# RetentiveKV: State-Space Memory for Uncertainty-Aware Multimodal KV Cache Eviction

**Sihao Liu[1], YuFan Xiong[2], Zhonghua Jiang[1], Zhaode Wang[2], chengfei lv[2], Shengyu Zhang[3] †**
[1]Zhejiang University
[2]Alibaba [3] Shanghai Institute for Advanced Study of Zhejiang University
{lyosihao,jiangzhonghua,sy_zhang}@zju.edu.cn
xiomg@webmail.hzau.edu.cn
{zhaode.wzd,chengfei.lcf}@taobao.com

## Abstract

Multimodal Large Language Models face severe challenges in computational efficiency and memory consumption due to the substantial expansion of the visual KV cache when processing long visual contexts. Existing KV cache compression methods typically rely on the "persistence of importance" hypothesis to prune tokens. However, this approach proves fragile in multimodal settings due to two key issues: 1) Visual tokens display "deferred importance," initially exhibiting low salience but becoming pivotal during later decoding, which can lead to premature eviction. 2) Discrete pruning disrupts the inherent spatial continuity of visual cues. To address these challenges, we propose RetentiveKV, an entropy-driven KV cache optimization method that reformulates KV eviction from "discrete context truncation" to "continuous memory evolution" based on State Space Models. Our method leverages information entropy to quantify the information potential of low-attention tokens and integrates tokens scheduled for eviction into a continuous state space through entropy-guided state transitions, enabling their dynamic reactivation when semantic relevance arises during subsequent decoding. Extensive experiments on multimodal benchmarks demonstrate that RetentiveKV achieves 5.0 × KV cache compression and 1.5 × decoding acceleration.

## 1 Introduction

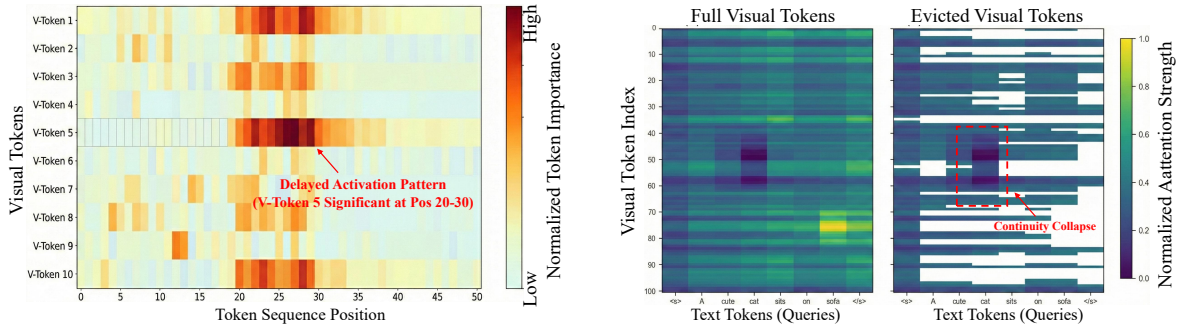
Multimodal Large Language Models (MLLMs) have emerged as the dominant paradigm for understanding and generating cross-modal content. However, the increasing demand for processing long visual contexts and high-resolution inputs results in a substantial expansion of visual tokens and the KV cache. This imposes severe challenges on computational efficiency and memory consumption during inference. Consequently, the efficient optimization of token sequences and the KV cache has become critical for accelerating autoregressive inference in MLLMs.

Prior research in Large Language Models (LLMs) has extensively explored importance-centric KV cache compression methods. These approaches generally leverage token-level importance metrics to selectively evict KV pairs with minimal contributions, thereby mitigating memory consumption and computational complexity. While existing methods demonstrate considerable promise in text-only settings, they prove fragile in multimodal contexts: 1) **Deferred Importance Matter**: Conventional KV compression relies on the "persistence of importance" hypothesis (Zhang et al., 2023; Liu et al., 2023), assuming that tokens with high initial attention remain critical throughout decoding. This assumption is widely inherited by multimodal methods (Chen et al., 2024a; Lin et al., 2025). However, we observe a distinct "Deferred Importance" phenomenon: visual tokens often exhibit low initial salience but become pivotal at later timesteps. As shown in Figure 1(a), attention distribution in Visual Question Answering (VQA) tasks is temporally non-uniform. Early decoding prioritizes the semantic comprehension of the textual query, while the attention mechanism only pivots toward the cross-modal dependencies of visual tokens when generating key responses requiring specific visual attributes. 2) **Visual Continuity Collapse**: Visual tokens are characterized by intrinsic spatial continuity and strong inter-patch correlations. As illustrated in Figure 1(b), the discrete eviction of KV pairs inevitably fragments these continuous representations. While existing multimodal methods (Li et al., 2025a; Jiang et al., 2025a) attempt to mitigate this by modulating eviction ratios based on modality preferences, this strategy essentially relies on quantitative reallocation. It fails to fundamentally address the critical issue of spatial discontinuity, as the underlying eviction

†Corresponding author.



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 18717–18727




(a) Temporal Importance Heatmap of Visual Token.

(b) Impact of Token Eviction on Cross-model Attention.

