MadaKV: Adaptive Modality-Perception KV Cache Eviction for Efficient Multimodal Long-Context Inference

Kunxi Li[1*], **Zhonghua Jiang**[1*], **Zhouzhou Shen**[2], **Zhaode Wang**[3], **Chengfei Lv**[3], **Shengyu Zhang**[1†], **Fan Wu**[4], **Fei Wu**[1]
[1]Zhejiang University, [2]Southeast University, [3]Alibaba, [4]Shanghai Jiao Tong University
{kunxili, jiangzhonghua, sy_zhang, wufei}@zju.edu.cn, zhouzhoushen@seu.edu.cn,
{zhaode.wzd, chengfei.lcf}@alibaba-inc.com, fwu@cs.sjtu.edu.cn

Abstract

This paper introduces MadaKV, a modality-adaptive key-value (KV) cache eviction strategy designed to enhance the efficiency of multimodal large language models (MLLMs) in long-context inference. In multimodal scenarios, attention heads exhibit varying preferences for different modalities, resulting in significant disparities in modality importance across attention heads. Traditional KV cache eviction methods, which are tailored for unimodal settings, fail to capture modality-specific information, thereby yielding suboptimal performance. MadaKV addresses these challenges through two key components: modality preference adaptation and hierarchical compression compensation. By dynamically sensing modality information within attention heads and adaptively retaining critical tokens, MadaKV achieves substantial reductions in KV cache memory footprint and model inference decoding latency (1.3 to 1.5 times improvement) while maintaining high accuracy across various multimodal long-context tasks. Extensive experiments on representative MLLMs and the MileBench benchmark demonstrate the effectiveness of MadaKV compared to existing KV cache eviction methods.

1 Introduction

In recent years, leveraging the transformer architecture, autoregressive language models have shown remarkable progress in handling long-context inputs across various tasks (Liu et al., 2024a; Tworkowski et al., 2024; Touvron et al., 2023; Jiang et al., 2023). However, the autoregressive decoding mechanism, which necessitates the consideration of every preceding token when generating a new one, incurs quadratic complexity, posing computational challenges. The KV cache approach addresses this by caching key and value tensors from past tokens, reducing the decoding process to linear time complexity. However, as the context length increases, the KV cache's memory footprint increases significantly (Shi et al., 2024). Building on prior research that highlights the sparsity within attention mechanisms, where only a subset of tokens significantly influences outcomes (Zhao et al., 2019; Beltagy et al., 2020; Choromanski et al., 2020; Wang et al., 2020), recent studies have focused on identifying the importance of each token and discarding the KV cache of less significant tokens (Zhang et al., 2024b; Cai et al., 2024; Chen et al., 2024).

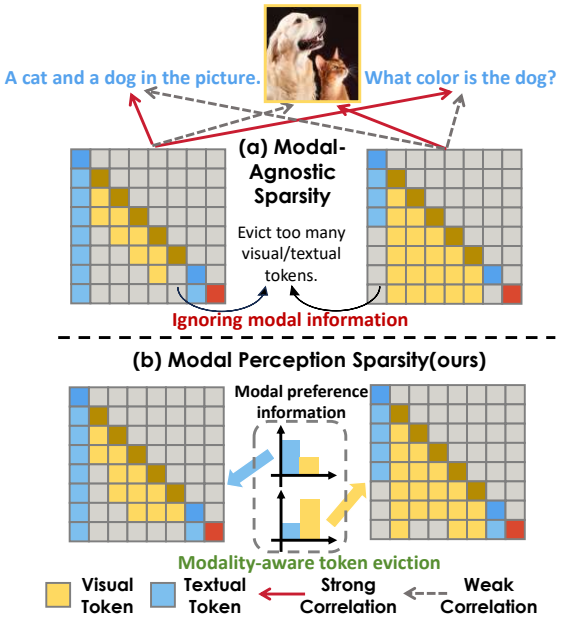

Figure 1: Comparison between Modal-Agnostic Sparsity (a) and MadaKV's Modal Perception Sparsity (b).

In the field of Multimodal Large Language Models (MLLMs) (Liu et al., 2024b; Li et al., 2024; Wang et al., 2024), these challenges are also present, yet traditional single-modality approaches are suboptimal. The information density encapsulated within tokens varies significantly across

*These authors contributed equally.
†Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13306–13318
July 27 - August 1, 2025

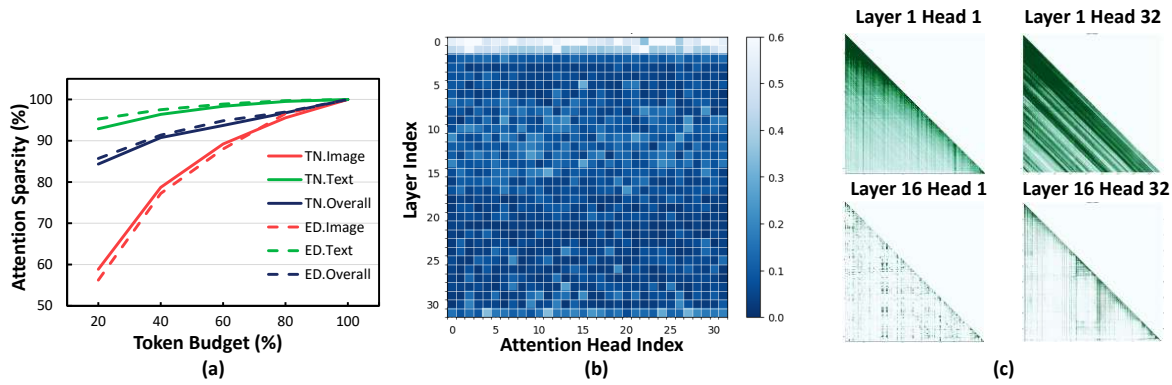

Figure 2: (a) The sparsity differences among tokens of different modalities. TN: Text Needle In A Haystack; ED: Conversational Embodied Dialogue. (b) The distribution of attention scores assigned to text tokens across different attention heads. A higher score indicates that more attention is allocated to text tokens, while a lower score suggests that more attention is directed towards visual tokens. (c) The attention score matrices across different layers.

modalities; for instance, textual tokens often encode semantic concepts concisely, while visual tokens require fine-grained spatial representations spanning hundreds of tokens, highlighting the need for modality-specific granularity when evaluating token importance. Moreover, the complex interactions between multimodal tokens lead to substantial variations in attention patterns across different input instances. As illustrated in Figure 1, traditional unimodal approaches, lacking awareness of multimodal information, may inadvertently evict critical multimodal tokens, thereby generating erroneous responses.

