both computation and storage. Secondly, when compressing streaming videos, it is almost impossible to predict future video content and potential user questions. This case renders many existing compression methods unapplicable, such as those based on global redundancy Ju et al. (2026); Shao et al. (2025) or text-guided estimation Zhang et al. (2026); Li et al. (2025). Therefore, we raise a question that

[1] Evaluated under streaming protocol.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

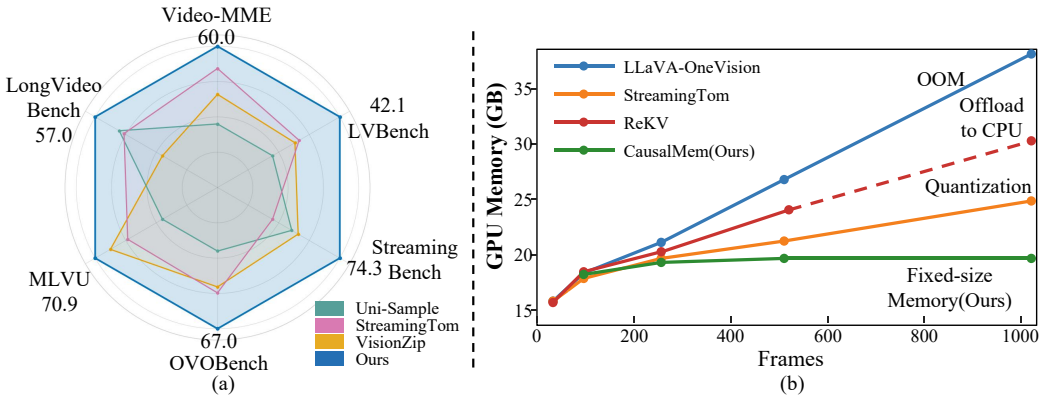



Figure 1: (a) CausalMem consistently outperforms existing visual compression methods across streaming and offline video understanding benchmarks using LLaVA-OneVision. (b) Compared with existing streaming video methods, CausalMem maintains a fixed-size memory bank, enabling streaming video understanding with bounded computation and storage overhead.

*"Is it possible to achieve the dynamic and budgeted visual memory bank for efficient streaming video understanding, much like the human brain?"*

In this paper, we study this problem from the perspective of constructing a dynamic and fixed-budget visual memory bank, and propose a novel training-free method for existing MLLMs, termed **CausalMem**. In principle, the target of CausalMem is to maximize the information amount of the budgeted memory bank, thereby achieving effective streaming understanding of MLLMs with bounded inputs. In practice, we introduce a novel memory update mechanism based on a carefully built *online semantic basis* to evaluate the significance of streaming semantics to memory bank. In this way, CausalMem can greatly reduce the computation and memory overhead of streaming video understanding by using bounded and critical visual memories.

To validate CausalMem, we apply it to three MLLMs, namely LLaVA-OneVision Li et al. (2024), Qwen2.5-VL Bai et al. (2025b) and Qwen3-VL Bai et al. (2025a), and conduct extensive experiments on both streaming and offline video understanding benchmarks. The experiments show that CausalMem achieves obvious performance gains than the compared streaming methods, *e.g.*, $+3.2\%$ and $+3.0\%$ average performance gains on streaming and offline benchmarks, respectively, while preserving the principal visual semantics of videos with only $12k$ visual token budget, *e.g.*, achieving more than $20\times$ compression ratio for hour-long videos while consuming the storage of only 82MB.

Overall, our contributions are three-fold:

- We study the efficient streaming video understanding from the perspective of a dynamic yet fixed budget memory bank under strict causality, and propose a novel and training-free approach for existing MLLMs, termed CausalMem.
- We equip CausalMem with an innovative memory mechanism based on carefully constructed *online semantic basis*, which ensures the information maximization of the memory bank.
- The proposed CausalMem obtains competitive or even better performance than existing efficient video compression methods on both streaming and offline benchmarks, and also exhibits its superiority in streaming video semantic compression with fixed token budgets.

## 2 Related Work

**Offline Video Understanding.** Recent Multimodal Large Language Models Liu et al. (2023); Bai et al. (2025b); Team et al. (2024) have shown strong capabilities in video understanding. However, long videos introduce excessive visual tokens, resulting in high computational and memory costs. Existing methods address this issue from two main perspectives. One adopts Retrieval-Augmented Generation (RAG) to select task-relevant frames based on vision–text similarity, but requires user queries in advance. The other performs visual token compression via pruning or merging Ju et al. (2026); Shao et al. (2025); Bolya et al. (2022); Shang et al. (2025), either based on global video content or query-aware relevance. Despite their effectiveness, they require whole video content or queries in advance and are thus unsuitable for streaming scenarios.

**Streaming Video Understanding.** To overcome these limitations, recent works study streaming video understanding, where visual tokens are processed only based on observed frames. Most methods aim to reduce visual tokens through memory or compression mechanisms. KV cache compression methods, such as StreamMem Yang et al. (2025b), ReKV Di et al. (2025) and LiveVLM Ning et al. (2025) compress the KV-caches of visual tokens to reduce token length. Other approaches introduce structured memory, such as FluxMem Xie et al. (2026), which organizes visual tokens into hierarchical memory with a First-In-First-Out strategy to manage information of different levels. In addition, StreamingTOM Chen et al. (2026b) combines token compression before LLM with KV cache quantization to further reduce tokens. Some methods further incorporate query-aware retrieval. For instance, LiveVLM Ning et al. (2025) and StreamingTOM Chen et al. (2026b) retrieve relevant visual tokens conditioned on the user query. Despite these advances, existing methods primarily rely on local redundancy within the observed visual tokens or require query guidance, which limits their effectiveness in fully capturing the global semantic structure of the video stream.

# 3 METHOD

## 3.1 OVERVIEW

Given a video $V$ and a question $Q$, an MLLM aims to generate the answer by maximizing the conditional generation objective:

$$p(Y \mid \mathbf{V}, Q) = \prod_{i=1}^{L} p(y_i \mid G, \mathbf{V}, Q, Y_{<i}), \tag{1}$$

where $G$ denotes the generator of MLLM, $p(\cdot)$ denotes the probabilities of the predicted word, $Y = \{y_1, ..., y_L\}$ is the answer sequence and $L$ is its length. $\mathbf{V}$ is a subsequence of visual tokens encoded from the video $V$ and $Y_{<i}$ denotes the answer subsequence before the i-$th$ step.

Due to the limited context length and memory budget of MLLMs, some methods Ju et al. (2026); Shao et al. (2025) apply token compression to reduce the length of $\mathbf{V}$, *e.g.*, by pruning or merging the redundant tokensJu et al. (2026); Shao et al. (2025); Fan et al. (2026). Meanwhile, some endeavors Chen et al. (2026a); Shu et al. (2025) compress the KV caches of MLLMs generated by visual tokens. However, these compression methods usually need to operate on the complete video or ignore the upper limit of both computation and storage, which are inferior in streaming scenarios.

To this end, we aim to maintain a visual memory bank $\mathbf{M} \in \mathbb{R}^{b \times d}$ with a fixed budget (length) $b$, $d$ is the visual token feature dimension, and generate the answer based on this memory bank:

$$p(Y \mid \mathbf{V}, Q) = \prod_{i=1}^{L} p(y_i \mid G, \mathbf{M}, Q, Y_{<i}) \tag{2}$$

The memory bank keeps a compact set of visual tokens while being dynamically updated as each new frame $I$ arrives. Therefore, the problem we study becomes how to maintain $\mathbf{M}$ online so that it preserves the most informative tokens of the observed video stream under strict causality.

