# BiKV: Bi-granularity KV-Cache Optimization with Adjacent-layer Reuse and Modality-Specific Pruning

Anonymous IJCNN submission

***Abstract*—Multimodal large language models require efficient inference under complex and long-context scenarios. KV-cache is widely used to accelerate decoding and reduce memory consumption. However, existing approaches often fail to accurately identify critical tokens, overlook inter-layer redundancy, and ignore modality-specific characteristics. To address these limitations, we propose BiKV, a unified KV-cache optimization framework that jointly operates at inter-layer and intra-layer levels. At the inter-layer level, BiKV reuses key–value pairs across highly similar adjacent layers to eliminate redundant computation. At the intra-layer level, it leverages cross-modal attention to select informative tokens, ensuring the preservation of essential multimodal information. Experiments on representative long-context tasks from the MileBench benchmark demonstrate that BiKV significantly reduces memory usage and inference latency while maintaining competitive reasoning performance, validating its effectiveness for efficient multimodal inference. We will make our data/code available upon acceptance.**

***Index Terms*—KV Cache, Cross-modal Attention, Inter-layer Reuse, Efficient Inference**

## I. Introduction

In recent years, with the widespread application of Multimodal Large Language Models (MLLMs) [1]–[5] in vision–language tasks, achieving efficient inference under limited computational resources has become a critical challenge. Such models typically need to process high-dimensional image features jointly with textual sequences, imposing substantial demands on both memory and computation. Prior work has focused on reducing memory requirements through the selective pruning of tokens, aiming to preserve those that are most significant [6], [7]. Consequently, a central problem in this area lies in accurately determining which tokens should be retained and which can be safely discarded, while preserving the original inference performance as much as possible. This requires a careful balance between efficiency and accuracy, as well as robust criteria for token importance that generalize across different modalities, tasks, and input distributions.

Against this backdrop, various Key–Value (KV) cache optimization techniques have been proposed. H2O [6] leverages the cumulative attention score of each token during decoding as an importance metric to progressively prune redundant tokens. SnapKV [7], in contrast, applies pooling operations to smooth the attention distribution, followed by nearest-neighbor voting to identify critical tokens. However, these methods [6]–[8]were originally designed for unimodal text, overlooking the inherent heterogeneity between visual and textual tokens in multimodal contexts—image tokens usually exhibit stronger spatial structures and higher semantic density, whereas text

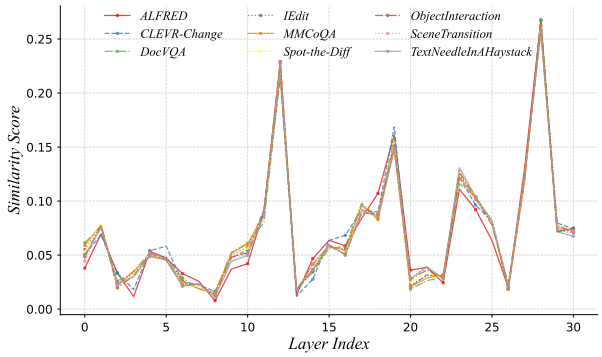


Fig. 1. Similarity between each adjacent layer on different datasets.

tokens tend to carry long-range sequential dependencies. Treating them identically often undermines optimization effectiveness. The LOOK-M [9] approach partially addresses this issue by prioritizing text tokens based on retention scores, while clustering and pruning image tokens. Nevertheless, its reliance on heuristic, modality-specific priors introduces clear limitations: (i) within the text modality, a large portion of tokens may still be redundant or carry low informational content, making a text-prioritized retention strategy suboptimal; (ii) its recent-window selection mechanism fails to capture cross-modal semantic correspondences, weakening the alignment between visual and textual information. Recently, pruning methods based on cross-modal attention have emerged. While these approaches leverage cross-modal dependencies for model reduction, the definition of their pruning strategies is often inherently complex, which to some extent increases the overall computational overhead during inference.