Figure 1: (a) Visualization of Deferred-Critical Tokens: The presence of delayed activation patterns, where visually salient features remain dormant during early decoding phases, defies the static heavy-hitter assumption. This temporal misalignment triggers erroneous eviction in importance-based pruning. (b) Visualization of Pruning-Induced Spatial Discontinuity: We compare baseline and pruned attention maps. The instability of visual signals triggers excessive pruning (white voids). This fragmentation disrupts the spatial continuity of visual representations.

mechanism remains structure-agnostic.

In this paper, drawing inspiration from State Space Models (SSMs), we transform KV cache eviction from "discrete context truncation" to "continuous memory evolution". Instead of permanently discarding tokens with low attention scores, our approach assimilates them into a continuous state space governed by information entropy. This mechanism preserves visual tokens exhibiting high uncertainty during early decoding stages and updates them continuously as decoding progresses, all while incurring only $O(1)$ memory overhead. Furthermore, by replacing the pair-token softmax kernels of standard attention with recursive convolution kernels derived from state transitions, SSMs inherently maintain the spatial continuity of visual cues, mitigating visual continuity collapse.

To this end, we propose RetentiveKV, a framework that reformulates KV cache eviction via entropy-driven SSMs. Specifically, RetentiveKV comprises three innovations for incorporating evicted information into a dynamic state space: 1) Entropy-guided KV Retention Estimator: Traditional KV eviction typically adheres to the "persistence of importance" principle, which assumes that tokens exhibiting high attention scores in current steps will remain critical for subsequent decoding. In contrast, our method incorporates an additional "prospective uncertainty" principle, designed to estimate the "undetermined cross-modal potential" of tokens exhibiting low attention scores. We leverage the entropy of the attention distribution from textual to visual tokens to assess this potential. High entropy indicates that a textual token attends diffusely across visual tokens, implying that the visual context holds a high degree of uncertainty and thus potential relevance for subsequent decoding steps. 2) Entropy-Guided State Transition: This mechanism reformulates the discrete KV eviction process into a continuous state evolution, which leverages selective state to approximate the per-token interactions of self-attention. Compared to weighted merging strategies, this formulation preserves the ability to selectively attend to specific historical segments while naturally incorporating positional and spatial information through recursive state updates. 3) State Modeling and Retrieval: Inspired by the hierarchical nature of human memory (Baddeley, 2000), we propose a dual-state architecture to resolve the conflict between visual continuity and semantic discreteness. Specifically, we design Visual-Dominant States to preserve the spatial topology of visual patches, and Recall-Oriented States to capture long-term semantic dependencies. These states are continuously updated via entropy-guided transitions. During inference, a Query-Conditioned Retrieval mechanism dynamically queries these states, retrieving and reactivating "deferred-critical" information only when semantic relevance to the current query is detected.

Our contributions are summarized as follows:

1. We reformulate KV eviction as continuous state evolution, assimilating 'deferred-critical' tokens with undetermined potential into a continuous state representation to preserve spatial continuity.

2. We propose RetentiveKV, which leverages

attention entropy to quantify future relevance and compresses evicted tokens into selective states, enabling their dynamic reactivation for decoding.

3. We conduct extensive experiments on multimodal benchmarks, demonstrating that RetentiveKV achieves substantial reductions in memory consumption and computational overhead.

## 2 Related Work

### 2.1 KV Cache Eviction

The autoregressive decoding mechanism of LLMs necessitates caching Key and Value (KV) states to avoid redundant re-computation. However, the linear growth of the KV cache with respect to sequence length imposes a significant memory bottleneck. To mitigate this, importance-centric eviction strategies (Li et al., 2025b; Jiang et al., 2025b) leverage the "heavy-hitter" phenomenon, observing that a small subset of tokens accumulates the majority of attention mass. While these seminal works effectively exploit attention sparsity to prune redundant KV pairs, subsequent research has refined the definition of token importance. For instance, StreamingLLM (Xiao et al., 2023) uncovers the "attention sink" phenomenon, advocating for the persistence of initial tokens alongside the most recent local context. Extending the scope to MLLMs, the memory bottleneck is further compounded by the expansion of visual tokens generated from high-resolution images and long-form videos. Furthermore, SnapKV (Li et al., 2024) introduces a voting mechanism that clusters key context chunks exhibiting high responsiveness to the prompt. In this domain, FastV (Chen et al., 2024a) investigates the layer-wise distribution of visual importance, uncovering the phenomenon of "visual attention inefficiency". It implements a depth-adaptive eviction policy that discards visual KV states in deeper layers. Similarly, LOOK-M (Wan et al., 2024) introduces a modality-aware compression framework, which utilizes pivotal merging to aggregate spatially correlated visual tokens into compact representations. Meda (Wan et al., 2025) dynamically allocates KV cache budgets with information entropy across different transformer layers. SAINT (Jeddi et al., 2025) employs a graph-theoretic similarity metric to prune redundant visual tokens within early network layers.