In this context, we propose a novel memory update scheme based on a carefully-built *online semantic basis* $\mathbf{B} \in \mathbb{R}^{q \times d}$ , where $q$ is the maximum number of semantic basis vectors in $\mathbf{B}$ and $q \ll t \times N$, where $t$ and $N$ denote the number of streaming frames and visual tokens per frame, respectively.

In practice, $\mathbf{B}$ is used to dynamically model the principal semantics of the observed video stream, based on which we can estimate the redundancy of streaming visual tokens and update the memory bank. Then we compress the visual tokens into $\mathbf{M}$ according to their redundancy once $\mathbf{M}$ achieves the upper limit. Note that, the updating process of $\mathbf{B}$ and $\mathbf{M}$ are both causal, *i.e.* unpredictable to future frames and the question. Therefore, $M$ can keep a fixed budget under strict causality.

## 3.2 MEMORY UPDATE WITH ONLINE SEMANTIC BASIS

The main challenge of CausalMem is how to effectively identify and write the key semantics of video streamings into a fixed-budget memory bank, and the curx lies in maximizing the information amount of the memory bank. In this case, we propose a novel memory update mechanism based on *online semantic basis* evaluation.

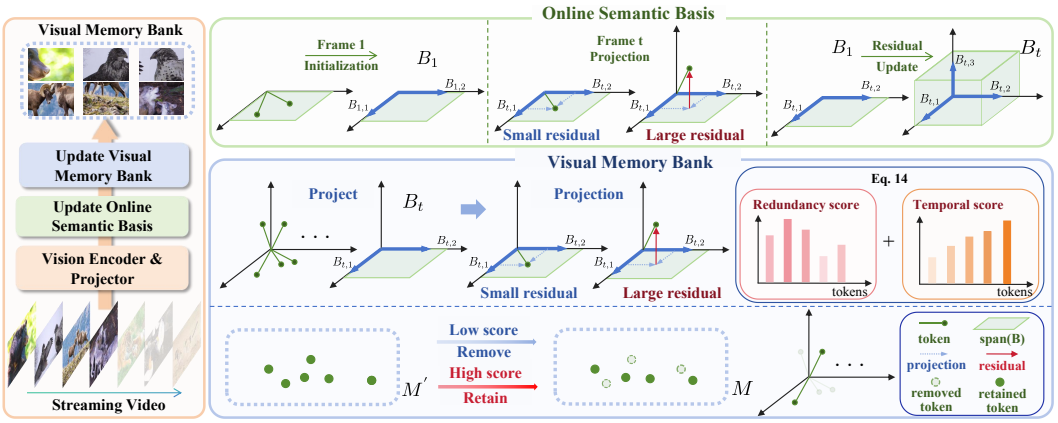



Figure 2: Overview of CausalMem. Given a streaming video, CausalMem encodes incoming frames and maintains a fixed-size visual memory bank under strict causality. An online semantic basis is dynamically updated to capture the principal semantics of the observed stream, and token residuals with respect to this basis are used to estimate redundancy. The memory bank is then updated by retaining tokens with low redundancy and temporal recency, thereby maximizing its information.

### 3.2.1 Online Memory Evaluation with Semantic Basis

In practice, given the video stream at the $t$-th timestamp, we maintain an *Online Semantic Basis* $\mathbf{B}_t \in \mathbb{R}^{q_t \times d}$, where $q_t \leq q$ and $q$ is the max number of basis vectors. Let the visual tokens of the $t$-th frame be denoted as $\mathbf{X}_t \in \mathbb{R}^{N \times d}$. At the 1-*st* step, we initialize the semantic basis from $\mathbf{X}_1$ via compact *Singular Value Decomposition (SVD)* Golub & Van Loan (2013):

$$\mathbf{X}_1 \approx \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{W}_1^\top. \tag{3}$$

Here, $\mathbf{U}_1 \in \mathbb{R}^{N \times q_1}$ and $\mathbf{W}_1^\top \in \mathbb{R}^{q_1 \times d}$ denote the left and right singular vectors, respectively, while $\mathbf{\Sigma}_1 \in \mathbb{R}^{q_1 \times q_1}$ contains the corresponding singular values indicating the importance of each component. We retain the top $q_1 \in (0, q]$ singular vectors and initialize the basis as

$$\mathbf{B_1} = \mathbf{W}_1 = [w_1, w_2, ..., w_{q_1}] \in \mathbb{R}^{q_1 \times d}, \tag{4}$$

where $w_i \in \mathbb{R}^d$ denotes the $i$-th right singular vector. Meanwhile, we introduce a *basis activity score* $\mathbf{s}_t \in \mathbb{R}^{q_t}$ to track the usage frequency of each basis vector, which is initialized by 1 and used to guide the update of the semantic basis $\mathbf{R}$.

For the evolving video stream, the basis matrix is updated at each time step. Given a new frame, we project its visual tokens $\mathbf{X}_t$ onto the current basis $\mathbf{B}_{t-1}$ and reconstruct them by

$$\hat{\mathbf{X}}_t = \mathbf{X}_t \mathbf{B}_{t-1}^\top \mathbf{B}_{t-1}, \tag{5}$$

where $\hat{\mathbf{X}}_t \in \mathbb{R}^{N \times d}$ represents the reconstructed token features from their projection onto the subspace spanned of the semantic basis $\mathbf{B}_{t-1}$. Then, the residual matrix $\mathbf{R}_t = \mathbf{X}_t - \hat{\mathbf{X}}_t$ captures the information that cannot be represented by $\mathbf{B}_{t-1}$. In this case, we can measure the redundancy of each token $\mathbf{X}_{t,i}$ via its residual norm, defined by

$$e_{t,i} = \|\mathbf{R}_{t,i}\|_2, \tag{6}$$

where $\|\cdot\|_2$ denotes the $L_2$ norm. A larger $e_{t,i}$ indicates that the tokens contain more novel and informative semantics, also indicating less redundancy.

To update basis $\mathbf{B}_{t-1}$, we select the top-$k$ least redundant tokens to form candidate tokens $\mathbf{X}_t^{\text{cand}} \in \mathbb{R}^{k \times d}$, whose residual $\mathbf{R}_t^{\text{cand}} \in \mathbb{R}^{k \times d}$ can be calculated akin to Eq. equation 5 and Eq. equation 6. Then, we remove the semantics already captured by $\mathbf{B}_{t-1}$ in $\mathbf{R}_t^{\text{cand}}$, resulting in $\hat{\mathbf{R}}_t$:

$$\hat{\mathbf{R}}_t = \mathbf{R}_t^{\text{cand}} \left( \mathbf{I}_d - \mathbf{B}_{t-1}^\top \mathbf{B}_{t-1} \right), \tag{7}$$

where $\mathbf{I}_d \in \mathbb{R}^{d \times d}$ denotes the identity matrix. We then perform *QR decomposition*Golub & Van Loan (2013) on $\hat{\mathbf{R}}_t$:

$$\hat{\mathbf{R}}_t = \bar{\mathbf{Q}}_t \bar{\mathbf{R}}_t, \tag{8}$$

where $\bar{\mathbf{Q}}_t \in \mathbb{R}^{k\times d}$ provides the orthonormal new basis. We update the basis and activity scores as:

$$\mathbf{B}_t^{'} = [\mathbf{B}_{t-1}, \bar{\mathbf{Q}}_t], \;\; s_t^{'} = [s_{t-1}, \bar{s}_t], \tag{9}$$

where $\bar{s}_t \in \mathbb{R}^k$ denotes the activity scores of $\bar{\mathbf{Q}}_t$. And $\mathbf{B}_t^{'}$ remains row-orthonormal throughout the whole processing, the proof is provided in Appendix A.1.