To systematically address these challenges, we propose a token selection mechanism based on cross-modal attention scores [10], [11] named Modality-Specific Pruning (MSP). By quantifying the interaction strength between textual and visual tokens, our method dynamically identifies tokens that play a pivotal role in multimodal reasoning, thereby enhancing cross-modal fusion quality. Moreover, inspired by [12] techniques in large language models (LLMs), we observe substantial feature redundancy across adjacent layers in multimodal networks. As illustrated in Figure 1, when running inference with LLaVA across different datasets, some similarity indicators of the neighboring layers are significantly higher than those of the other neighboring layers, especially in middle and upper layers. Building on this observation, we design an adjacent-layer KV cache reuse (ALR) mechanism that skips redundant key–value projections, reducing memory consumption by ap-

proximately 56% and nearly doubling inference speed in long-context scenarios, all with negligible performance degradation.

Our approach demonstrates significant advantages across multiple multimodal tasks, including Visual Question Answering (VQA) [13], image captioning, and information retrieval. For instance, in VQA, the model is able to focus cross-modal attention on image regions relevant to the question , while suppressing irrelevant background noise. This enables the model to emphasize key contextual information, thereby improving both accuracy and inference efficiency in practical applications.

## II. RELATED WORKS

### A. Cache Optimization for LLM

The autoregressive decoding nature of LLMs necessitates the maintenance of a KV Cache to avoid redundant computations. However, the linear growth of the KV cache with sequence length imposes significant memory pressure and I/O overhead, becoming a primary bottleneck for long-context generation and system throughput. Current research to mitigate this bottleneck can be categorized into three main trajectories:

Architectural Innovations: From a design perspective, methods like Multi-Query Attention (MQA) [14] and Grouped-Query Attention (GQA) [15] reduce the number of KV heads to shrink the cache footprint. Moreover, Multi-head Latent Attention (MLA), introduced in DeepSeek-V2 [16], employs low-rank compression to project KV vectors into a latent space, further reducing memory requirements while maintaining expressive power.

Dynamic Pruning and Eviction: Leveraging the sparsity of attention distributions, strategies such as H2O [6] and StreamingLLM [8] identify and retain only essential tokens. While H2O focuses on high-contribution "heavy hitters," StreamingLLM maintains "attention sinks" and recent window tokens to achieve a constant cache size for infinite-length generation.

Low-Bit Quantization: Quantization techniques like KIVI [17] and various FP8/INT8 schemes [18] compress the numerical precision of KV pairs. These methods significantly reduce the memory cost with minimal impact on model accuracy, effectively breaking the physical memory limits of hardware.

### B. Challenges in MLLMs

In MLLMs, the optimization of KV caches faces distinct challenges due to the asymmetric information distribution across modalities. Standard LLM pruning techniques often fail to account for the dense, high-redundancy nature of visual tokens compared to sparse textual tokens. Recent work [9], [19] attempts to address this by observing cross-modal attention patterns and prioritizing the eviction or merging of visual tokens while preserving textual context. However, static eviction policies for visual tokens may lack robustness across diverse scenarios, particularly when processing dense textual instructions alongside complex imagery. Unlike prior works that rely on rigid modal priority, our proposed method considers the inherent redundancy within both text and image modalities. By integrating the Transformer architecture's capacity for information filtering, our approach effectively retains task-relevant information across varying contexts, demonstrating superior performance in complex multimodal reasoning tasks.

## III. METHOD

We propose a KV-Cache optimization method named BiKV to reduce the storage overhead during multi-modal large model inference while preserving the integrity of task-relevant information. The overall framework of BiKV is illustrated in Figure 2, and the method primarily consists of two core modules: Adjacent-layer Token Reuse and Modality-Specific Token Pruning.

### A. Adjacent-Layer Reuse

To investigate the inter-layer redundancy of the model, we randomly sample a small portion of data for each subtask and perform a standard inference process on this subset. During this process, we compute the similarity between each pair of adjacent layers, measured specifically by the average Cosine similarity between their key-value vectors. The inter-layer similarity reflects, to some extent, the degree of information redundancy, thus providing guidance for subsequent layer reuse strategies. For attention layer pairs with high similarity, we consider them to contain substantial redundant information, and optimizing them will not lead to significant degradation in model performance. Therefore, computational resources can be effectively reduced by optimizing these components.