More recently, LOOK-M (Wan et al., 2024) applies the KV cache eviction strategies to MLLMs. However, this approach empirically assigns fixed modality prioritization, overlooking the significance of differences among modalities in various task contexts (examples are provided in the appendix). We analyze the attention scores assigned to visual and text tokens by each attention head on multimodal long-context scenarios. As shown in Figure 2 (b), the proportion of scores assigned to different modalities varies significantly across attention heads, indicating a preference for modality-specific information. This insight suggests that a KV cache eviction strategy for MLLMs is necessary to accommodate the varying importance of modalities in different contexts.

This paper introduces **MadaKV**, a plug-and-play modality-adaptive KV cache compression strategy, designed to mitigate the inference costs of MLLMs. Specifically, our approach comprises two integral components: *modality preference adaptation* (MPA) and *hierarchical compression compensation* (HCC). Firstly, MPA dynamically discerns modality preference patterns through real-time analysis of cross-modal contextual interactions, thereby facilitating the learning of specific modality importance for attention head and guiding the eviction of multimodal KV caches accordingly. Second, recognizing that multimodal attention patterns vary significantly across layers, tasks, and instances, we introduce a progressive hierarchical compression compensation. This mechanism dynamically adjusts the eviction ratio based on the complexity of modality information in the context. Additionally, it incorporates a compensation scheme to balance the cache budget across layers. The compensation is accumulated across layers, summarizing the eviction of modality information from previous layers and providing alignment guidance for the eviction strategy in the current layer. This ensures that the overall cache budget is maintained while preventing error propagation due to modality information compression.

We conduct extensive experiments on representative MLLMs, including LLaVA-v1.5 (Liu et al., 2024b) and Qwen2.5-VL, and evaluate their performance across a variety of multimodal long-context tasks within the MileBench (Song et al., 2024a): temporal multi-image tasks, semantic multi-image tasks, needle in a haystack task, and image retrieval tasks. MadaKV achieves better accuracy for a given degree of KV cache sparsity than baselines Specifically, MadaKV achieves a 1.3 to 1.5 times improvement in model inference decoding latency and reduces the KV Cache Memory Footprint by 80% to 95%, while maintaining performance on multimodal long-context tasks.

2 Related work

Memory efficient inference has always been an important research direction in the field of deep learning, with early studies mainly focusing on techniques such as activation checkpoints (Chen et al., 2016; Zhou et al., 2025d), offloading (Ren et al., 2021), and dynamic memory management (Rhu et al., 2016). Although these methods alleviate memory pressure to some extent, they often introduce additional latency or hardware dependency, limiting their widespread adoption in practical applications. For large language models (LLMs) (Floridi and Chiriatti, 2020; Achiam et al., 2023; Meta, 2024; Yang et al., 2024a; Kong et al., 2025; Zhou et al., 2025b,a) based on transformers (Waswani et al., 2017; Wang et al., 2025; Zhou et al., 2025c, 2024), KV cache, as a key component in autoregressive generation tasks, stores the key and value vectors calculated in the attention mechanism, thereby avoiding redundant calculations during the decoding process. However, as the model size and sequence length increase, the memory usage of KV cache gradually becomes a performance bottleneck (Shi et al., 2024). Recent research has explored various methods for optimizing KV cache, such as quantization methods (Hooper et al., 2024; Yang et al., 2024c; Sheng et al., 2023) that reduce memory usage by lowering the accuracy of cache values (such as from FP16 to INT8), and eviction strategies (Zhang et al., 2024b; Yang et al., 2024b) that selectively delete less important key-value pairs (Xiao et al., 2023a; Han et al., 2024). Although these techniques have achieved significant results in single-modal models, they are not directly applicable to multimodal models due to the unique characteristics of cross-modal attention.

MLLMs, such as Flamingo (Alayrac et al., 2022), Qwen-VL (Bai et al., 2023), and LLaVA (Liu et al., 2023) integrate information from multiple modalities achieving state-of-the-art performance in tasks such as visual question answering and image captioning. Due to the fact that MLLMs typically require a large number of tokens to store multimodal inputs (Yin et al., 2023), such as images, this introduces additional complexity when managing KV Cache. Recent studies have proposed methods such as sparse attention (Child et al., 2019) and low rank approximation (Song et al., 2024b) to reduce the computational overhead of MLLMs, but have not solved the memory usage problem of KV Cache in multimodal settings.

Compressing KV Cache in MLLMs presents unique challenges. Cross-modal attention patterns are often different from traditional LLMs, which limits the effectiveness of conventional KV cache optimization techniques in MLLMs. Some recent studies such as FastGen (Ge et al., 2023) and H2O (Zhang et al., 2024b) propose optimizing the efficiency of KV Cache usage through lightweight model analysis and adaptive cache compression strategies. LOOK-M (Wan et al., 2024) prioritizes evicting or merging the KV Cache of image tokens by observing cross-modal attention patterns, while significantly compressing the KV Cache while maintaining contextual quality without the need for fine-tuning the model. However, the method of prioritizing the eviction of visual tokens exhibits limited applicability in multimodal settings. Departing from prior work, our method introduces context-aware modality prioritization. This adaptive mechanism mitigates critical information loss and demonstrates robust effectiveness across a wider spectrum of multimodal scenarios.