Meanwhile, to adapt to the evolving video stream, we dynamically update the activity score $s_t$. Specifically, we denote the $i$-th token of $t$-th frame as $\mathbf{X}_{t,i}$. For the $j$-th basis vector of $\mathbf{B}_t$ at the $t$-th time step, we define it as $\mathbf{B}_{t,j}$ and its instantaneous activity as $a_{t,j}$:

$$a_{t,j} = \frac{1}{N}\sum_{i=1}^{N} |\mathbf{B}_{t,j}\mathbf{X}_{t,i}^{\top}|. \tag{10}$$

Then we update the activity score of the $j$-th vector $s_{t,j}$ using an exponential moving average:

$$s_{t,j} = \gamma s_{t-1,j} + (1-\gamma)a_{t,j}, \tag{11}$$

where $a_{t,j}$ measures the instantaneous activation of the $j$-th basis vector in the current frame, and $s_{t,j}$ denotes the temporally smoothed activity score that smoothes historical activity with recent observations. Here, $\gamma \in [0, 1]$ is the smoothing factor.

Once the number of vectors in $\mathbf{B}_t^{'}$ exceeds the predefined limit $q$, we retain the $q$ most active basis vectors according to $s_t^{'}$. We denote $\mathcal{J}_t$ as the retained index set. $\mathbf{B}_t$ and $\mathbf{s}_t$ are updated by:

$$\mathbf{B}_t \leftarrow \mathbf{B}_t^{'}(\mathcal{J}_t), \quad \mathbf{s}_t \leftarrow \mathbf{s}_t^{'}(\mathcal{J}_t), \quad \text{where} \quad \mathcal{J}_t = \arg\max_{|\mathcal{J}|=q} \sum_{j\in\mathcal{J}} s_{t,j}. \tag{12}$$

A larger $s_{t,j}$ indicates that the corresponding basis vector has been used more consistently in recent frames. Therefore, we retain $q$ most active basis vectors as the updated basis. In this way, the basis remains a compact yet informative semantic representation with bounded size while evolving online.

### 3.2.2 Dynamic Visual Memory Bank

As the video stream evolves, the number of visual tokens grows linearly over time. To satisfy a fixed token budget, we maintain a dynamic visual memory bank $\mathbf{M}_t \in \mathbb{R}^{b_t\times d}$ at the $t$-th time step. which maximizes the information preserved in the video stream. Here, $b_t \le b$, $b_t$ is the number of tokens stored in $\mathbf{M}_t$ and $b$ is the fixed budget of the memory bank. The memory bank $\mathbf{M}_t$ stores a compact set of visual tokens that are either informative or temporally recent from the observed video stream and is updated online whenever a new frame arrives.

At the $t$-th time step, we first insert all tokens in $\mathbf{X}_t$ into the previous memory bank $\mathbf{M}_{t-1}$, yielding a temporary memory bank $\mathbf{M}_t' = [\mathbf{M}_{t-1}, \mathbf{X}_t]$. If $|\mathbf{M}_t'| > b$, we compress $\mathbf{M}_t'$ guided by the online semantic basis $\mathbf{B}_t$.

Specifically, for each token $\mathbf{X}_i^{'} \in \mathbf{M}_t'$, we first compute its redundancy score $\tilde{e}_i$ according to the residual between $\mathbf{X}_i^{'}$ and its projection onto $\mathbf{B}_t$, similar to Eq.(7):

$$\tilde{e}_i = \frac{e_i - e_{\min}}{e_{\max} - e_{\min} + \epsilon}, \quad \text{where} \quad e_i = \left\| x_i - x_i\mathbf{B}_t^{\top}\mathbf{B}_t \right\|_2, \tag{13}$$

Here, $e_{\min}$ and $e_{\max}$ are the minimum and maximum redundancy scores among all tokens in $\mathbf{M}_t'$, respectively, and $\epsilon$ is a small constant to avoid division by zero.

In addition to redundancy, we also incorporate a temporal recency prior to encourage the memory bank to adapt smoothly to the evolving video stream. Therefore, for each token $\mathbf{X}_i^{'} \in \mathbf{M}_t^{'}$, we compute the final retention score as $f_i$ :

$$f_i = (1-\alpha)\tilde{e}_i + \alpha\tilde{\tau}_i, \quad \text{where} \quad \tilde{\tau}_i = \frac{t_i + 1}{t + 1}, \tag{14}$$

Here, $t_i$ denotes the timestamp of the source frame index of token $\mathbf{X}_i^{'}$ and $\alpha$ balances the redundancy score and the temporal recency score.

We then retain $b$ tokens with the highest scores from $\mathbf{M}_t'$ to form the updated memory bank $\mathbf{M}_t$:

$$\mathbf{M}_t \leftarrow \mathbf{M}_t^{'}(\mathcal{O}_t), \quad \text{where} \quad \mathcal{O}_t = \arg\max_{|\mathcal{O}|=b} \sum_{i\in\mathcal{O}} f_i. \tag{15}$$

Table 1: Comparison with existing methods on two streaming benchmarks using LLaVA-OneVision. The best and second-best results are in **bold** and underline respectively. **Keep / Input** denotes the numbers of candidate tokens and tokens actually fed into the LLM for inference, respectively. Some methods retrieve a subset of tokens from the candidate for inference. $x \times f$ denotes the number of tokens retained per frame.

| Method | Keep / Input Tokens | OVO-Bench real-time | | | | | | | StreamingBench real-time | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | OCR | ACR | ATR | STU | FPD | OJR | Avg. | OP | CR | CS | ATP | EU | TR | PR | SU | ACP | CT | Avg. |
| LLaVA-OneVision | 12k/12k | 67.1 | **61.5** | 74.1 | 45.5 | 69.3 | 65.2 | 63.8 | 80.4 | 74.2 | 76.0 | 80.7 | 72.7 | 71.7 | 67.6 | 65.5 | 65.7 | 45.1 | 71.1 |
| VisionZip | 50×f/50×f | 68.5 | 55.0 | **77.6** | 48.3 | 74.3 | 65.2 | 64.8 | 79.3 | **83.6** | 79.7 | 80.3 | 68.4 | 72.1 | 72.2 | 62.4 | 67.7 | 41 | 71.6 |
| ReKV | 196×f/12k | – | – | – | – | – | – | 57.3 | 74.4 | 78.9 | 78.6 | 77.1 | 68.3 | 67.9 | 67.6 | 62.6 | 64.3 | 44.6 | 69.1 |
| LiveVLM | 12k/12k | 67.1 | 60.5 | 70.7 | 46.1 | 71.3 | 61.9 | 62.9 | 81.5 | 78.1 | **83.3** | 79.1 | **69.6** | 74.1 | **75.0** | **69.1** | 67.7 | 40.4 | 72.9 |
| StreamingTom | 50×f/12k | 67.8 | **61.5** | 75.9 | 47.8 | 68.3 | 66.3 | 64.6 | 79.2 | 80.1 | 78.2 | 79.4 | 67.7 | 68.9 | 69.4 | 60.0 | 64.6 | **46.3** | 70.1 |
| **CausalMem (Ours)** | 12k/12k | **71.8** | **61.5** | 76.7 | **48.9** | **76.2** | **66.8** | **67.0** | **82.6** | 79.7 | 83.2 | **83.2** | **69.6** | **78.4** | **75.0** | 67.3 | **70.9** | 39.4 | **74.3** |

Tokens that are not selected into $\mathbf{M}_t$ are directly removed. In this way, $\mathbf{M}_t$ always keeps the fixed budget while preserving tokens that maximize the information of the video stream.

