Improving Recall in Efficient Visual Language Models

Anonymous ACL submission

Abstract

Associative recall has emerged as a critical weakness in efficient language models and, as we demonstrate, is also a core bottleneck in efficient visual language models (VLMs). In this work, we show that efficient VLMs—exemplified by VisualRWKV—suffer from significant deficits in recall, particularly in text-centric tasks such as TextVQA and document understanding. Quantitatively, the baseline VisualRWKV-7B still trails the Transformer-based LLaVA-1.5-7B by 7.2 accuracy points on the TextVQA benchmark. We attribute this gap to a fundamental architectural limitation: insufficient input feature quality. To address this, we propose two effective processing strategies to enhance visual feature representations. First, our model incorporates SigLIP, DINOv2, and SAM to improve feature richness across resolutions, enabling the retention of multi-scale visual information without increasing the number of input visual tokens. Second, we introduce a segmentation-recombination strategy that supports ultra-high-resolution inputs (up to 4096×4096), allowing for precise and detailed feature extraction. These improvements significantly enhance recall performance and feature quality, enabling VisualRWKV-Boost-1.6B to outperform the larger baseline VisualRWKV-7B. Moreover, the performance gap on TextVQA compared with LLaVA-1.5-7B is reduced from 7.2 to just 1.9 accuracy points, paving the way for more scalable and efficient VLM architectures.

1 Introduction

The emergence of efficient language models with linear time complexity, such as Mamba (Gu and Dao, 2023) and RWKV (Peng et al., 2023, 2024, 2025), has opened new avenues for scaling large language models (LLMs) with reduced computational overhead. These models achieve competitive performance on various NLP tasks while offering significant efficiency advantages over traditional attention-based architectures. However, recent studies have identified a critical limitation: their poor performance on tasks requiring associative recall (Arora et al., 2023). This capability—essential for retrieving relevant information from long and complex sequences—remains a key bottleneck for efficient language models.

In this work, we investigate this limitation in the context of visual language modeling and show that associative recall is also a core weakness of efficient visual language models (VLMs). Specifically, we study VisualRWKV (Hou et al., 2024), a representative efficient VLM, and observe that it performs comparably or even better than Transformer-based counterparts on general visual-language tasks such as VQA (Antol et al., 2015) and ScienceQA (Lu et al., 2022a), as shown in Table 3. However, on text-centric benchmarks like TextVQA (Singh et al., 2019)—which demand precise recall of visual textual content—VisualRWKV underperforms significantly, trailing LLaVA-1.5 (Liu et al., 2023a) by 7.2 accuracy points (see Figure 1). These findings suggest that associative recall is not only a challenge in pure language modeling but also a limiting factor in visual-language modeling.

We attribute this performance gap primarily to the limited quality of image feature representations, which constrains the model's ability to retrieve relevant visual information—especially in text-rich scenarios such as TextVQA. To address this, we propose two complementary strategies:

- First, we improve the richness and quality of image features by integrating a ensemble of state-of-the-art vision encoders, including SigLIP (Zhai et al., 2023), DINOv2 (Oquab et al., 2023), and SAM (Kirillov et al., 2023). These encoders are further supported by a segmentation-recombination pipeline that enables ultra-high-resolution image inputs (up