3 Method

3.1 Preliminary

In the inference process of MLLMs, KV Cache is a key memory mechanism used to store the Key and Value in the attention mechanism, thereby accelerating the generation process and reducing computational overhead. The use of KV Cache mainly involves two stages: prefilling and decoding.

Prefill phase: During this stage, the model ingests an input sequence $X = [x_1^t, x_2^t, ..., x_{n-1}^v, x_n^v]$, where n represents the length of the input sequence, x^t represents a text token x^v represents a visual token. To establish the foundational context for efficient decoding, the model computes Key and Value vectors:

$$K = XW_k, \quad V = XW_v, \tag{1}$$

where $W_k \in \mathbb{R}^{d\times d}$ and $W_v \in \mathbb{R}^{d\times d}$ are the weight matrices of the attention module, respectively. d represents the hidden dimension of the model. Subsequently, the Key and Value of all tokens are stored in the KV Cache:

$$\text{KV Cache} = \{K, V\}, \tag{2}$$

Decoding phase: During iterative token generation, the model leverages the KV Cache and existing context to predict the next token. At generation step t, let the input of the attention module be

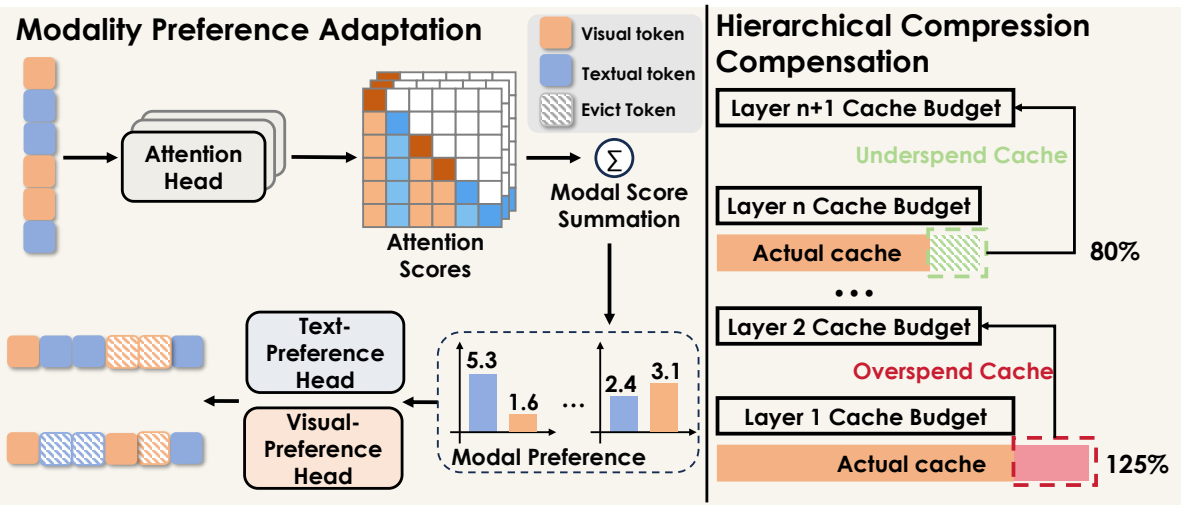

Figure 3: Overview of MadaKV. Modality Preference Adaptation identifies the modality preferences of attention heads and guides the computation of cache budgets for each modality. Hierarchical Compression Compensation employs a compensation mechanism to coordinate budgets across layers. This component ensures that the trade-offs between compression and information retention are balanced across different layers of the model.

x^t. The model firstly computes its corresponding Query, Key and Value projections:

$$Q_t = x_t W_q, \quad K_t = x_t W_k, \quad V_t = x_t W_v. \quad (3)$$

Then, the KV Cache gets updated as:

$$K = [K, K_t], \quad V = [V, V_t]. \quad (4)$$

This expanded cache enables efficient attention computation for the subsequent token prediction. The t-th step attention outputs O_t is computed :

$$O_t = \text{softmax}\left(\frac{Q_t \cdot K^T}{\sqrt{d_k}}\right) V \quad (5)$$

While the KV Cache mechanism effectively eliminates redundant key-value recomputation during MLLMs inference, its substantial memory footprint necessitates efficient compression strategies.

3.2 Observation

In this section, we explore the attention patterns of MLLMs in multimodal long-context scenarios, presenting experimental findings. The study is conducted on the LLaVA-v1.5-7B (Liu et al., 2024b) using the Milebench (Song et al., 2024a) benchmark.

Unlike traditional LLMs long-context, multimodal long contexts involve not only interactions among tokens within the same modality but also more complex interactions between tokens of different modalities. However, there is a lack of comprehensive analysis of modal behaviors within multimodal long contexts. We investigate the modal behaviors within the MLLMs attention patterns in multiple dimensions, including token level, head level of attention, layer level, and task-specific context.

Token level. Prior studies (Zhang et al., 2024b; Xiao et al., 2023b; Chen et al., 2024) have demonstrated that the attention mechanisms of LLMs exhibit inherent sparsity, characterized by a small portion of tokens contributing the majority of attention scores. This sparsity allows for the eviction of unnecessary KV embeddings, thereby reducing the memory demands of KV cache. Building on these insights, we investigate attention mechanism sparsity in MLLMs. Specifically, we conduct zero-shot inference on two datasets: Text Needle (TN) and Conversational Embodied Dialogue (ED), analyzing attention score distributions under varying token budgets. To provide a more granular understanding, we partition tokens into modality-specific subsets and quantify the sparsity for each subset independently, thus extending the analysis beyond global token sparsity. All attention scores are aggregated through layer-wise and head-wise averaging.