When a question $Q$ arrives at the $t$-th time step, the response $Y$ is generated conditioned on the question and the visual memory bank $\mathbf{M}_t$:

$$p(Y \mid \mathbf{M}_t, Q) = \prod_{i=1}^{L} p(y_i \mid G, \mathbf{M}_t, Q, Y_{<i}). \tag{16}$$

The illustration of CausalMem is given in Fig. 2, and the algorithm is provided in Appendix 1.

# 4 Experiment

## 4.1 Benchmarks and Metrics

We conduct extensive experiments on both streaming and offline video understanding benchmarks to validate the effectiveness of CausalMem. For streaming evaluation, we adopt OVO-Bench Niu et al. (2025) and StreamingBench Lin et al. (2026), both of which evaluate real-time video understanding under streaming settings. Specifically, OVO-Bench focuses on timestamp-aware streaming reasoning, while StreamingBench emphasizes continuous long-form video comprehension. Following prior works, we report results on their real-time subsets. For offline evaluation, we use Video-MME Fu et al. (2025a), LongVideoBench Wu et al. (2024), MLVU Zhou et al. (2025), and LVBench Wang et al. (2025). These benchmarks cover a wide range of video durations and reasoning complexities. Video-MME and LongVideoBench contain videos ranging from several seconds to nearly one hour, while MLVU extends the maximum duration to approximately two hours. LVBench further evaluates extreme long-video understanding with videos ranging from 1 to 2 hours.

## 4.2 Implementation

We mainly compare CausalMem with representative streaming video understanding methods, including VisionZip Yang et al. (2025a) ReKV Di et al. (2025), LiveVLM Ning et al. (2025), StreamingTOM Chen et al. (2026b) and HoliTOM Shao et al. (2025) Following their settings, we apply CausalMem to LLaVA-OneVision-7B Li et al. (2024), Qwen2.5-VL-7B Bai et al. (2025b) and Qwen3-VL-8B Bai et al. (2025a). For streaming benchmarks, all videos are sampled at 0.5 fps using LLaVA-OneVison, while Qwen-VL series both adopt 1 fps. For offline benchmarks, LLaVA-OneVision uses 0.5 fps for videos shorter than 30 minutes and 0.2 fps for the longer. We set the total visual token budget $b$ to $12k$ for LLaVA-OneVision and $6k$ for Qwen-VL series. Additionally, we limit the maximum tokens per frame to 256 for Qwen-VL-series models to prevent excessive token usage. The maximum number of basis vectors is set to $q = 64$, and the maximum number of candidate tokens used for basis update in each frame is set to $m = 8$. The activity smoothing factor $\gamma$ is set to 0.9. In addition, the balancing factor $\alpha$, which controls the trade-off between redundancy and temporal relevance, is set to 0.8 for streaming benchmarks, while the temporal score is disabled for offline evaluation. All experiments are conducted on NVIDIA A800 GPUs.

## 4.3 Quantitative Analysis

**Comparison on Streaming Benchmarks.** We first compare our CausalMem with existing video compression methods on two streaming benchmarks, as reported in Tab. 1. As can be seen,

Table 2: Comparison with existing streaming methods on offline benchmarks using LLaVA-OneVision. Streaming methods, denoted as *, are evaluated in the streaming setting under strict causality.

| Method | Keep / Input Tokens | VideoMME | MLVU | | | | LongVideo Bench | LVBench | Avg |
|---|---|---|---|---|---|---|---|---|---|
| | | | Single detail | Multi detail | Holistic | Avg | | | |
| LLaVA-OneVision | 12k/12k | 58.5 | 68.6 | 39.1 | 79.3 | 64.7 | 56.5 | 38.4 | 54.5 |
| ReKV* | 196×f/12k | 58.3 | 72.9 | 42.2 | 83.1 | 68.5 | 55.8 | 39.8 | 55.6 |
| LiveVLM* | 12k/12k | 59.6 | - | - | - | 68.1 | 56.1 | 39.5 | 55.8 |
| StreamingTOM* | 50×f/12k | 59.9 | 72.3 | 43.3 | 82.1 | 68.2 | 56.4 | 40.5 | 56.3 |
| HoliTOM | 12k/12k | 59.5 | 73.5 | 44.2 | 81.5 | 68.9 | 56.0 | 41.1 | 56.4 |
| ForestPrune | 12k/12k | 59.6 | 73.5 | 45.7 | 82.0 | 69.4 | 56.6 | 41.7 | 56.8 |
| **CausalMem(Ours)*** | 12k/12k | **60.0** | **74.0** | **49.5** | **83.6** | **70.9** | **57.0** | **42.1** | **57.5** |

Table 3: Comparison with advanced MLLMs on streaming and offline benchmarks. * denots tokens number.

| Method | Input | OVO-Bench | StreamingBench | | | | | | | | | | | MLVU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Real-Time | OP | CR | CS | ATP | EU | TR | PR | SU | ACP | CT | Avg | |
| Gemini 1.5 Pro (Team et al., 2024) | – | 69.3 | 79.0 | 80.5 | 83.5 | 79.7 | **80.0** | 84.7 | 77.8 | 64.2 | 72.0 | 48.7 | 75.7 | - |
| GPT-4o (Hurst et al., 2024) | – | 64.5 | 77.1 | 80.5 | 83.9 | 76.5 | 70.2 | 83.8 | 66.7 | 62.2 | 69.1 | 49.2 | 73.3 | 64.6 |
| Dispider (Qian et al., 2025) | 1fps | 54.6 | 74.9 | 75.5 | 74.1 | 73.1 | 74.4 | 59.9 | 76.1 | 62.9 | 62.2 | 45.8 | 67.6 | 61.7 |
| ViSpeak (Fu et al., 2025b) | 1fps | 66.3 | 79.8 | **88.3** | 83.3 | 81.1 | 76.4 | 75.1 | 70.4 | 65.9 | **77.3** | 34.2 | 74.4 | 54.1 |
| TimeChat-Online (Yao et al., 2025) | 1fps | 61.4 | 80.8 | 79.7 | 80.8 | 83.3 | 74.8 | 78.8 | 78.7 | 64.2 | 68.8 | **58.0** | 75.3 | 65.4 |
| LLaVA-OneVision (Li et al., 2024) | 12k* | 63.8 | 79.3 | 76.6 | 76.2 | 81.3 | 75.9 | 71.5 | 71.3 | 66.9 | 63.4 | 43.1 | 71.3 | 64.7 |
| **+CausalMem(Ours)** | 12k* | 67.0 | 82.6 | 79.7 | 83.2 | 83.2 | 69.6 | 78.4 | 75.0 | 67.3 | 70.9 | 39.4 | 74.3 | 70.9 |
| Qwen2.5-VL (Bai et al., 2025b) | 1fps | 63.3 | 78.3 | 80.5 | 79.8 | 82.4 | 75.5 | 80.4 | 74.1 | 62.6 | 67.6 | 51.1 | 73.9 | 70.2 |
| **+CausalMem(Ours)** | 6k* | 67.8 | 80.7 | 78.9 | **87.3** | **85.5** | 79.1 | 84.0 | 81.5 | **70.2** | 69.4 | 43.0 | 76.9 | 72.8 |
| Qwen3-VL (Bai et al., 2025a) | 1fps | 68.3 | **83.4** | 75.0 | 82.5 | 85.2 | 72.8 | **86.2** | 82.4 | 62.9 | 68.9 | 56.9 | 76.6 | 75.0 |
| **+CausalMem(Ours)** | 6k* | **70.1** | 83.2 | 74.6 | 84.9 | 85.3 | 75.7 | 85.0 | **85.1** | 66.7 | 71.5 | 56.7 | **77.8** | **75.5** |