### 2.2 State Space Models

State Space Models originate from classical control theory and signal processing, where they model continuous systems through latent state evolution. However, applying vanilla SSMs to deep learning was historically hindered by prohibitive computational costs and the difficulty of handling long-range dependencies. HiPPO (High-Order Polynomial Projection Operators) (Gu et al., 2020) pioneered a theoretical framework to solve this by projecting continuous signals onto orthogonal polynomial bases, mathematically guaranteeing the optimal retention of history. Building on this, S4 (Gu et al., 2021) introduced a parameterization of the state matrix $\mathbf{A}$ as a diagonal plus low-rank structure. This innovation allowed the recurrence to be computed efficiently via parallel scans in the GPU, reducing the complexity from quadratic $O(L^2)$ to linear or log-linear $O(L \log L)$. The most significant recent advancement is the selective SSM (Gu and Dao, 2024; Sun et al., 2023). Unlike prior models, where parameters are time-invariant, selective SSM makes these parameters input-adaptive. This allows the model to selectively propagate or forget information based on the current token, effectively solving the "selection mechanism" problem that static SSMs faced. Fundamentally, an SSM maps an input sequence $x_t \in \mathbb{R}$ to an output sequence $y_t \in \mathbb{R}$ through an implicit latent state $h_t \in \mathbb{R}^N$. This process can be formulated as:

$$h_{t+1} = \mathbf{A}h_t + \mathbf{B}x_t,$$
$$y_t = \mathbf{C}h_t,$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ represents the state evolution matrix, while $\mathbf{B} \in \mathbb{R}^{N\times 1}$ and $\mathbf{C} \in \mathbb{R}^{1\times N}$ are projection parameters for the input and state.

## 3 Method

### 3.1 SSM-based KV Cache Retention

Standard Transformers rely on the Scaled Dot-Product Attention mechanism, defined as:

$$\text{Atten}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V. \quad (1)$$

This requires maintaining a KV cache that grows linearly with sequence length $L$, resulting in $O(L)$ memory complexity and $O(L^2)$ computational complexity during decoding.

In this work, we leverage the dual computation paradigm of SSM, which reformulates the attention mechanism as SSM. By removing the nonlinear Softmax operation and utilizing the associative property of matrix multiplication, the output computation can be rewritten as:

$$\text{SSM}(Q, K, V) = Q_t \left( \sum_{i=1}^{t} \gamma^{t-i} (K_i^\top V_i) \right), \quad (2)$$

here, we introduce a decay factor $\gamma \in (0, 1]$ (analogous to the diagonal of the matrix $\bar{\mathbf{A}}$ in SSMs) to modulate the retention of historical information. We can thus define a matrix-valued hidden state $S_t \in \mathbb{R}^{d \times d}$, which serves as a compressed representation of the KV cache. The recurrence rule for RetentiveKV is formally defined as:

$$\begin{aligned} h_{t+1} &= \gamma h_t + K_t^\top V_t, \\ O_t &= Q_t S_t. \end{aligned} \quad (3)$$

In this formulation, $S_t$ effectively absorbs the visual tokens $(K_t, V_t)$ into the state space. Unlike the standard KV cache, which appends tokens to a growing list, this formulation updates the state in-place. This ensures that the memory footprint for the visual context remains constant, regardless of the visual resolution or sequence length.

### 3.2 Overview of RetentiveKV

The RetentiveKV framework consists of three important components to incorporate evicted tokens into entropy-driven state spaces: 1) Entropy-Guided KV Retention Estimator, which leverages entropy-based measures to quantify the uncertainty and prospective relevance of KV pairs for future decoding; 2) Entropy-Guided State Transition, which absorbs evicted KV cache into a modality-specific state space and leverages token-level entropy variations to modulate retention and decay for state evolution. 3) Query-Conditioned State Retrieval, which is responsible for selectively recalling task-relevant multimodal information from the state space during autoregressive decoding.

### 3.3 Observation

Recent studies (Shi et al., 2025; Xiong et al., 2025) in the textual modality have demonstrated that information entropy serves as an important indicator for token compression. Tokens with higher entropy typically correspond to critical decision points in autoregressive decoding, reflecting peaks in contextual shifts that influence the likelihood of subsequent states. Building upon these insights, we investigate the applicability of information entropy as a guiding metric for multimodal token compression. In the MLLMs, the entropy dynamics are complicated by the cross-modal interactions between visual and textual modalities. Therefore, we focus on the information entropy of the attention distribution from textual tokens to visual tokens, which we refer to as **Cross-Modal Attention Entropy**. As shown in Figure 2, empirical analysis reveals that importance-centric KV eviction induces a layer-dependent increase in cross-modal attention entropy. Some layers exhibit anomalously elevated entropy compared to their non-evicted counterparts, indicating that these layers experience heightened uncertainty regarding cross-modal alignment after KV cache eviction. The layers exhibiting significant increases in cross-modal attention entropy are mainly located in the middle-to-upper depths of the model. These layers are widely regarded as playing a critical role in abstract semantic modeling and cross-modal alignment. These observations motivate the adoption of cross-modal attention entropy as an important metric for multimodal KV cache compression.

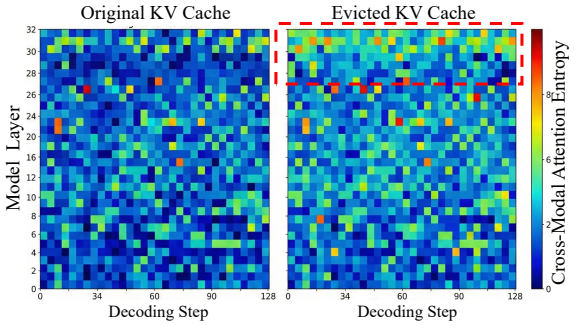



Figure 2: Entropy Shifts under KV Eviction.