The findings are depicted in Figure 2 (a). We observe that the attention mechanisms of MLLMs exhibit significant sparsity: with merely 20% of tokens capture nearly 90% of attention scores, which demonstrates that a compact token subset can effectively approximate full KV embeddings. Notably, modality-specific sparsity patterns exhibit marked differences. Specifically, textual tokens display sharp attention concentration, whereas visual

tokens exhibit diffuse attention allocation. This observation aligns with our earlier discussion on information density: visual tokens' lower information density correlates with their flatter attention distributions. This sparsity heterogeneity reveals the limitations of prior KV cache compression methods for LLMs that treat all tokens equally and underscores the need for modality-specific compression strategies to optimize performance.

Attention head level. In Figure 2 (c), we present the attention score matrices of different attention heads (with additional examples provided in the appendix). It can be observed that, similar to LLM, the attention heads of MLLMs exhibit distinct attention patterns for the same sample. To investigate the behavior of different modalities across attention heads, we analyze the distribution of attention scores for each modality within each head. As shown in Figure 2 (b), there are variations in how attention heads allocate attention scores to different modalities. Typically, an attention head tends to favor one modality by assigning it a higher proportion of attention scores. We refer to this phenomenon as modality preference. Intuitively, modality preference indicates the modality that an attention head is adept at processing. This insight suggests that we should retain more tokens corresponding to the preferred modality.

Layer level. The attention patterns of MLLMs across different layers are similar to those of LLM, as illustrated in Figure 2 (c). In the initial layers of the model, the attention scores are distributed evenly, while in the subsequent layers, they are concentrated on a few tokens. This observation suggests that we should retain more KV embeddings in the initial layers and fewer in the later layers, which aligns with the claims made by Cai et al. (2024). However, Liu et al. (2024c) argue that important tokens exhibit greater variability in the higher layers, thus requiring a larger cache to reduce cache misses. We also observe a corresponding phenomenon in the attention score matrix of MLLMs, where the distribution of attention scores is more uniform in the final layer compared to the middle layers. These observations lead us to recognize the importance of tokens in both the initial and final layers, necessitating the preservation of more KV embeddings.

Task level. As previously discussed in the earlier sections, the significance of modalities varies across different tasks. For instance, as illustrated in Figure 5, in the Text Needle task, textual tokens hold greater importance, whereas in Image Retrieval, visual tokens dominate. This disparity necessitates the pursuit of a modality-adaptive compression strategy.

3.3 MadaKV

In this section, we present a modality-adaptive KV cache compression strategy, encompassing Modality Preference Adaptation and Hierarchical Compression Compensation mechanisms. This approach dynamically adjusts to the varying significance of modalities, ensuring efficient cache management while maintaining performance.

Modality Preference Adaptation. The modality preference characterizes the system's inherent bias in allocating attention resources across different modalities. This insight suggests that to retain as much contextual information as possible, we should align with the modality preference and evict modality tokens at different granularities. Formally, we define the preference metric as the sum of the importance of modality tokens:

$$w_v = \sum_{i \in X_v} \psi(i), \quad w_t = \sum_{i \in X_t} \psi(i) \tag{6}$$

where X_v and X_t denote the sets for visual and textual tokens respectively. In long-context scenarios, evaluating token importance using cumulative attention can be biased (Chen et al., 2024). Here, we opt for proxy tokens to assess token importance more fairly:

$$\psi(i) = \sum_{j \in \mathcal{P}} \alpha_{j \to i} \tag{7}$$

where $\mathcal{P}$ denotes the proxy token set, and $\alpha_{j \to i}$ represents the attention score from the proxy token j to token i. We select a few tokens from the end of the prompt as proxy tokens, as they typically represent task-specific questions.

We now introduce how to compute the cache budget for each modality using modality preference information. For the h-th attention head in the l-th layer, the budget for each modality is determined based on preference metric:

$$\varphi_v^{l,h} = \frac{w_v}{w_v + w_t} \varphi^l, \quad \varphi_t^{l,h} = \frac{w_t}{w_v + w_t} \varphi^l \tag{8}$$

where the cache budget for the l-th layer of the model as φ^l. Our method conducts modality-specific KV cache eviction for each attention head,

effectively improving cache utilization. This flexibility allows our approach to be seamlessly adapted to a wide range of application scenarios.

Hierarchical Compression Compensation. We present a Hierarchical Compression Compensation (HCC) strategy that adaptively modulates token eviction policies across different layers based on the input instance. This design is motivated by the fundamental observation that different layers inherently possess diverse sparsity characteristics, which significantly impacts the model's overall performance.

Prior work has shown that evicting tokens in shallow layers can lead to cascading errors that propagate and amplify through the network (Zhang et al., 2024a). As a result, conservative eviction is recommended in these layers. Conversely, Liu et al. (2024c) argue that important tokens exhibit greater variability in higher layers, necessitating larger caches to reduce cache misses. In this work, we integrate these insights from prior research. Rather than empirically predefining the importance of each layer, we propose an adaptive approach that considers both the complexity of modality information in the current layer and the compression status of preceding layers. This ensures that the changes brought about by eviction remain within acceptable bounds, thereby balancing the trade-off between cache efficiency and information retention. Specifically, we define the sparsity of each modality within the attention heads as follows:

$$\begin{aligned} k_v^{l,h} &= \min\Big\{|\mathcal{C}_v| \;\Big|\; \sum\nolimits_{i\in\mathcal{C}_v} \psi(i) \geq \theta w_v\Big\}, \\ k_t^{l,h} &= \min\Big\{|\mathcal{C}_t| \;\Big|\; \sum\nolimits_{i\in\mathcal{C}_t} \psi(i) \geq \theta w_t\Big\}, \end{aligned} \tag{9}$$