Let the key states and value states of the $l$-th layer be $\mathbf{K}^{(l)} \in \mathbb{R}^{n \times d_h}$ and $\mathbf{V}^{(l)} \in \mathbb{R}^{n \times d_h}$, respectively, where $n$ is the sequence length and $d_h$ is the header's dimensionality. We first flatten the key-value vectors into a one-dimensional representation and then employ the Cosine Similarity to measure the feature similarity between two layers during the initial inference phase:

$$S^{(l,l+1)} = \frac{1}{2}\Big(\mathrm{Sim}(\mathbf{K}^{(l)}, \mathbf{K}^{(l+1)}) + \mathrm{Sim}(\mathbf{V}^{(l)}, \mathbf{V}^{(l+1)})\Big) \quad (1)$$

This similarity $S^{(l,l+1)}$ characterizes the consistency between the two layers within the representation space. $\mathcal{P} = \{(l, l+1) \mid l = 1, 2, \ldots, L-1\}$,A high similarity between two layers indicates that their contributions to downstream decoding are highly overlapping, and repetitively storing their Key-Value (KV) caches would lead to GPU memory redundancy. The top-m most similar layer pairs are selected to construct the reuse layer set, which provides guidance for eliminating inter-layer redundancy during subsequent inference.

$$\mathcal{P}_m = \text{Top-m}_{(l,l+1)\in\mathcal{P}} \; S^{(l,l+1)} \quad (2)$$

For each dataset corresponding to a distinct task, we analyze the reuse set by considering the top $i$ elements ($i = 1, 2, \ldots, m$). Specifically, we iteratively select the reuse pairs with the highest occurrence frequencies and organize them