### 3.4 Entropy-Guided KV Retention Estimator

Entropy-Guided KV Retention Estimator introduces the cross-modal attention entropy to measure the distributional uncertainty of attention from textual tokens to visual tokens. Let $\alpha_t^{l,i}$ denotes the standard attention score for token $i$ at decoding step $t$ for the $l$-th layer, $p_v(\cdot)$ represents the cross-modal attention scores selected from the $\alpha_t^{l,i}$. The cross-modal attention entropy is defined by the Shannon

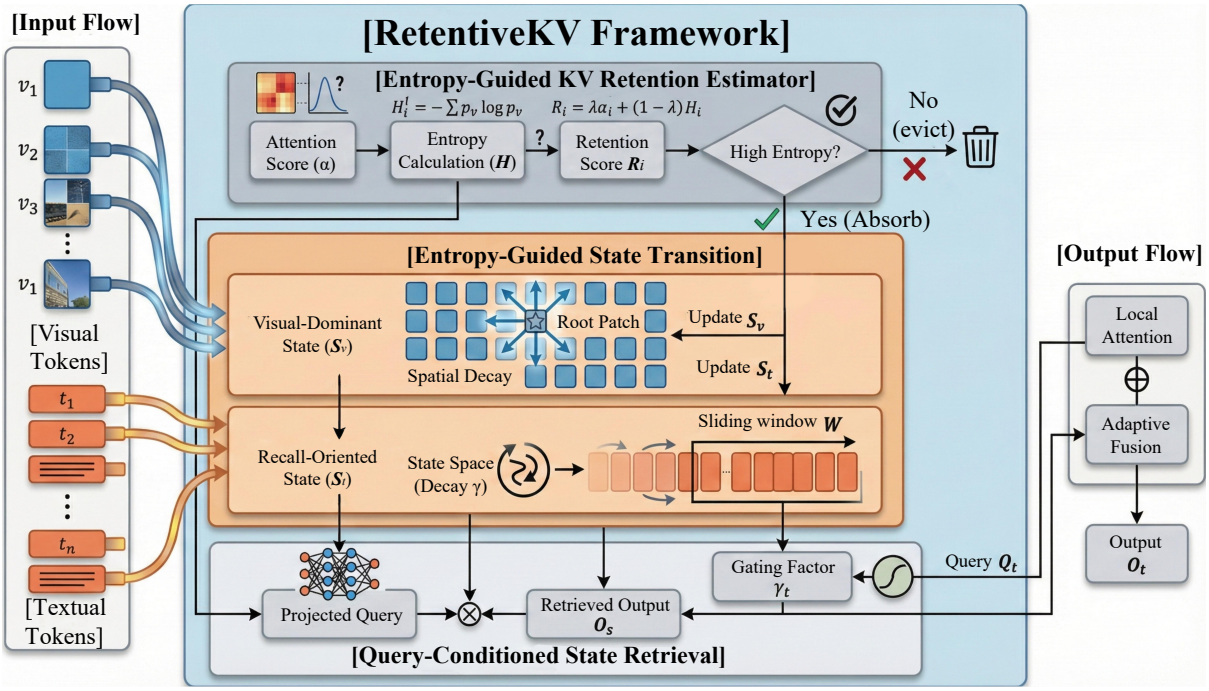



Figure 3: Overview of the proposed RetentiveKV framework. The architecture coordinates efficient long-context multimodal reasoning through three core components: (1) An Entropy-Guided KV Retention Estimator that identifies deferred-critical tokens by analyzing layer-wise entropy shifts ($H_t$) and accumulated attention ($\alpha_t$); (2) An Entropy-Guided State Transition mechanism that absorbs evicted KV pairs into modality-adaptive state space; and (3) A Query-Conditioned State Retrieval module that dynamically fuses retrieved long-term context ($O_s$) with local attention outputs via a learnable gating factor ($\gamma_t$) during autoregressive decoding.

entropy over the attention distribution. Formally:

$$H_t^{l,i} = -p_v(\alpha_t^{l,i}) \cdot \log p_v(\alpha_t^{l,i}). \tag{4}$$

We define a retention score $R^{l,i}$ to decide whether the KV pair of token $i$ should be retained in the state at layer $l$. Our scoring function integrates both the immediate importance and the prospective uncertainty. Formally, the retention score is calculated as a weighted combination of the standardized attention score and the cross-modal attention entropy:

$$R_t^{l,i} = \lambda \alpha_t^{l,i} + (1-\lambda) H_t^{l,i}, \tag{5}$$

where $\lambda \in [0,1]$ balances the contribution of immediate importance and prospective utility. KV pairs with $R_t$ above a retention threshold $\tau$ are absorbed into the continuous state space.