where $\mathcal{C}_v$ and $\mathcal{C}_t$ represent subsets of visual and text tokens respectively, and θ indicates the threshold value. The budget compensation for the l-th layer is defined as follows:

$$K^l = \sum_{h=1}^{H} (k_v^{l,h} + k_t^{l,h} - \varphi^l), \tag{10}$$

where H denotes the number of attention heads in the model's attention mechanism and φ^l indicates the allocated cache budget for the l-th layer. A positive K^l value indicates that the current layer has exceeded its budget, while a negative value means the current layer has a saved budget. This budget compensation is accumulated across layers and influences the budget allocation for subsequent layers:

$$\varphi^{l+1} = \varphi^l - \frac{K^l}{L-l}, \tag{11}$$

where L denotes the total number of layers in the model. The inter-layer compression compensation allows the model to adaptively adjust its strategy for the current layer based on the modality sparsity of the current layer and the compression status of previous layers. Compared to prior heuristic approaches, this approach is more generalizable and effective.

4 Experiments

4.1 Setting

We sample nine tasks from the MileBench benchmark (Song et al., 2024a), which is the first benchmark specifically designed to test the long-context multimodal capabilities of MLLMs. MileBench covers a wide range of general scenarios, including temporal multi-image tasks, semantic multi-image tasks, needle-in-a-haystack tasks, and image retrieval tasks. On average, each sample in MileBench contains 15.2 images and 422.3 words.

To evaluate MadaKV, we conduct comprehensive experiments on widely-adopted long-context MLLMs: LLaVA-v1.5-7/13B (Liu et al., 2024b) and Qwen2.5-VL-7B. We compare MadaKV against several representative KV cache eviction methods. We include StreamingLLM (Xiao et al., 2023a), H2O (Zhang et al., 2024b), and SnapKV (Chen et al., 2024), which are all text-based KV cache eviction methods. Additionally, we compare against LOOK-M (Wan et al., 2024), an eviction method designed for multimodal scenarios.

4.2 Experiment Results

In Table 1, we present a comprehensive comparison of MadaKV against various eviction methods in the context of multimodal long-context scenarios. The experimental results highlight MadaKV's effectiveness in managing KV caches under memory constraints while maintaining high performance across diverse tasks. Specifically, MadaKV achieves an 80% reduction in memory usage, with only a slight drop in average accuracy compared to full caching. This demonstrates MadaKV's ability to significantly reduce memory footprint with minimal performance trade-offs.

Compared to baseline eviction methods, MadaKV consistently outperforms across most

Method	TN	IEdit	MMCoQA	STD	ALFRED	CLEVR-C	DocVQA	ST	OI	Average
LLaVA-v1.5-7B										
Full Cache	9.68	7.98	33.50	16.32	15.18	16.62	46.00	63.50	48.50	28.59
StreamingLLM	3.12	3.59	26.00	11.77	3.73	10.44	42.50	43.00	44.00	20.91
H2O	2.50	5.51	28.00	15.73	14.86	14.07	44.00	63.50	44.00	25.80
SnapKV	3.27	6.03	29.00	14.82	14.40	15.37	45.50	**64.00**	45.50	26.43
LOOK-M	3.34	6.51	29.50	15.79	13.96	14.12	45.50	63.00	46.50	26.47
MadaKV	**9.38**	**6.97**	**31.00**	**15.85**	**15.06**	**16.76**	**47.00**	**64.00**	**48.00**	**28.22**
LLaVA-v1.5-13B										
Full Cache	25.34	9.01	40.50	15.70	18.56	15.98	55.50	74.50	52.00	34.12
StreamingLLM	7.81	2.63	28.50	14.05	1.72	6.55	49.80	68.00	50.00	25.45
H2O	7.81	8.70	34.50	15.21	15.03	15.06	53.00	74.00	51.00	30.48
SnapKV	7.92	8.52	35.00	15.53	16.84	15.37	53.50	**74.50**	51.00	30.91
LOOK-M	10.00	8.77	37.50	15.46	15.25	14.12	53.00	74.00	50.50	30.96
MadaKV	**23.43**	**9.41**	**38.50**	**15.62**	**17.76**	**16.76**	**55.00**	**74.50**	**51.50**	**33.61**
Qwen2.5-VL-7B										
Full Cache	11.25	29.45	44.50	28.36	37.53	42.46	62.50	63.00	61.00	63.34
StreamingLLM	3.46	25.12	40.00	26.14	34.07	27.58	60.00	59.00	60.50	55.98
H2O	3.97	29.13	40.50	26.79	37.78	36.03	**62.50**	61.50	61.00	59.70
SnapKV	4.55	29.25	41.50	26.54	37.41	39.41	61.50	60.00	60.50	60.11
LOOK-M	4.23	28.73	42.00	**28.97**	36.66	38.66	62.00	61.50	61.50	60.71
MadaKV	**10.73**	**29.39**	**43.50**	27.30	**37.95**	**40.57**	**62.50**	**62.00**	**62.50**	**62.74**

Table 1: Performance of eviction strategies. The best results are highlighted in **bold**.

datasets. Notably, MadaKV substantially outperforms single-modality text-based KV eviction methods. For instance, in the TN task, MadaKV achieves a 6.11% improvement in performance. This finding underscores a critical limitation of text-based KV eviction methods in multimodal long-context scenarios: they often overlook modality-specific information, leading to improper eviction of KV caches and generating erroneous responses. Moreover, MadaKV outperforms LOOK-M, a method designed for multimodal scenarios. Unlike LOOK-M, which prioritizes retaining text tokens, MadaKV adaptively determines the importance of tokens from different modalities based on the specific task context. This approach prevents excessive eviction of tokens from any single modality, thereby enhancing the model's robustness when handling complex multimodal long-context inputs.