CausalMem consistently improves the baseline models and achieves competitive or superior performance compared with existing methods. Specifically, CausalMem improves the baseline from $63.8\%$ to $67.0\%$ on OVO-Bench and from $71.1\%$ to $74.3\%$ on StreamingBench. Moreover, CausalMem achieves competitive or even better performance than SOTA streaming methods like StreamMem and StreamingTOM. CausalMem surpasses StreamingTOM and LiveVLM by absolute margins of 4.2% and 1.4% in average accuracy on StreamingBench, respectively. Furthermore, CausalMem achieves strong performance across multiple subcategories on both streaming benchmarks. For example, CausalMem outperforms StreamingTOM on STU and FPD subtasks of OVO-Bench by absolute margins of $1.1\%$ and $7.9\%$, respectively, which shows the robustness of CausalMem on diverse streaming tasks. In summary, CausalMem is an effective and training-free method for MLLMs that maintains a compact yet informative memory for streaming video understanding.

**Comparison on Offline Benchmarks.** We also evaluate CausalMem using LLaVA-OneVision on offline benchmarks compared with both streaming and offline methods in Tab. 2. For offline evaluation, CausalMem still follows the streaming processing setting under strict causality, where frames are unknown to future information. The results show that CausalMem also achieves the competitive or even better performance on four offline benchmarks. For instance, CausalMem surpasses StreamingTOM by $3.0\%$ and $1.6\%$ absolutely on MLVU and LVBench, respectively. It demonstrates the robustness of CausalMem across benchmarks targeting different scenarios, *e.g.*, the comprehensive tasks of MLVU and local tasks of LVBench. It further proves that CausalMem maximizes the diversified information of the video stream in the visual memory bank, which makes it perform well in different scenarios.

**Comparison with SOTA Video-MLLMs.** We further apply CausalMem to Qwen2.5-VL and Qwen3-VL, and compare to existing advanced Video-MLLMs on two streaming benchmarks in Tab. 3. We can observe that CausalMem achieves highly competitive performance compared with existing advanced Video-MLLMs, even though it is a plug-and-play method. For instance, CausalMem help Qwen2.5-VL to achieve $67.8\%$ on OVO-Bench, outperforming TimeChat-Online by $6.4\%$ absolutely. It shows that CausalMem effectively strengthens the streaming video understanding ability of MLLMs via maintaining a compact and dynamically updated visual memory bank. Furthermore, CausalMem improves Qwen3-VL to the level of Gemini 1.5 Pro and GPT-4o, *e.g.*, surpassing GPT-4o by $5.6\%$ and $4.5\%$ on OVO-Bench and StreamingBench, respectively.

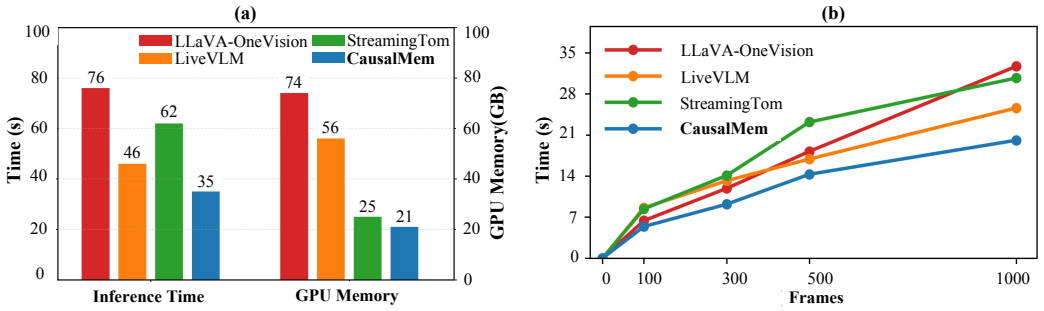



Figure 3: (a) Comparison with existing methods in terms of inference time and GPU memory overhead and the video is sampled at 1FPS. (b) The inference time varies with the number of input frames.

Table 5: Ablation studies of CausalMem. *SB/OVO*, *Hist./Inst.*, and *Red./Temp.* denote StreamingBench/OVOBench, Historical/Instantaneous, and Redundancy/Temporal Scores, respectively. $*$: final setting.

| (a) Basis Size $q$ | | | | (b) Update Tokens $m$ | | | | (c) Activity Score $s$ | | | | (d) Retention Score $f$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $q$ | SB | OVO | MLVU | $m$ | SB | OVO | MLVU | Setting | SB | OVO | MLVU | Setting | SB | OVO | MLVU |
| 16 | 74.0 | 65.5 | 70.7 | 4 | 74.1 | 65.5 | 70.7 | w/ Hist. | 74.1 | 65.3 | 70.4 | w/ Red. | 71.8 | 65.1 | 70.8 |
| 64$*$ | **74.3** | **65.7** | **70.9** | 8$*$ | **74.3** | **65.7** | **70.9** | w/ Inst. | 74.0 | 65.2 | 70.6 | w/ Temp. | 74.0 | 65.3 | 60.2 |
| 512 | 73.7 | 65.2 | 70.1 | 32 | 74.0 | 65.1 | 70.4 | Both$*$ | **74.3** | **65.7** | **70.9** | Both$*$ | **74.3** | **65.7** | **70.9** |

**Analysis of Fine-grained Historical Information Retention.** In Tab. 4, we examine CausalMem's capability in preserving fine-grained historical information. We use the fine-grained analysis, backward tracking or Need-in-Haystack tasks of StreamingBench, OVO and MLVU for evaluations. We observe that CausalMem consistently outperform exsiting methods, *e.g.*, gain $+2.9\%$ than StreamingTom on Backward-Tracing task. These gains show that CausalMem effectively retains both fine-grained and important early visual evidence over long video streams, which further demonstrates the effectiveness of its semantic-basis based memory update mechanism.

**Analysis of Efficiency.** We compare inference time, GPU memory usage, and scalability with existing streaming methods in Fig.3. As shown in Fig.3-(a), CausalMem achieves the lowest GPU memory and inference time. This advantage mainly stems from the fixed-budget memory design. Unlike LLaVA-OneVision, whose visual tokens grow linearly with frames increasing, CausalMem keeps visual memory bank fixed and feeds a compact token sequence to LLM. Moreover, CausalMem avoids the extra overhead of KV-cache retrieval or quantization in existing streaming methods. Fig.3-(b) further shows that CausalMem scales better as the number of frames increases, exhibiting the slowest growth in inference time compared with other methods. These results demonstrate that the fixed-budget memory bank effectively reduces both memory usage and inference time for long streaming videos.