### 3.5 Entropy-Guided State Transition

The objective of Entropy-Guided State Transition is to integrate evicted KV pairs into a continuous state space and govern the evolution of the state space. Differing from standard SSMs with a fixed transition matrix, our method introduces cross-modal attention entropy as a dynamic transition coefficient. Formally, we define the continuous state at decoding step $t$ as $\mathbf{S}_t$. The state transition is formulated as a recursive update equation:

$$\mathbf{S}_t = \mathbf{H}_t \odot \mathbf{S}_{t-1} + \mathbf{A}_t \odot (\mathbf{k}_t^\top \mathbf{v}_t), \tag{6}$$

where $\mathbf{A}$ denotes the absorption matrix that controls the injection of the current evicted KV pair, modulated by the normalized accumulated attention of the current tokens, and $\mathbf{H}$ denotes the retention matrix that determines how much of the accumulated state is decayed, modulated by the token-level information entropy. Formally, let the normalized accumulated attention score and the sigmoid-transformed cross-modal attention entropy of token $i$ be $\alpha_i$ and $\sigma(H_i)$, respectively. We define:

$$\mathbf{H}_t[n,m] = \begin{cases} \sigma(H_t)^{n-m}, & n \geq m \\ 0, & n < m \end{cases}, \tag{7}$$

$$\mathbf{A}_t[n,m] = \begin{cases} \alpha_t^{n-m}, & n \geq m \\ 0, & n < m \end{cases}. \tag{8}$$

The Entropy-Guided State Transition mechanism defined above serves as the foundational KV retention kernel for RetentiveKV. While the above

equation defines the general form of state evolution, its deployment varies across different phases of autoregressive decoding.

### 3.6 Modality-Agnostic Initial State Modeling

The core challenge during the prefill stage lies in the parallel processing of high-resolution visual tokens and long-sequence textual instructions, which requires the model to manage a large-scale multimodal KV cache concurrently. In the prefill stage, we adopt the following dual-state strategies:

#### 3.6.1 Visual-Dominant State ($\mathbf{S}_V$)

Visual tokens possess an inherent two-dimensional spatial topology and continuity. To preserve these geometric priors, we model the visual-dominant state as a spatially-aware State Space. For each input image $\mathcal{I}$, we maintain an independent state $\mathbf{S}_V^{\mathcal{I}}$, which is decomposed into two orthogonal substates representing horizontal and vertical scanning directions. The initialization of the state space is conditioned on the root patch $\mathbf{p}_{x_m,y_m}$ with the maximum accumulated attention score. For a patch located at grid coordinates $(x_n, y_n)$ being evicted from the discrete cache, its integration into the continuous state is modulated by its spatial displacement from the root patch. We define the update of the visual-dominant state as:

$$\begin{aligned}\mathbf{S}_{V,t}^{\mathcal{I}} = \mathbf{H}_t^{|x_n-x_m|+|y_n-y_m|} \odot \mathbf{S}_{V,t-1}^{\mathcal{I}} \\ + \mathbf{A}_t^{|x_n-x_m|+|y_n-y_m|} \odot (\mathbf{k}_t^\top \mathbf{v}_t).\end{aligned} \tag{9}$$

#### 3.6.2 Recall-Oriented State ($\mathbf{S}_T$)

For the textual modality, we maintain a fixed-length working sliding window of size $W$, where the most recent $W$ key-value pairs are preserved in full-precision. These instantaneous memories act as high-resolution semantic anchors, preventing the accumulation of approximation errors induced by early-stage eviction. For tokens that are outside the sliding window or exiting the sliding window during decoding, we identify the textual tokens that are not classified as "Heavy-Hitters" (Zhang et al., 2023) but exhibit high attention entropy and are integrated into the Recall-Oriented State.

### 3.7 Query-Conditioned State Retrieval

To effectively leverage the information retained in the continuous state space during decoding, we introduce a query-conditioned state retrieval mechanism, which conditions on the current query states to selectively re-inject features from the visual-dominant state ($\mathbf{S}_\mathrm{V}$) and the recall-oriented state $\mathbf{S}_\mathrm{T}$ into the attention computation.

At each decoding step $t$, given the current query vector $\mathbf{q}_t$, we retrieve the contextual features from the dual state spaces by projecting the query into the continuous state space. The retrieved value representation $\mathbf{v}^{(t)}$ is defined as:

$$\mathbf{O}_\mathbf{S}^{(t)} = \mathrm{Norm}(\mathbf{q}_t \mathbf{S}_V^{(t-1)} + \gamma_t \cdot \mathbf{q}_t \mathbf{S}_T^{(t-1)}), \tag{10}$$

where $\mathbf{S}_V^{(t-1)}$ denotes the visual-dominant state constructed by concatenating the instance-specific visual states $\mathbf{S}_V^{\mathcal{I}}$ corresponding to input images, $\mathrm{Norm}(\cdot)$ denotes Layer Normalization, which is applied to stabilize the numerical distribution of the fused multimodal features. We introduce an activation gating factor $\gamma_t$, which adaptively controls the contribution of the retrieved state information. Formally, we define the activation gating factor as:

$$\gamma_t = \sigma\left(W_r \cdot H_t + b_r\right), \tag{11}$$

where $H_t$ denotes the average cross-modal attention entropy at decoding step $t$, and $W_r$ and $b_r$ are learnable parameters. The sigmoid function $\sigma(\cdot)$ constrains $\gamma_t \in (0, 1)$, enabling smooth and stable modulation of the retrieved state contribution.