4.3 Influence of Various Cache Budgets

To evaluate the effectiveness of MadaKV under varying cache budgets, we conduct experiments on the LLava-v1.5-7B model with cache budgets ranging from 5% to 60%. We select three subtasks for assessment: Spot-the-Diff, MMCoQA, and Text Needle In A Haystack. The results are presented in Figure 4. Most achieve methods performance comparable to full caching when using a 50% cache budget. This suggests that multimodal long-context scenarios contain substantial redundant information, which can be pruned to reduce memory usage without significant performance loss. MadaKV consistently outperforms the baselines across all cache budgets. Notably, in the Text Needle In A Haystack task, MadaKV's performance with a 20% cache budget matches that of LOOK-M using a 60% cache budget. Additionally, MadaKV shows significant improvements over baselines when the cache budget is below 10%. These findings demonstrate MadaKV's ability to accurately identify critical information within KV caches, thereby minimizing context loss even under stringent memory constraints.

Method	Budget	Decoding Latency	GPU Memory
Full Cache	100%	27.85 ms/token	1.63 GiB
MadaKV	20%	19.57 ms/token	0.41 GiB
MadaKV	5%	**17.16 ms/token**	**0.16 GiB**

Table 2: Model Speed and KV Cache GPU Memory Usage. The best results are highlighted in **bold**.

4.4 Efficiency Analysis

We delve into the efficiency of our proposed method, as detailed in Table 2. Specifically, we examine the decoding speed and memory usage of model inference both with and without our method. To ensure the reliability and robustness of our find-

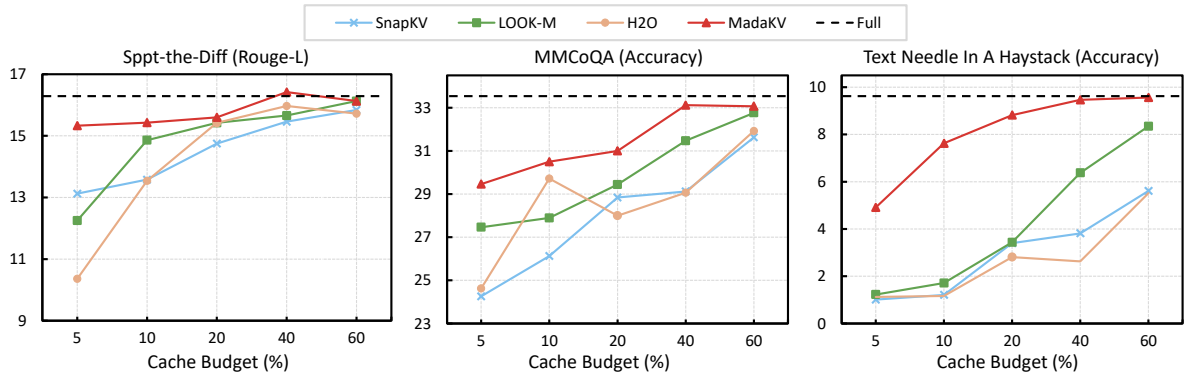

Figure 4: Comparison results for various cache budgets.

MPA	HCC	TN	IEdit	ALFRED
✘	✘	2.47	3.55	14.32
✔	✘	6.58	5.72	14.86
✘	✔	5.51	5.19	14.61
✔	✔	**9.38**	**6.97**	**15.06**

Table 3: Ablation study of the effect of individual module. MPA: modality preference adaptation, HCC: hierarchical compression compensation. The best results are highlighted in **bold**.

ings, we conduct tests on decoding latency and GPU memory usage using 20 randomly selected data entries. All speed tests were performed on a single NVIDIA A100 Tensor Core GPU.

As shown in Table 2, MadaKV exhibits significantly lower decoding latency compared to the model with a full cache. This advantage is particularly evident in long generation tasks, where the efficiency of our method is further accentuated. Moreover, we analyze the speed and GPU memory usage of KV Cache under two budget scenarios: 20% and 5%. These results were derived from the mean values obtained during the inference process of 20 randomly sampled data points. Our findings reveal that the average GPU memory consumption is nearly proportional to the cache budget. Specifically, at a 20% KV Cache budget, memory usage during the decoding stage is reduced by approximately 80% compared to a Full Cache scenario. This highlights the substantial memory savings achieved through our method.

4.5 Ablation Study

The Effect of Modality Preference Adaptation. Modality Preference Adaptation is used to determine the modality preferences of attention heads, indicating which modality each head tends to allocate more attention scores to. This information guides the model in deciding how many tokens to retain for each modality. In Table 3, we report the impact of MPA on model performance. Removing this strategy leads to a noticeable drop in performance. These results demonstrate the effectiveness of MPA in multimodal long-context scenarios. As previously discussed, the importance of different modalities varies greatly across tasks, making it impractical to simply designate one modality as universally more critical than others. MPA allows the model to dynamically adjust the number of tokens retained for each modality based on its actual importance. This dynamic adjustment improves the model's ability to integrate multimodal information while maintaining computational efficiency.

The Effect of Hierarchical Compression Compensation. As illustrated in Table 3, Hierarchical Compression Compensation significantly enhances model performance. For instance, in the TN task, HCC leads to a 3.07% improvement, underscoring its effectiveness. This finding highlights the importance of allocating different cache budgets to different layers in multimodal long-context scenarios. Our approach dynamically adjusts cache resource allocation by considering both the sparsity of the current layer and the cache budget usage of previous layers. This strategy not only effectively reduces redundant computation but also ensures that critical information is retained and efficiently processed throughout the long context.

5 Conclusion

In this study, we introduce MadaKV, a modality-adaptive KV cache eviction strategy designed to optimize the inference efficiency of MLLMs in long-context scenarios. MadaKV leverages modal-

ity preferences to guide the granularity of token eviction by attention heads and employs inter-layer compensation to dynamically adjust eviction ratios across layers based on overall modality complexity. Through extensive experiments on a diverse set of multimodal long-context tasks using the representative MileBench benchmark, we demonstrate the effectiveness of MadaKV. Looking ahead, we plan to explore the integration of MadaKV with other MLLMs inference acceleration techniques to further enhance efficiency and performance.