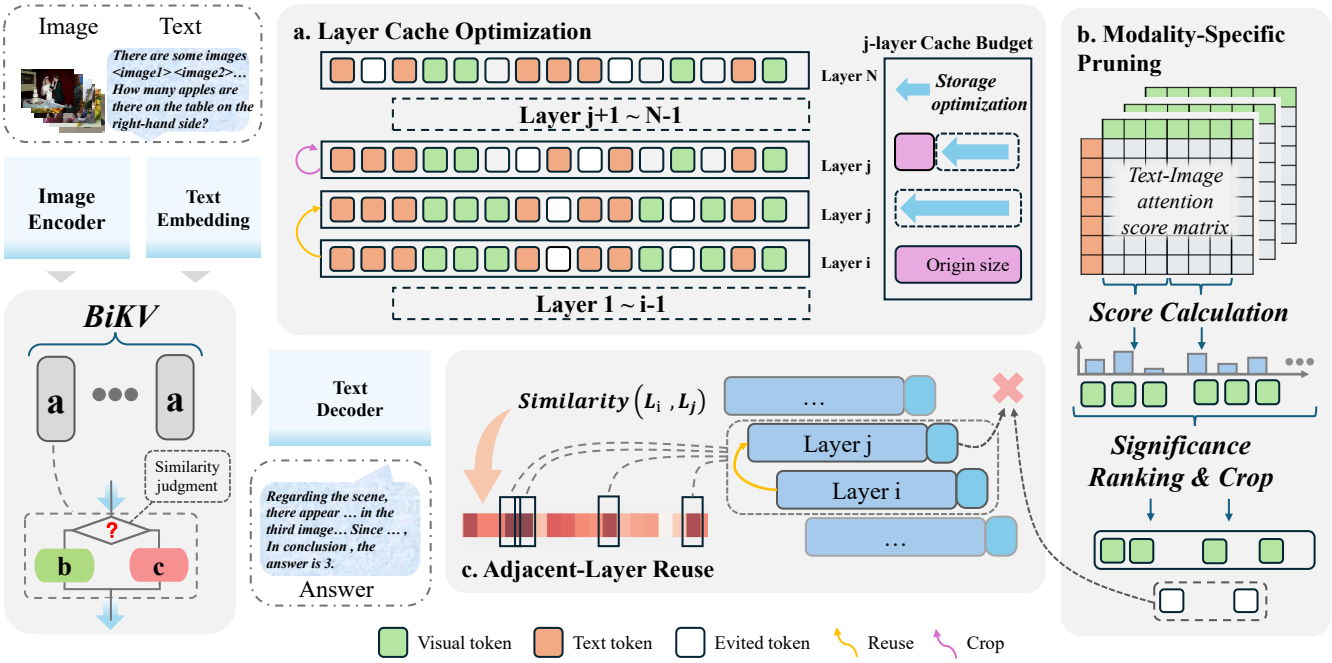


Fig. 2. **Overview of the BiKV framework.** BiKV consists of two functional modules. (a) For each layer of the model, execute one of the two methods: (b) Modality-specific pruning (Pruning of visual tokens is illustrated; the same operation applies to text tokens); (c) Adjacent-layer reuse by directly reusing the key-value pairs computed from lalst layer.

into a candidate reuse queue in descending order. Once the reuse depth for a given experiment is determined, we extract the top-ranked reuse pairs from this queue and adopt them as the reuse strategy for that experiment.

Therefore, during the main inference phase, For these $m$ layer pairs, the KV of the subsequent layer is not redundantly stored but directly reused from the preceding layer.

$$\mathbf{K}^{(l+1)} \leftarrow \mathbf{K}^{(l)}, \mathbf{V}^{(l+1)} \leftarrow \mathbf{V}^{(l)} \quad \forall (l, l+1) \in \mathcal{P}_m \tag{3}$$

This selective reuse strategy can significantly reduce redundant storage while preserving the expressive power of the information.

### *B. Modality-Specific Pruning*

For layers without a reuse strategy, a considerable amount of redundant information still remains to be optimized. Such redundancy is mainly reflected in the key-value vectors within the same layer, where features from different modalities often exhibit overlapping information [8] [21], thereby introducing additional computational overhead during inference. To address this issue, we propose a modality-based key-value pruning method specifically designed for optimizing intra-layer key-value pairs. This approach not only focuses on compressing representations within each modality but also fully accounts for the interactions between textual and visual information. In particular, we design a mutual selection mechanism between text tokens and visual tokens, where cross-modal attention scores are leveraged to identify and preserve the most contributive tokens while discarding those with limited impact on downstream predictions. In this way, the method effectively reduces redundant features while maintaining the model's representational capacity, thereby improving inference efficiency.

For every layer that is not reused in the model, We denote $A_{\mathrm{t}\to\mathrm{v}} \in [0,1]^{L_t \times L_v}$ and $A_{\mathrm{v}\to\mathrm{t}} \in [0,1]^{L_v \times L_t}$ as the visual-text and text-visual cross-attention respectively and define the bidirectional cross-modal attention of visual tokens and text token as:

$$\begin{aligned} A_{\mathrm{t}\to\mathrm{v}} &= \mathrm{Softmax}\left(\frac{Q_{\mathrm{t}} K_{\mathrm{v}}^{\top}}{\sqrt{d}}\right), \\ A_{\mathrm{v}\to\mathrm{t}} &= \mathrm{Softmax}\left(\frac{Q_{\mathrm{v}} K_{\mathrm{t}}^{\top}}{\sqrt{d}}\right) \end{aligned} \tag{4}$$

Such interaction highlights tokens with higher contributions to cross-modal alignment.We select a token set $\mathcal{T}_{\text{keep}}$ according to the above attention scores, retaining those with higher task relevance and then retain the original text and image tokens in proportions of $\alpha$ and $\beta$, respectively.

$$\begin{aligned} \mathcal{T}_{\text{keep}} &= \text{Top-}\alpha\Big(\sum_{|v|} A_{v\to t}\Big), \\ \mathcal{V}_{\text{keep}} &= \text{Top-}\beta\Big(\sum_{|t|} A_{t\to v}\Big) \end{aligned} \tag{5}$$

To guarantee robust inference performance, the complete tail window $\mathcal{I}_{\text{recent}}$ is preserved, The optimized key-value index $\mathcal{I}_{\text{opt}}$ set is obtained by concatenating the resulting sequences:

$$\mathcal{I}_{\text{opt}} = \text{Concat}\big(\mathcal{T}_{\text{keep}},\ \mathcal{V}_{\text{keep}},\ \mathcal{I}_{\text{recent}}\big) \tag{6}$$