We integrate the retrieved state information with the attention outputs from the working sliding window and the non-evicted KV cache. The final attention output $\mathbf{O}_t$ is expressed as:

$$\mathbf{O}_t = \mathrm{Attn}_{local}(\mathbf{q}_t, \mathbf{K}_{local}, \mathbf{V}_{local}) + \cdot \mathbf{O}_\mathbf{S}^{(t)}. \tag{12}$$

## 4 Experiment

### 4.1 Experiment setting

We evaluate the performance of RetentiveKV across three diverse and representative MLLM architectures: LLaVA-v1.5-7B, which serves as a widely recognized baseline for multimodal instruction tuning; Qwen3-VL-4B and Qwen3-VL-8B, representing state-of-the-art efficiency in general multimodal tasks; To verify the generalizability of our methods, we conduct a comprehensive evaluation across eight benchmarks covering multiple domains: MMMU (Yue et al., 2024) for expert-level reasoning, DocVQA (Mathew et al., 2020) and TextVQA (Singh et al., 2019) for document and scene-text understanding, MathVista (Lu et al., 2024) for mathematical visual reasoning, MMStar

Table 1: Performance comparison of different KV compression (The best results are highlighted in **bold**.)

| Method | TextVQA | DocVQA | MathVista | MMStar | MMMU | BLINK | MMCoQA | ALFRED |
|---|---|---|---|---|---|---|---|---|
| ***LLaVA-v1.5-7B*** | | | | | | | | |
| Full Cache | 62.30 | 61.24 | 60.40 | 62.52 | 43.34 | 64.04 | 35.50 | 16.32 |
| H2O | 60.20 | 56.42 | **59.30** | 61.66 | 42.39 | 62.62 | 28.00 | 14.49 |
| SnapKV | 61.00 | 58.52 | 59.10 | 62.13 | 42.33 | **63.14** | 30.50 | 14.57 |
| SAINT | 60.00 | 57.84 | 58.60 | 61.84 | 42.28 | 62.64 | 28.00 | 14.64 |
| Meda | 60.80 | 57.26 | 59.20 | 62.07 | 42.36 | 63.04 | 28.50 | 14.87 |
| LOOK-M | 60.60 | 56.90 | 59.10 | 62.25 | 42.15 | 62.49 | 29.50 | 14.33 |
| RetentiveKV | **61.20** | **59.05** | **59.30** | **62.42** | **42.45** | 63.12 | **31.50** | **15.82** |
| ***Qwen3-VL-8B*** | | | | | | | | |
| Full Cache | 62.00 | 63.15 | 63.10 | 62.73 | 42.33 | 64.65 | 38.00 | 16.62 |
| H2O | 60.80 | 58.75 | 61.30 | 62.33 | 40.66 | **64.42** | 31.50 | **16.36** |
| SnapKV | 61.30 | 60.12 | 62.20 | 62.36 | 41.38 | 64.31 | 32.50 | 16.12 |
| SAINT | 60.60 | 60.44 | 60.50 | 62.38 | 41.22 | 63.82 | 32.00 | 16.15 |
| Meda | 60.40 | 60.52 | 61.80 | 62.41 | 41.10 | 63.55 | 31.50 | 15.98 |
| LOOK-M | 61.00 | 59.42 | 61.40 | 62.33 | 41.12 | 63.83 | 31.50 | 16.22 |
| RetentiveKV | **61.40** | **62.21** | **62.60** | **62.58** | **41.57** | 64.22 | **33.10** | 16.34 |
| ***Qwen3-VL-4B*** | | | | | | | | |
| Full Cache | 57.70 | 59.97 | 50.40 | 60.66 | 42.11 | 64.95 | 32.50 | 15.98 |
| H2O | 55.00 | 56.03 | 48.00 | 60.36 | 41.05 | 63.55 | 28.00 | 14.67 |
| SnapKV | 55.90 | 57.51 | **49.50** | 60.41 | 41.16 | 63.79 | **29.50** | 15.64 |
| SAINT | 55.90 | 57.42 | 48.30 | 60.21 | 41.16 | 63.48 | 28.50 | 15.19 |
| Meda | 55.60 | 57.51 | 48.80 | 60.38 | 41.38 | 63.62 | 28.50 | 15.26 |
| LOOK-M | 56.10 | 56.64 | 48.50 | 60.25 | 41.32 | 63.21 | 28.00 | 15.32 |
| RetentiveKV | **56.40** | **58.64** | 49.30 | **60.58** | **41.66** | **63.84** | **29.50** | **15.87** |

(Chen et al., 2024b) and BLINK for holistic perception, and MMCoQA (Song et al., 2024) and ALFRED(Song et al., 2024) for conversational and embodied AI tasks. We compare RetentiveKV against several competitive KV cache compression baselines, including importance-centric pruning methods: H2O (Zhang et al., 2023), SnapKV (Li et al., 2024), and modality-aware compression techniques: LOOK-M (Wan et al., 2024), Meda (Wan et al., 2025), SAINT (Jeddi et al., 2025).

### 4.2 Experiment Results

Table 1 reports the overall performance comparison between RetentiveKV and state-of-the-art KV cache compression methods. The results demonstrate that RetentiveKV consistently outperforms prior KV eviction and compression strategies. In fine-grained visual perception tasks (*e.g.,* DocVQA and TextVQA), RetentiveKV achieves +2.1 gains over the importance-centric baselines. These improvements stem from the entropy-guided state evolution, which preserves high-uncertainty visual tokens and enables their delayed reactivation, effectively mitigating premature eviction in importance-based methods. For long-context conversations and Needle-in-a-Haystack tasks such as MMCoQA and ALFRED, RetentiveKV improves over modality-aware baselines by +3.0 points. Unlike modality-aware methods that merely adjust compression ratios, RetentiveKV preserves deferred-critical visual information via continuous state evolution, enabling retrieval under long-horizon reasoning. The consistent performance gains observed across different model architectures (from LLaVA to Qwen3-VL) and parameter scales (4B to 8B) underscore the architectural agnosticism of our approach.