6 Limitation

The primary limitations of our approach are as follows: First, due to resource constraints, we have not yet conducted experiments on models with larger parameter sizes (e.g., 34B, 70B) or on datasets with extremely long contexts. However, based on our current experimental analysis, MadaKV appears to be scalable and adaptable to a variety of application scenarios. Additionally, this study has focused on the common visual and textual modalities within MLLMs and has not yet explored other modalities (e.g., video, audio). These areas will be prioritized as future directions for exploration.

7 Acknowledgements

This work was supported by the Key Research and Development Program of Zhejiang Province (No. 2024C03270), and the National Natural Science Foundation of China (No. 62402429, U24A20326, 62441236). This work was also partially supported by the ZJU Kunpeng&Ascend Center of Excellence, the Ningbo Yongjiang Talent Introduction Programme (No. 2023A-397-G) and the Young Elite Scientists Sponsorship Program by CAST (No. 2024QNRC001).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Baobao Chang, Junjie Hu, et al. 2024. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

Yilong Chen, Guoxia Wang, Junyuan Shang, Shiyao Cui, Zhenyu Zhang, Tingwen Liu, Shuohuan Wang, Yu Sun, Dianhai Yu, and Hua Wu. 2024. Nacl: A general and effective kv cache eviction framework for llms at inference time. *arXiv preprint arXiv:2408.03675*.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. 2023. Model tells you what to discard: Adaptive kv cache compression for llms. *arXiv preprint arXiv:2310.01801*.

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2024. Lm-infinite: Zero-shot extreme length generalization for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3991–4008.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. 2024. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *arXiv preprint arXiv:2401.18079*.

Mehrdad Hosseinzadeh and Yang Wang. 2021. Image change captioning by learning from an auxiliary task.

In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2725–2734. Computer Vision Foundation / IEEE.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, pages 709–727. Springer.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4024–4034. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zixiao Kong, Xianquan Wang, Shuanghong Shen, Keyu Zhu, Huibo Xu, and Yu Su. 2025. Scholargec: Enhancing controllability of large language model for chinese academic grammatical error correction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24339–24347.

Dongxu Li, Junnan Li, and Steven Hoi. 2024. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36.

Yongqi Li, Wenjie Li, and Liqiang Nie. 2022. Mmcoqa: Conversational question answering over text, tables, and images. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4220–4231. Association for Computational Linguistics.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024a. World model on million-length video and language with blockwise ringattention. *CoRR*.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. 2024c. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. 2021. {Zero-offload}: Democratizing {billion-scale} model training. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*, pages 551–564.

Minsoo Rhu, Natalia Gimelshein, Jason Clemons, Arslan Zulfiqar, and Stephen W Keckler. 2016. vdnn: Virtualized deep neural networks for scalable, memory-efficient neural network design. In *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 1–13. IEEE.

Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Beidi Chen, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. 2023. Flexgen: High-throughput generative inference of large language models with a single gpu. In *International Conference on Machine Learning*, pages 31094–31116. PMLR.

Luohe Shi, Hongyi Zhang, Yao Yao, Zuchao Li, and Hai Zhao. 2024. Keep the cost down: A review on methods to optimize llm's kv-cache consumption. *arXiv preprint arXiv:2407.18003*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10737–10746. Computer Vision Foundation / IEEE.

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

Lin Song, Yukang Chen, Shuai Yang, Xiaohan Ding, Yixiao Ge, Ying-Cong Chen, and Ying Shan. 2024b. Low-rank approximation for sparse attention in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13763–13773.

Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. 2019. Expressing visual relationships via language. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1873–1883. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. 2024. Focused transformer: Contrastive training for context scaling. *Advances in Neural Information Processing Systems*, 36.

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*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Xianquan Wang, Likang Wu, Zhi Li, Haitao Yuan, Shuanghong Shen, Huibo Xu, Yu Su, and Chenyi Lei. 2025. Mitigating redundancy in deep recommender systems: A field importance distribution perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, page 1515–1526, New York, NY, USA. Association for Computing Machinery.

A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. 2017. Attention is all you need. In *NIPS*.

Bo Wu, Shoubin Yu, Zhenfang Chen, Josh Tenenbaum, and Chuang Gan. 2021. STAR: A benchmark for situated reasoning in real-world videos. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

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

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023b. Efficient streaming language models with attention sinks. *ArXiv*, abs/2309.17453.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Dongjie Yang, XiaoDong Han, Yan Gao, Yao Hu, Shilin Zhang, and Hai Zhao. 2024b. Pyramidinfer: Pyramid kv cache compression for high-throughput llm inference. *arXiv preprint arXiv:2405.12532*.

June Yong Yang, Byeongwook Kim, Jeongin Bae, Beomseok Kwon, Gunho Park, Eunho Yang, Se Jung Kwon, and Dongsoo Lee. 2024c. No token left behind: Reliable kv cache compression via importance-aware mixed precision quantization. *arXiv preprint arXiv:2402.18096*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Rongzhi Zhang, Kuang Wang, Liyuan Liu, Shuohang Wang, Hao Cheng, Chao Zhang, and Yelong Shen. 2024a. Lorc: Low-rank compression for llms kv cache with a progressive compression strategy. *arXiv preprint arXiv:2410.03111*.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. 2024b. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36.

Guangxiang Zhao, Junyang Lin, Zhiyuan Zhang, Xuancheng Ren, Qi Su, and Xu Sun. 2019. Explicit sparse transformer: Concentrated attention through explicit selection. *arXiv preprint arXiv:1912.11637*.