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT METHODS. SOME RESULTS ORIGINATE FROM MADAKV [20].

| Method | Evaluation Subtasks | | | | | | | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| | TN | IEdit | MMCoQA | STD | ALFRED | CLEVR-C | DocVQA | ST | OI | |
| ***LLaVA-1.5-7B*** | | | | | | | | | | |
| Origin | 9.68 | 7.98 | 33.5 | 16.32 | 15.42 | 16.62 | 46 | 66.5 | 48.5 | 28.95 |
| StreamingLLM | 3.12 | 3.59 | 26 | 11.77 | 3.73 | 10.44 | 42.5 | 43 | 44 | 20.91 |
| H2O | 2.5 | 5.51 | 28 | 15.73 | 14.86 | 14.07 | 44 | 63.5 | 44 | 25.8 |
| SnapKV | 3.27 | 6.03 | 29 | 14.82 | 14.4 | 15.37 | 45.5 | 64 | 45.5 | 26.43 |
| LOOK-M | 3.34 | 6.51 | 29.4 | 15.79 | 13.96 | 14.12 | 45.5 | 63 | 46.5 | 26.47 |
| BiKV | **9.38** | 5.16 | **29.5** | **17.5** | **17.31** | **17.34** | **46.5** | **67.5** | **48** | **28.61** |
| ***LLaVA-1.5-13B*** | | | | | | | | | | |
| Origin | 24.38 | 9.05 | 40.5 | 15.62 | 18.36 | 16.12 | 46.5 | 75.5 | 47 | 32.56 |
| StreamingLLM | 7.81 | 2.63 | 28.5 | 14.05 | 1.72 | 6.55 | 46.5 | 68 | 47 | 24.75 |
| H2O | 7.81 | 8.7 | 34.5 | 14.9 | 15.03 | 15.06 | 46.5 | 74 | 47 | 29.28 |
| SnapKV | 7.92 | 8.52 | 35 | 15.53 | 16.84 | 15.37 | 46 | 74.5 | 47 | 29.63 |
| LOOK-M | 10 | 8.77 | 37.5 | 15.46 | 15.25 | 14.12 | 47 | 74 | 47 | 29.9 |
| BiKV | **19.69** | 7.84 | **39** | **15.59** | **17.61** | **15.68** | 47 | **75** | 47 | **31.61** |

This strategy maintains long-range dependency modeling while reducing storage redundancy. The experimental results indicate that the proposed modality-aware pruning strategy effectively reduces memory usage and computational cost.

The entire reasoning process can be summarized as follows: Firstly, a similarity judgment is performed. For cases of high similarity, adjacent layer reuse is implemented. For non-reused layers, we apply a pruning method based on cross-modal attention. Finally, the optimized key-value pairs are obtained.

TABLE II
COMPARISON OF DIFFERENT CACHING METHODS.

| **Method** | **Config** | | **Decoding Latency (ms/token)** | **Used Memory (GB)** |
|---|---|---|---|---|
| Full Cache | R0(1,1) | | 40.95 | 1.78 |
| BiKV | R0 | (0.9,0.1) | 21.63 | 0.77 |
| | R2 | | 21.61 | 0.73 |
| | R4 | | 21.58 | 0.67 |
| | R4 | (0.9,0.2) | 21.75 | 0.77 |
| | | (0.8,0.2) | 21.72 | 0.72 |
| | | (0.9,0.1) | 21.58 | 0.67 |
| | | (0.8,0.1) | 21.54 | 0.63 |

## IV. EXPERIMENTS

### A. Experimental Setup

To evaluate the effectiveness of the proposed BiKV method, we conduct experiments on MileBench [22], the first benchmark specifically designed for long-context scenarios, which consists of multiple sub-tasks for simulating distinct settings. Compared with other benchmarks [23] [24], it features richer contextual information, longer textual sequences, and a greater number of images, which makes it well-suited for evaluating BiKV in real-world scenarios.