Table 2: Comparison of decoding latency and GPU memory usage for different cache budgets.

| **Method** | **Budget** | **Latency** | **Memory** |
|---|---|---|---|
| Full Cache | 100% | 32.15 ms/token | 2.24 GiB |
| RetentiveKV | 50% | 27.84 ms/token | 1.18 GiB |
| | 35% | 24.42 ms/token | 0.84 GiB |
| | 20% | 21.42 ms/token | 0.46 GiB |
| | 5% | **18.42** ms/token | **0.23** GiB |

### 4.3 Efficiency Analysis

To investigate the computational efficiency of our approach, we evaluate the decoding latency and

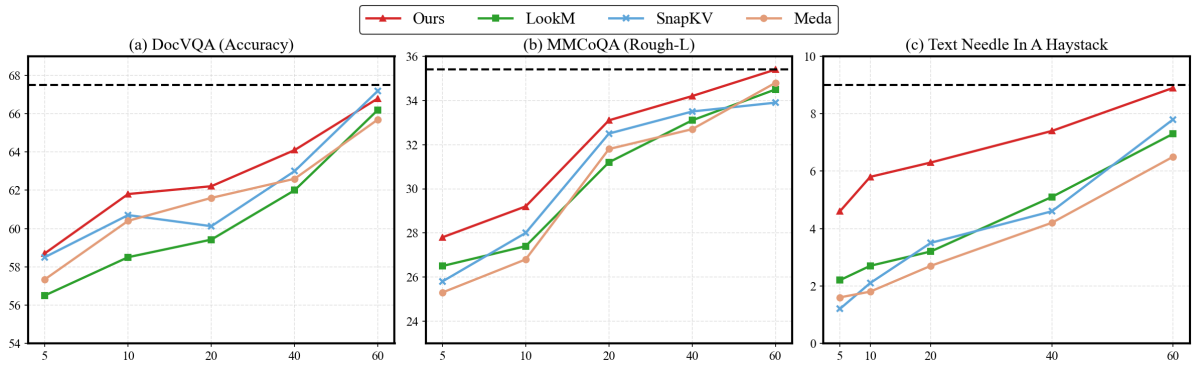



Figure 4: Comparison results for various cache budgets.

GPU memory footprint across varying compression ratios on a single NVIDIA A100 Tensor Core GPU. As presented in Table 2, RetentiveKV demonstrates superior resource efficiency compared to the Full Cache baseline. Specifically, under the most constrained budget of 5%, our method achieves a remarkable reduction in memory consumption, dropping from 2.24 GiB to 0.23 GiB. Simultaneously, it accelerates inference speed by 1.75×, reducing latency from 32.15 ms/token to 18.42 ms/token. Empirical observations reveal that GPU memory occupancy is directly proportional to the retained KV cache budget, suggesting that RetentiveKV effectively compresses context without introducing significant computational overheads.

### 4.4 Ablation study

To evaluate the individual contributions of our proposed components, we conduct extensive ablation studies. As detailed in Table 3: 1) Removing state retention and Query-Conditioned Retrieval (QR) incurs the most severe performance degradation from 4.6% to 5.4%. This mechanism allows the model to recall specific visual and textual cues that were previously evicted, mitigating the irreversible information loss inherent in discrete pruning. 2) Replacing the Modality-Agnostic state (MA) with a unified state space for all modalities degrades performance by 3.1% on average. This disparity confirms the presence of cross-modal interference in unified structures. The specialized visual-dominant state prevents high-density textual sequences from overwhelming sparse visual cues. 3) Replacing the entropy-driven metric (EM) with a standard importance-centric metric leads to a degradation of 2.9% on MMCoQA and 4.6% on DocVQA. The marginal reduction suggests that entropy-driven selection is effective for long-context conversations.

Table 3: Ablation study on key components of RetentiveKV.

| EM | MA | QR | DocVQA | MMCoQA |
|---|---|---|---|---|
| ✓ | ✓ | | 58.70 | 60.91 |
| ✓ | | ✓ | 60.20 | 61.55 |
| | ✓ | ✓ | 59.80 | 61.38 |
| ✓ | ✓ | ✓ | **61.40** | **64.22** |

### 4.5 Influence of Various Cache Budgets

To evaluate the effectiveness of RetentiveKV under varying cache budgets, we conduct experiments on the *Qwen3-VL-8B* model with cache budgets ranging from 5% to 60%. As shown in Figure 4, we observe that conventional importance-based eviction methods suffer from abrupt performance degradation when the cache budget is restricted to extremely low ratios (*e.g.,* 5%–10%). This sharp decline confirms our hypothesis regarding Deferred Importance. In contrast, RetentiveKV exhibits robust performance stability. As illustrated in Figure 4(a) and (b), our method achieves a significant accuracy margin over the second-best baseline (SnapKV) on both DocVQA and MMCoQA tasks within the 10%–20% budget interval. The performance disparity is maximized in the "Text Needle In A Haystack" task (Figure 4(c)). At a constrained 5% budget, RetentiveKV doubles the retrieval score of Meda and significantly surpasses SnapKV. This indicates that our entropy-driven state transition mechanism effectively compresses visual semantics into continuous states rather than discretely eliminating them, thereby preserving the spatial continuity and retrieving long-range dependencies required for later reasoning.