Table 4: Performance comparison on Fine-grained Historical tasks.**TRU/OP/BT** denote Text-Rich Understanding/Object Perception/Backward Tracing and Needle is Needle-in-haystack.

| Method | SB TRU | SB OP | OVO BT | MLVU Needle |
|---|---|---|---|---|
| LLaVA-OneVision | 71.7 | 80.4 | 43.7 | 74.1 |
| +ReKV | 67.9 | 79.3 | 45.3 | 76.5 |
| +LiveVLM | 74.1 | 81.5 | 44.9 | 75.0 |
| +StreamingTom | 68.9 | 79.2 | 43.8 | 75.9 |
| +CausalMem | **78.4** | **82.6** | **46.7** | **78.6** |

**Ablation Study.** We conduct ablation studies of the key designs of CausalMem in Tab. 5. The four subtables respectively correspond to (a) the basis size $q$ of the online semantic basis, (b) the number of tokens $m$ used to update the basis, (c) the basis activity score $s_{t,j}$ in Eq.11, and (d) the token retention score Eq.14. We first ablate the number of basis vectors $q$ in Tab. 5 (a). We can observe that the performance improves as $q$ increases from 16 to 64, then gradually decreases with larger $q$. It shows that a small basis may underrepresent the principal semantics of the video stream, leading to inaccurate redundancy estimation, while an overly large basis may absorb noisy or overly detailed information. Therefore, we adopt a moderate value of $q = 64$ to balance semantic coverage and the discriminative capability of redundancy estimation. Second, we ablate the number of tokens $m$ used

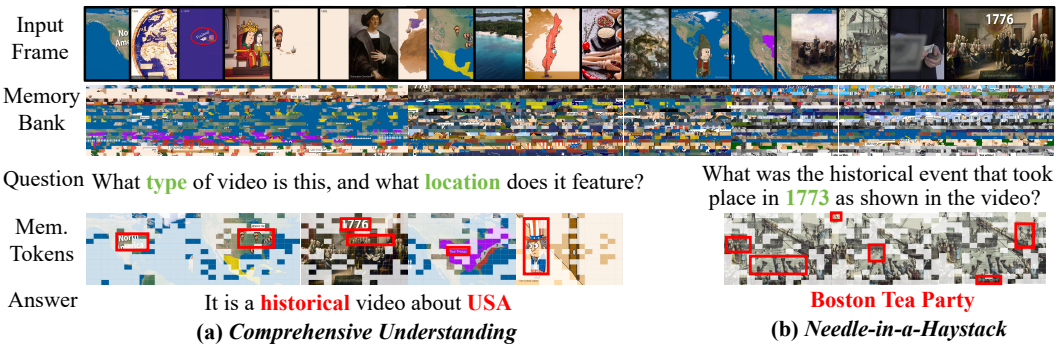



Figure 4: Visualizations of CausalMem. The highlighted tokens indicate the key memories preserved for answering questions. These results show that CausalMem can maintain an informative and compact visual memory bank under a fixed token budget while supporting long-streaming video understanding.

to update the online semantic basis per frame in Tab.5-(b). As shown, the performance first slightly improves when increasing $m$ from 4 to 8, and then gradually decreases with larger $m$. This suggests that a moderate number of tokens provides sufficient new semantics for basis update, while too few tokens may make it difficult to cover the new semantics in a new frame, and too many tokens may introduce noisy or overly fine-grained information.

Furthermore, we analyze the effect of the *activity score* of the semantic basis in Eq. equation 11 , as shown in Tab. 5-(c). We can observe that removing score leads to suboptimal performance. This suggests that relying only on historical activity makes the semantic basis difficult to adapt to the evolving video stream, while using only the instantaneous activity of the current frame may cause the basis to rapidly forget historical semantics. In contrast, combining historical activity with current-frame instantaneous activity enables the basis to adapt to new semantics while preserving previously observed semantics. We also analyze the effect of *Token Retention Score* of Eq. equation 14 in Tab. 5-(d). Without the redundancy score, CausalMem still obtains competitive performance on streaming benchmarks, mainly due to the strong temporal prior of these benchmarks ( Lin et al. (2026); Niu et al. (2025)). However, its performance drops significantly on the offline benchmark MLVU, showing that the redundancy score is crucial for maintaining globally informative and semantically diverse tokens. By combining both scores, CausalMem achieves the best overall performance, demonstrating its complementary effects in maintaining an informative and temporal recency memory bank.

### 4.4 Qualitative Analysis

Fig. 4 visualizes CausalMem maintaining the visual memory bank during streaming video processing. Under a fixed memory budget, CausalMem adaptively preserves informative visual tokens while discarding redundant regions. The retained tokens capture salient objects, distinctive scenes, and newly emerging semantic content, enabling effective and efficient long streaming video understanding. In Fig. 4, CausalMem supports both comprehensive understanding of overall video content and needle-in-a-haystack understanding of fine-grained historical events. The highlighted tokens demonstrate that the preserved memory contains sufficient evidence for accurate question answering, achieving a favorable balance between memory compactness and semantic preservation.

## 5 Conclusion

In this paper, we study the efficient streaming video understanding from the perspective of a dynamic and fixed-budget visual memory bank, and propose a novel and training-free approach for MLLMs, termed CausalMem. To build such a visual memory bank, CausalMem introduces an online and evolving semantic basis to estimate the principal semantics of streaming tokens, based on which the memory tokens in the bank are written and updated. Extensive experiments on two representative MLLMs and a set of streaming and offline benchmarks show its advantages to existing methods, and also confirm its superiority in semantic preservation for streaming videos.

### AI USE STATEMENT

We used generative AI tools only for organizing and restructuring existing code during the research process. Generative AI tools were not used for code generation, research ideation, methodology development, experimental design, data analysis, manuscript writing, or editing. All research ideas, technical contributions, experimental results, and the final manuscript were developed and prepared by the authors. We have reviewed all AI-assisted outputs and take full responsibility for the final content of this work.

## REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

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.

Tao Chen, Shaobo Ju, Qiong Wu, Chenxin Fang, Kun Zhang, Jun Peng, Hui Li, Yiyi Zhou, and Rongrong Ji. Towards effective and efficient long video understanding of multimodal large language models via one-shot clip retrieval. *arXiv preprint arXiv:2512.08410*, 2025.

Tao Chen, Kun Zhang, Qiong Wu, Xiao Chen, Chao Chang, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Scaling the long video understanding of multimodal large language models via visual memory mechanism. *arXiv preprint arXiv:2603.29252*, 2026a.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

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, 2026b.

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. *arXiv preprint arXiv:2503.00540*, 2025.

Ziyang Fan, Keyu Chen, Ruilong Xing, Yulin Li, Li Jiang, and Zhuotao Tian. Flashvid: Efficient video large language models via training-free tree-based spatiotemporal token merging. *arXiv preprint arXiv:2602.08024*, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025a.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Yi-Xing Peng, Kun-Yu Lin, Xihan Wei, Jian-Fang Hu, Xiaohua Xie, and Wei-Shi Zheng. Vispeak: Visual instruction feedback in streaming videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21778–21788, 2025b.

Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

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

Shaobo Ju, Baiyang Song, Tao Chen, Jiapeng Zhang, Qiong Wu, Chao Chang, HuaiXi Wang, Yiyi Zhou, and Rongrong Ji. Forestprune: High-ratio visual token compression for video multimodal large language models via spatial-temporal forest modeling. *arXiv preprint arXiv:2603.22911*, 2026.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Yulin Li, Haokun Gui, Ziyang Fan, Junjie Wang, Bin Kang, Bin Chen, and Zhuotao Tian. Less is more, but where? dynamic token compression via llm-guided keyframe prior. *arXiv preprint arXiv:2512.06866*, 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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pp. 947–952. IEEE, 2019.