For our study, we select nine representative sub-tasks from MileBench, including ALFRED, CLEVR-C, DocVQA [13], ObjectInteraction(OI), IEdit, MMCoQA, Spot-the-Diff(STD), SceneTransition(ST) and TextNeedleInAHaystack(TN), covering a diverse range of scenarios such as visual relation reasoning, natural language understanding, image retrieval, and document question answering, which comprehensively evaluates the model's capabilities in capturing visual details, generating responses, reasoning over complex scenarios, achieving cross-modal consistency, and performing causal inference. We evaluate BiKV using LLaVA-v1.5-7B/13B [25] [26], and compare its performance against prior KV-cache optimization approaches such as H2O [6], StreamingLLM [8], SnapKV [7], and LOOK-M [9], covering both unimodal and multimodal settings. All experiments were conducted on RTX 4090D and vGPU-48GB.

### B. Experiment Result

As shown in Table I, From the perspective of overall average performance, BiKV is able to maintain performance comparable to, and in some cases slightly superior to, that of the original model even after applying memory optimization. This demonstrates that BiKV effectively balances inference efficiency and overall model performance while substantially reducing KV storage requirements. At the level of individual subtasks, BiKV exhibits particularly strong performance on multiple vision–language understanding tasks that heavily rely on fine-grained image–text interactions. Under both the 7B and 13B model configurations, BiKV consistently achieves the best or second-best results among compression methods, indicating that it does not merely preserve performance but can even yield performance gains on certain tasks through more efficient information retention mechanisms. Compared with methods that optimize only a single modality, BiKV demonstrates clear advantages across nearly all subtasks, especially in cross-modal reasoning scenarios. This suggests that KV pruning

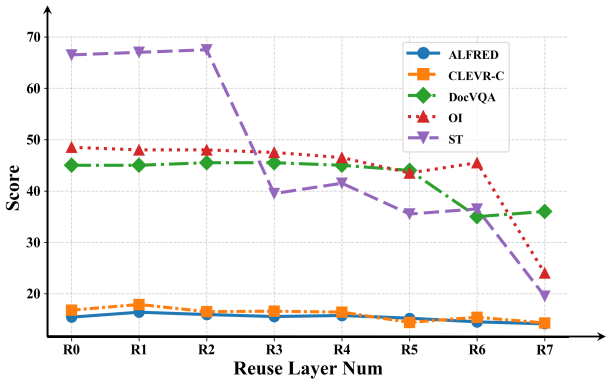


Fig. 3. The results with different numbers of reused layers.

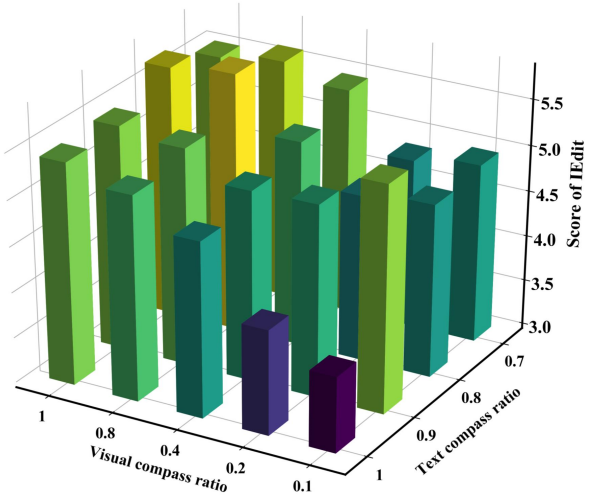


Fig. 4. The results under different compression rates.

strategies based solely on textual information or local context are prone to discarding critical cues, whereas BiKV's bimodal redundancy elimination strategy better preserves discriminative cross-modal context. Furthermore, although LOOK-M leverages textual priors and achieves moderate improvements on some tasks, BiKV consistently outperforms it in overall performance as well as on the majority of subtasks. This advantage can be attributed to BiKV's bidirectional selection mechanism between text and image tokens, which jointly integrates query and contextual information and thus enables more comprehensive modeling of cross-modal dependencies. Overall, the experimental results further confirm that, for KV compression in multimodal large models, explicitly modeling the heterogeneity and interactions across different modalities is crucial for achieving efficient and high-performance inference.