Yiyun Zhou, Wenkang Han, and Jingyuan Chen. 2025a. Revisiting applicable and comprehensive knowledge tracing in large-scale data. *arXiv preprint arXiv:2501.14256*.

Yiyun Zhou, Zheqi Lv, Shengyu Zhang, and Jingyuan Chen. 2024. Cuff-KT: Tackling learners' real-time learning pattern adjustment via tuning-free knowledge state-guided model updating.

Yiyun Zhou, Zheqi Lv, Shengyu Zhang, and Jingyuan Chen. 2025b. Cuff-kt: Tackling learners' real-time learning pattern adjustment via tuning-free knowledge state guided model updating. *Preprint*, arXiv:2505.19543.

Yiyun Zhou, Zheqi Lv, Shengyu Zhang, and Jingyuan Chen. 2025c. Disentangled knowledge tracing for alleviating cognitive bias. In *Proceedings of the ACM on Web Conference 2025*, pages 2633–2645.

Yiyun Zhou, Chang Yao, and Jingyuan Chen. 2025d. Cola: Collaborative low-rank adaptation. *Preprint*, arXiv:2505.15471.

A Appendix

A.1 The Significance of Modalities Across Different Tasks

We provide an example to illustrate the substantial differences in modality importance across vari-

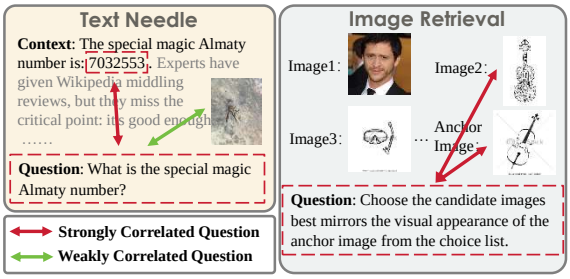

Figure 5: Examples of modal significance in different tasks.

ous tasks. As shown in Figure 5, the Text Needle task focuses on locating the corresponding answer within the text, while the Image Retrieval task centers on comparing the similarity between images. These examples highlight how the significance of modalities varies significantly depending on the specific task requirements.

A.2 Details of MileBench

MileBench (Song et al., 2024a) dataset is the first benchmark specifically designed to test the Multimodal Long-context capabilities of MLLMs. Milebench primarily includes 6,440 multimodal long-text samples, which are composed of 21 existing or self-constructed datasets, with an average of 15.2 images and 422.3 words per sample. The detailed information of dataset is presented in Table 4.

Dataset Abbr.	Task	Data Source
TN	Text Needle In A Haystack	TextNeedleInAHaystack
IEdit	Visual Relationship Expressing	IEdit (Tan et al., 2019)
MMCoQA	Multimodal Dialogue	MMCoQA (Li et al., 2022)
STD	Visual Change Captioning	Spot-the-Diff (Jhamtani and Berg-Kirkpatrick, 2018)
ALFRED	Conversational Embodied Dialogue	ALFRED (Shridhar et al., 2020)
CLEVR-C	Visual Change Captioning	CLEVR-Change (Hosseinzadeh and Wang, 2021)
DocVQA	Document QA	DocVQA (Mathew et al., 2021)
ST	Scene Transition	MovieNet (Huang et al., 2020)
OI	Object Interaction	STAR (Wu et al., 2021)

Table 4: Detailed Statistics and Taxonomy of dataset.

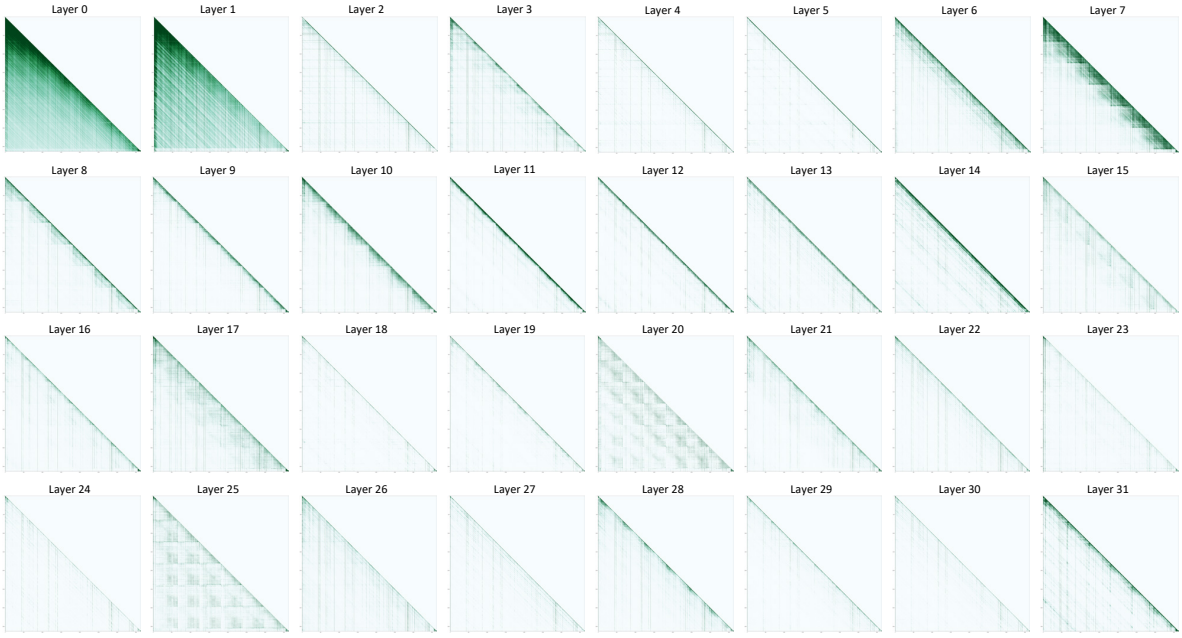

Figure 6: Attention patterns of LLava-v1.5-7B on ALFRED.