Zhenyu Ning, Guangda Liu, Qihao Jin, Wenchao Ding, Minyi Guo, and Jieru Zhao. Livevlm: Efficient online video understanding via streaming-oriented kv cache and retrieval. *arXiv preprint arXiv:2505.15269*, 2025.

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

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

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025.

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Holitom: Holistic token merging for fast video large language models. *arXiv preprint arXiv:2505.21334*, 2025.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26160–26169, 2025.

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

Bo Tong, Bokai Lai, Yiyi Zhou, Gen Luo, Yunhang Shen, Ke Li, Xiaoshuai Sun, and Rongrong Ji. Flashsloth: Lightning multimodal large language models via embedded visual compression. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14570–14581. IEEE, 2025.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025.

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

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

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 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.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, et al. Timechat-online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10807–10816, 2025.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22128–22136, 2025.

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

Zheyu Zhang, Ziqi Pang, Shixing Chen, Xiang Hao, Vimal Bhat, and Yu-Xiong Wang. One token per highly selective frame: Towards extreme compression for long video understanding. *arXiv preprint arXiv:2604.14149*, 2026.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025.

# A APPENDIX

## A.1 ORTHONORMALITY OF THE ONLINE SEMANTIC BASIS.

We represent the online semantic basis as $\mathbf{B}_t \in \mathbb{R}^{q_t \times d}$, where each row is a semantic basis vector. We show that the update preserves the row-orthonormality of $\mathbf{B}_t$, i.e.,

$$\mathbf{B}_t \mathbf{B}_t^\top = \mathbf{I}_{q_t}.$$

At initialization, $\mathbf{B}_1$ is constructed from the retained right singular vectors of $\mathbf{X}_1$. Since right singular vectors are orthonormal, we have

$$\mathbf{B}_1 \mathbf{B}_1^\top = \mathbf{I}_{q_1}.$$

Assume that at time step $t-1$, the basis is row-orthonormal:

$$\mathbf{B}_{t-1} \mathbf{B}_{t-1}^\top = \mathbf{I}_{q_{t-1}}.$$

Then

$$\mathbf{P}_{t-1} = \mathbf{B}_{t-1}^\top \mathbf{B}_{t-1}$$

is the orthogonal projector onto the subspace spanned by the rows of $\mathbf{B}_{t-1}$. Therefore,

$$\mathbf{P}_{t-1}^\perp = \mathbf{I}_d - \mathbf{B}_{t-1}^\top \mathbf{B}_{t-1}$$

is the projector onto its orthogonal complement.

For the candidate residual matrix $\mathbf{R}_t^{cand}$, we remove the components already represented by the current basis:

$$\hat{\mathbf{R}}_t = \mathbf{R}_t^{cand} \left(\mathbf{I}_d - \mathbf{B}_{t-1}^\top \mathbf{B}_{t-1}\right).$$

Thus,

$$\begin{aligned}
\hat{\mathbf{R}}_t \mathbf{B}_{t-1}^\top &= \mathbf{R}_t^{cand} \left(\mathbf{I}_d - \mathbf{B}_{t-1}^\top \mathbf{B}_{t-1}\right) \mathbf{B}_{t-1}^\top \\
&= \mathbf{R}_t^{cand} \mathbf{B}_{t-1}^\top - \mathbf{R}_t^{cand} \mathbf{B}_{t-1}^\top \mathbf{B}_{t-1} \mathbf{B}_{t-1}^\top \\
&= \mathbf{R}_t^{cand} \mathbf{B}_{t-1}^\top - \mathbf{R}_t^{cand} \mathbf{B}_{t-1}^\top \left(\mathbf{B}_{t-1} \mathbf{B}_{t-1}^\top\right) \\
&= \mathbf{R}_t^{cand} \mathbf{B}_{t-1}^\top - \mathbf{R}_t^{cand} \mathbf{B}_{t-1}^\top = \mathbf{0}.
\end{aligned}$$

Hence, every row of $\hat{\mathbf{R}}_t$ lies in the orthogonal complement of the current basis.

We then apply QR orthonormalization to obtain new row-orthonormal basis vectors $\bar{\mathbf{Q}}_t \in \mathbb{R}^{r_t \times d}$, where $r_t \leq k$, such that

$$\bar{\mathbf{Q}}_t \bar{\mathbf{Q}}_t^\top = \mathbf{I}_{r_t}.$$

Since the rows of $\bar{\mathbf{Q}}_t$ are linear combinations of the rows of $\hat{\mathbf{R}}_t$, they remain orthogonal to the previous basis:

$$\bar{\mathbf{Q}}_t \mathbf{B}_{t-1}^\top = \mathbf{0}.$$

The updated basis before pruning is obtained by concatenating along the basis-vector dimension:

$$\mathbf{B}_t' = \begin{bmatrix} \mathbf{B}_{t-1} \\ \bar{\mathbf{Q}}_t \end{bmatrix}.$$

Therefore,

$$\begin{aligned}
\mathbf{B}_t' \mathbf{B}_t'^\top &= \begin{bmatrix} \mathbf{B}_{t-1} \\ \bar{\mathbf{Q}}_t \end{bmatrix} \begin{bmatrix} \mathbf{B}_{t-1}^\top & \bar{\mathbf{Q}}_t^\top \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{B}_{t-1} \mathbf{B}_{t-1}^\top & \mathbf{B}_{t-1} \bar{\mathbf{Q}}_t^\top \\ \bar{\mathbf{Q}}_t \mathbf{B}_{t-1}^\top & \bar{\mathbf{Q}}_t \bar{\mathbf{Q}}_t^\top \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I}_{q_{t-1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{r_t} \end{bmatrix}.
\end{aligned}$$

Thus, $\mathbf{B}_t'$ is row-orthonormal. If the number of basis vectors exceeds the budget $q$, we retain a subset of rows according to the activity scores. Since any subset of a row-orthonormal matrix remains row-orthonormal, the pruned basis $\mathbf{B}_t$ also satisfies

$$\mathbf{B}_t \mathbf{B}_t^\top = \mathbf{I}_{q_t}.$$

By induction, the online semantic basis remains row-orthonormal at every time step.