In Table II, we present the effectiveness analysis of the BiKV method. Compared with the full-cache baseline, BiKV consistently reduces inference latency to around 21 ms/token, while lowering memory consumption from 1.78 GB to below 0.77 GB. This demonstrates that our approach not only nearly doubles the decoding speed but also significantly decreases the memory requirements during inference. Specifically, the ALR mechanism effectively mitigates cross-layer redundancy,

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT KV COMPRESSION METHODS ON QWEN2.5VL-7B.

| Qwen2.5VL | ALFRED | CLEVR-C | IEdit | ST | STD |
|---|---|---|---|---|---|
| Origin | 37.53 | 42.46 | 29.45 | 63.00 | 28.36 |
| Streaming LLM | 34.07 | 27.58 | 25.12 | 59.00 | 26.14 |
| H2O | 37.78 | 36.03 | 29.13 | 61.50 | 26.79 |
| SnapKV | 37.41 | 39.41 | 29.25 | 60.00 | 26.54 |
| Look-m | 36.66 | 38.66 | 28.73 | 61.50 | 28.97 |
| BiKV | 36.73 | 42.54 | 30.64 | 61.50 | 27.53 |

TABLE IV
ABLATION STUDY OF THE EFFECT OF TWO INDIVIDUAL MODULES.

| ALR | MSP | ALFERD | ST | DocVQA |
|---|---|---|---|---|
| × | × | 15.42 | 66.5 | 46 |
| ✓ | × | 15.93 | 67.5 | 45.5 |
| × | ✓ | 17.15 | 68.5 | 46 |
| ✓ | ✓ | 17.31 | 67.5 | 46.5 |

leading to substantial memory savings, while MSP further eliminates redundancy by balancing the retention of text and visual tokens. By combining these two strategies, BiKV compresses GPU memory usage and halves the decoding latency, confirming its effectiveness for efficient multimodal inference.

To further validate the effectiveness of our proposed method BiKV across different model families, we conduct additional experiments on the Qwen2.5VL [5] model using several representative benchmarks, as summarized in Table III. Overall, BiKV demonstrates consistently strong performance compared with other methods. Specifically, BiKV achieves the best or near-best results on challenging benchmarks such as ALFRED, CLEVR-C, and IEdit, indicating its superior ability to preserve critical visual and linguistic information under KV compression. Compared to the Origin and Streaming LLM baselines, BiKV maintains competitive accuracy while benefiting from more efficient memory usage. Moreover, when compared with other compression strategies, BiKV shows a clear advantage in overall robustness, particularly on CLEVR-C and IEdit, where reasoning over long contexts is essential. These results suggest that BiKV generalizes well to the Qwen2.5-VL model series and effectively balances efficiency and performance, further confirming the reliability and applicability of our approach across different vision-language architectures.

### C. Ablation Study

We conduct an ablation study to examine the roles of ALR and MSP in the optimization process. As shown in Table IV, applying ALR alone reveals that inter-layer redundancy has a notable impact on inference. Reusing highly similar key-value pairs effectively reduces such redundancy without degrading performance. In contrast, applying MSP alone demonstrates that pruning based on cross-modal attention is highly effective in filtering intra-layer redundancy, thereby enhancing the model's ability to capture salient multimodal interactions. This demonstrates that the two strategies are complementary:

ALR reduces redundancy across layers, while MSP focuses on intra-layer token efficiency. We also investigate the impact of different numbers of reuse layers and compression rates on the model's performance scores. As shown in Figure 3 and Figure 4, this analysis will guide us in adopting different parameter configurations for specific subtasks or application scenarios, aiming to achieve an optimal balance between performance and resource consumption.

We generally recommend setting the number of reused layers to 1-3, with the compression ratios for text and image tokens set to 0.7-0.9 and 0.1-0.3. However, certain subtasks are not sensitive to the number of reused layers or the compression ratio (as shown in Figure 3, CLEVR-C and ALFRED exhibit no significant performance degradation even with more reused layers, indicating a high similarity between the original and reused key-value pairs). Therefore, this configuration can be dynamically adjusted according to the specific characteristics of each subtask.