## 5 Conclusion

In this paper, we address the critical dilemma of Visual Continuity Collapse and Deferred Importance in multimodal KV cache compression. We introduce RetentiveKV, a framework that bridges the gap between discrete token pruning and continuous state modeling. Instead of viewing low-attention tokens as redundant noise, we reinterpret them through the information entropy, identifying those with high uncertainty as candidates for preservation within a continuously evolving state space.

## 6 Limitation

Despite the robust performance of RetentiveKV, two directions require further exploration. 1) Scaling Laws: Our current experiments validate efficacy on models up to 8B parameters. Future work will focus on investigating the behavior of entropy-driven retention in massive-scale MLLMs (>30B), particularly to determine if the "deferred importance" phenomenon intensifies as model capacity grows. 2) Towards Omni-modal Perception: While this paper addresses visual-linguistic challenges, the "continuous evolution" paradigm of RetentiveKV is inherently agnostic to data modality. As the field advances towards Omni-modal MLLMs, we see significant potential in applying RetentiveKV to real-time audio and video. Leveraging the continuous nature of SSMs to unify memory management across diverse modalities stands as a compelling direction for our future research.

## 7 Acknowledgements

This work was supported by the Key Research and Development Program of Zhejiang Province(No. 2025C01026), and the National Natural Science Foundation of China (No. 62402429, U24A20326, 62441236). This work was also partially supported by the Ningbo Yongjiang Talent Introduction Programme (2023A-397-G) and Young Elite Scientists Sponsorship Program by CAST (2024QNRC001). The author gratefully acknowledges the support of Zhejiang University Education Foundation Qizhen Scholar Foundation.

## References

Alan Baddeley. 2000. The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11):417–423.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024a. 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*, pages 19–35. Springer.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. 2020. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33:1474–1487.

Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Ahmadreza Jeddi, Negin Baghbanzadeh, Elham Dolatabadi, and Babak Taati. 2025. Similarity-aware token pruning: Your vlm but faster. *arXiv preprint arXiv:2503.11549*.

Zhonghua Jiang, Kui Chen, Kunxi Li, Keting Yin, Yiyun Zhou, Zhaode Wang, Chengfei Lv, and Shengyu Zhang. 2025a. Acckv: Towards efficient audio-video llms inference via adaptive-focusing and cross-calibration kv cache optimization. *Preprint*, arXiv:2511.11106.

Zhonghua Jiang, Kunxi Li, Yiyun Zhou, Sihao Liu, Zhaode Wang, Chengfei lv, and Shengyu Zhang. 2025b. Purekv: Plug-and-play kv cache optimization with spatial-temporal sparse attention for vision-language large models. *Preprint*, arXiv:2510.25600.

Kunxi Li, Zhonghua Jiang, Zhouzhou Shen, Zhaode-Wang ZhaodeWang, Chengfei Lv, Shengyu Zhang, Fan Wu, and Fei Wu. 2025a. Madakv: Adaptive modality-perception kv cache eviction for efficient multimodal long-context inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13306–13318.

Kunxi Li, Yufan Xiong, Zhonghua Jiang, Yiyun Zhou, Zhaode Wang, Chengfei Lv, and Shengyu Zhang. 2025b. Flowmm: Cross-modal information flow guided kv cache merging for efficient multimodal context inference. *Preprint*, arXiv:2511.05534.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. 2024. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970.

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. 2025. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5334–5342.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. 2023. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36:52342–52364.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Minesh Mathew, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. 2020. Docvqa: A dataset for vqa on document images. corr abs/2007.00398 (2020). *arXiv preprint arXiv:2007.00398*.

Jiahe Shi, Zhengqi Gao, Ching-Yun Ko, and Duane Boning. 2025. Earl: Entropy-aware rl alignment of llms for reliable rtl code generation. *arXiv preprint arXiv:2511.12033*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. 2024. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*.

Zhongwei Wan, Hui Shen, Xin Wang, Che Liu, Zheda Mai, and Mi Zhang. 2025. Meda: Dynamic kv cache allocation for efficient multimodal long-context inference. *arXiv preprint arXiv:2502.17599*.

Zhongwei Wan, Ziang Wu, Che Liu, Jinfa Huang, Zhihong Zhu, Peng Jin, Longyue Wang, and Li Yuan. 2024. Look-m: Look-once optimization in kv cache for efficient multimodal long-context inference. *arXiv preprint arXiv:2406.18139*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Jing Xiong, Jianghan Shen, Fanghua Ye, Chaofan Tao, Zhongwei Wan, Jianqiao Lu, Xun Wu, Chuanyang Zheng, Zhijiang Guo, Min Yang, and 1 others. 2025. Uncomp: Can matrix entropy uncover sparsity?—a compressor design from an uncertainty-aware perspective. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4179–4199.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others.

2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, and 1 others. 2023. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710.