### A.2 ALGORITHM

**Algorithm 1** CausalMem$(\cdot)$

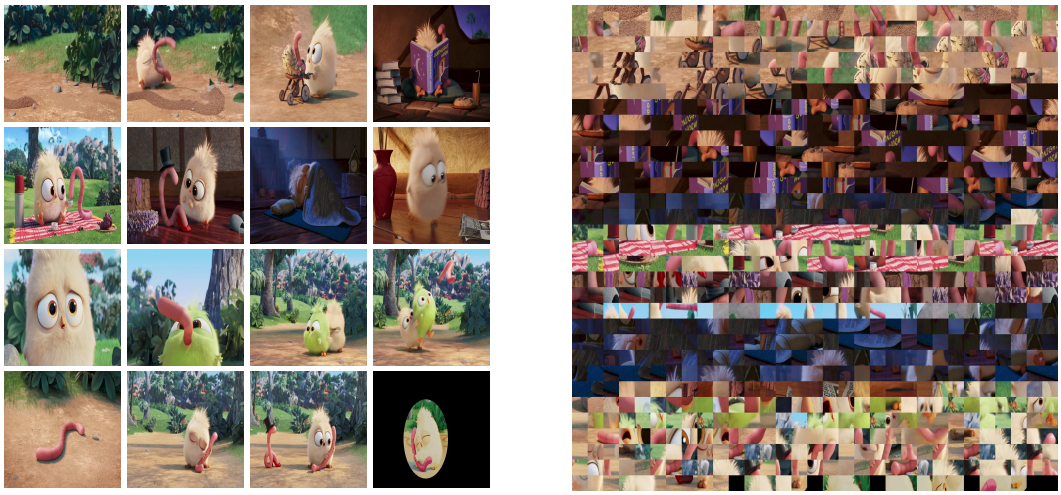
**Require:** Streaming video $\mathcal{V} = \{I_t\}_{t=1}^{T}$, question $Q$, MLLM generator $G$, memory budget $b$, basis budget $q$, basis update size $m$, balance factor $\alpha$, decay factor $\gamma$
**Ensure:** Answer $Y$, memory bank $\mathbf{M}_T$

$\mathbf{M}_0 \leftarrow \emptyset$, $\mathbf{B}_0 \leftarrow \emptyset$, $\mathbf{s}_0 \leftarrow \emptyset$
**for** $t = 1$ to $T$ **do**
  $\mathbf{X}_t \leftarrow \Phi(I_t)$
  **if** $\mathbf{B}_{t-1} = \emptyset$ **then**
    $\mathbf{B}_t \leftarrow \textsc{TopSVD}(\mathbf{X}_t, q_1), \quad \mathbf{s}_t \leftarrow \mathbf{1}_{q_1}$
  **else**
    $(\mathbf{R}_t, \mathbf{e}_t) \leftarrow \textsc{ResidualScore}(\mathbf{X}_t, \mathbf{B}_{t-1})$
    $\mathbf{X}_t^{\text{cand}} \leftarrow \textsc{TopK}(\mathbf{X}_t, \mathbf{e}_t, m)$
    $\bar{\mathbf{Q}}_t \leftarrow \textsc{Orthogonalize}(\mathbf{X}_t^{\text{cand}}, \mathbf{B}_{t-1})$
    $\mathbf{B}'_t \leftarrow [\mathbf{B}_{t-1}; \bar{\mathbf{Q}}_t], \quad \mathbf{s}'_t \leftarrow [\mathbf{s}_{t-1}; \mathbf{1}]$
    $\mathbf{s}'_t \leftarrow \textsc{UpdateActivity}(\mathbf{s}'_t, \mathbf{B}'_t, \mathbf{X}_t, \gamma)$
    $(\mathbf{B}_t, \mathbf{s}_t) \leftarrow \textsc{KeepTopBasis}(\mathbf{B}'_t, \mathbf{s}'_t, q)$
  **end if**
  $\mathbf{M}'_t \leftarrow [\mathbf{M}_{t-1}; \mathbf{X}_t]$
  **if** $|\mathbf{M}'_t| > b$ **then**
    $\mathbf{f}_t \leftarrow \textsc{RetentionScore}(\mathbf{M}'_t, \mathbf{B}_t, \alpha, t)$
    $\mathbf{M}_t \leftarrow \textsc{TopK}(\mathbf{M}'_t, \mathbf{f}_t, b)$
  **else**
    $\mathbf{M}_t \leftarrow \mathbf{M}'_t$
  **end if**
**end for**
$Y \sim G(\mathbf{M}_T, Q)$
**return** $Y, \mathbf{M}_T$

### A.3 Additional Visualization Results

We provide qualitative visualizations to illustrate how the proposed method processes long video streams and maintains informative visual content.

The results show that our method effectively preserves informative and semantically salient content while maintaining a compact representation of the video stream.



**Video frames** **Memory bank**

(a) Video 1




**Video frames** **Memory bank**

(b) Video 2

Figure 5: Qualitative visualization of raw video frames and the corresponding memory bank representations.

### A.4 Robustness of Residual-Based Novelty Estimation to Non-Semantic Perturbations

In the causal streaming setting, future user queries are unavailable. We use the residual norm to estimate the semantic novelty and complementary information of an incoming visual token with respect to the observed video history. Specifically, a large residual indicates that the token cannot be sufficiently represented by the current online semantic basis, suggesting that it contains information that is less redundant with previously observed content. To examine whether non-semantic variations could spuriously produce large residuals and dominate the memory update process, we conduct controlled robustness experiments on OVO-Bench under four types of perturbations: Gaussian noise, motion blur, background anomalies, and feature-level shifts.

For Gaussian noise, we randomly inject Gaussian noise into consecutive sampled frames, with the corrupted-frame ratio ranging from 10% to 60%. For each corrupted frame, the noisy region covers $1/8$ of the spatial area. For motion blur, we simulate camera-motion-induced appearance varia-

Table 6: OVO-Bench accuracy (%) under different non-semantic perturbations. The perturbation ratio denotes the proportion of consecutive sampled frames affected by each perturbation.

| Perturbation | 0% | 10% | 20% | 30% | 40% | 50% | 60% |
|---|---|---|---|---|---|---|---|
| Gaussian noise | **65.7** | 65.7 | 65.6 | 65.4 | 65.4 | 65.1 | 64.6 |
| Motion blur | **65.7** | 65.7 | 65.4 | 65.4 | 65.4 | 65.3 | 65.2 |
| Background anomalies | **65.7** | 65.9 | 65.8 | 65.7 | 65.6 | 65.5 | 65.4 |
| Feature-level shifts | **65.7** | 65.3 | 65.2 | 65.2 | 65.1 | 64.7 | 64.3 |

tions by applying horizontal linear blur with a kernel size of 9 to 10%–60% of consecutive sampled frames. For background anomalies, we inject random textures into the four corner regions of 10%–60% of consecutive sampled frames. Each corner region occupies $1/16$ of the frame area, resulting in visually unusual but task-irrelevant regions without modifying the main semantic content. Finally, to evaluate feature shifts directly at the representation level, we inject Gaussian noise into the encoded visual features of 10%–60% of consecutive sampled frames, again covering $1/8$ of the spatial region for each affected frame.

The results show that CausalMem is relatively stable under a range of non-semantic perturbations. Under Gaussian noise, the accuracy remains at 65.1% even when 50% of consecutive sampled frames are corrupted, corresponding to a decrease of only 0.6 percentage points from the unperturbed setting. Motion blur produces similarly small changes, with the accuracy decreasing from 65.7% to 65.2% at a 60% perturbation ratio. Background anomalies have a negligible effect, with the accuracy remaining at 65.4% even when 60% of consecutive sampled frames contain the injected anomalous regions. These results suggest that CausalMem does not simply prioritize visually unusual pixel-level patterns when updating the memory.

Feature-level shifts introduce a stronger disturbance because they directly modify the visual features used by the residual-based memory update. Nevertheless, the method remains relatively stable under moderate perturbations: the accuracy drops by 0.4 and 0.5 percentage points when 10% and 20% of consecutive sampled frames are perturbed, respectively. Even at a 60% perturbation ratio, the accuracy remains at 64.3%, corresponding to a 1.4 percentage-point decrease.

Overall, these experiments provide additional evidence that residual-based novelty estimation is not easily dominated by several common non-semantic variations. This supports the use of residual magnitude as a causal, query-agnostic signal for maintaining diverse and informative visual content within a fixed memory budget.