## V. CONCLUSION

In this work, we propose BiKV, a KV-Cache optimization method designed for long-context scenarios in multi-modal large models. By eliminating information redundancy from two dimensions, this approach facilitates the interaction and alignment of information across different modalities. Experimental results demonstrate that BiKV significantly reduces storage overhead while maintaining stable model performance and without significant degradation.

## REFERENCES


[1] Wenhai Wang, Zhe Chen, Yangzhou Liu, Yue Cao, Weiyun Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, and Jifeng Dai, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Large Vision-Language Models: Pre-training, Prompting, and Applications*, pp. 23–57. Springer, 2025.

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou, "Qwen-vl: A frontier large vision-language model with versatile abilities. corr, vol. abs/2308.12966 (2023)," 2023.

[3] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al., "Mobilevlm v2: Faster and stronger baseline for vision language model," *arXiv preprint arXiv:2402.03766*, 2024.

[4] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al., "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al., "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

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

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

[8] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis, "Efficient streaming language models with attention sinks," in *The Twelfth International Conference on Learning Representations*. The Twelfth International Conference on Learning Representations, 2024.

[9] Zhongwei Wan, Ziang Wu, Che Liu, Jinfa Huang, Zhihong Zhu, Peng Jin, Longyue Wang, and Li Yuan, "Look-m: Look-once optimization in kv cache for efficient multimodal long-context inference," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 4065–4078.

[10] Zhongwei Wan, Hui Shen, Xin Wang, Che Liu, Zheda Mai, and Mi Zhang, "Meda: Dynamic kv cache allocation for efficient multimodal long-context inference," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025, pp. 2485–2497.

[11] Xiaohuan Pei, Tao Huang, and Chang Xu, "Cross-self kv cache pruning for efficient vision-language inference," *arXiv preprint arXiv:2412.04652*, 2024.

[12] Yifei Yang, Zouying Cao, Qiguang Chen, Libo Qin, Dongjie Yang, Hai Zhao, and Zhi Chen, "Kvsharer: Efficient inference via layer-wise dissimilar kv cache sharing," *arXiv e-prints*, pp. arXiv–2410, 2024.

[13] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar, "Docvqa: A dataset for vqa on document images," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 2200–2209.

[14] Noam Shazeer, "Fast transformer decoding: One write-head is all you need," *arXiv preprint arXiv:1911.02150*, 2019.

[15] Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai, "Gqa: Training generalized multi-query transformer models from multi-head checkpoints," *arXiv preprint arXiv:2305.13245*, 2023.

[16] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al., "Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model," *arXiv preprint arXiv:2405.04434*, 2024.

[17] Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu, "Kivi: A tuning-free asymmetric 2bit quantization for kv cache," *arXiv preprint arXiv:2402.02750*, 2024.

[18] Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Yakun S Shao, Kurt Keutzer, and Amir Gholami, "Kvquant: Towards 10 million context length llm inference with kv cache quantization," *Advances in Neural Information Processing Systems*, vol. 37, pp. 1270–1303, 2024.

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

[20] Kunxi Li, Zhonghua Jiang, Zhouzhou Shen, ZhaodeWang ZhaodeWang, Chengfei Lv, Shengyu Zhang, Fan Wu, and Fei Wu, "MadaKV: Adaptive modality-perception KV cache eviction for efficient multi-modal long-context inference," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, Eds., Vienna, Austria, July 2025, pp. 13306–13318, Association for Computational Linguistics.

[21] Lin Song, Yukang Chen, Shuai Yang, Xiaohan Ding, Yixiao Ge, Ying-Cong Chen, and Ying Shan, "Low-rank approximation for sparse attention in multi-modal llms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13763–13773.

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

[23] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al., "Mmbench: Is your multi-modal model an all-around player?," in *European conference on computer vision*. Springer, 2024, pp. 216–233.

[24] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," *arXiv preprint arXiv:2307.16125*, 2023.

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34892–34916, 2023.

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 26296–26306.