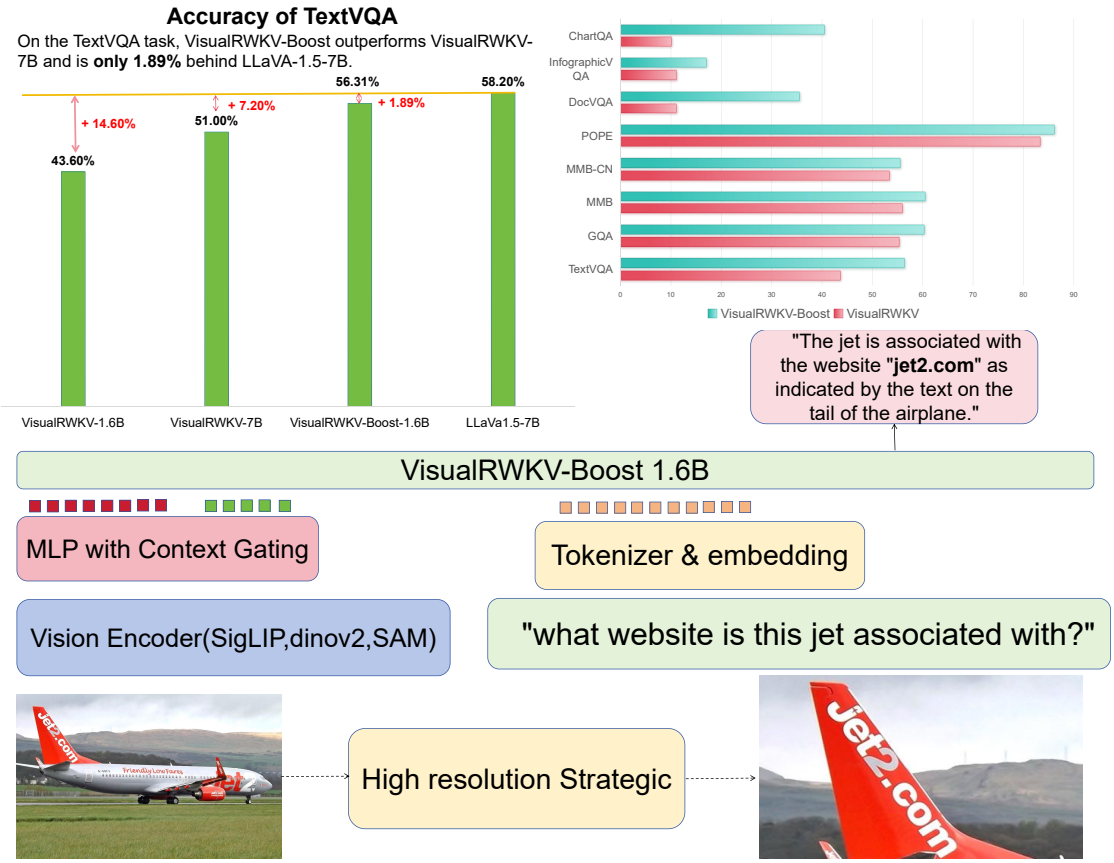

Figure 1: The performance enhancement and workflow of the VisualRWKV-Boost model when dealing with high-resolution images. The bar chart in the upper left corner compares the accuracy of different models on the TextVQA task, with VisualRWKV-Boost leading at an accuracy rate of 56.31%. The bar graph in the upper right corner contrasts the performance of VisualRWKV and VisualRWKV-Boost across multiple visual question answering benchmarks, highlighting the superior performance of VisualRWKV-Boost across various tasks. The flowchart at the bottom demonstrates the workflow of the VisualRWKV-Boost model: it uses visual encoders to extract image features, optimizes feature integration through MLP with context gating, and then converts the textual question into a format that the model can process. High-resolution processing helps the model accurately identify text in images, thereby improving the accuracy of responses.

to 4096×4096), thereby generating highly detailed, multi-scale visual representations while maintaining token efficiency.

- Second, we perform data scaling by incorporating a substantially larger and more diverse corpus of high-quality image-text pairs. This expanded training set improves the model's ability to align visual and textual modalities, thereby enhancing its robustness and generalization across visual-language tasks.

Combined, these improvements lead to a significant boost in associative recall performance. Our proposed VisualRWKV-Boost (1.6B) surpasses the larger 7B VisualRWKV baseline (Hou et al., 2024) by 5.3 accuracy points and substantially closes the performance gap with the Transformer-based LLaVA-1.5 (Liu et al., 2023a), reducing it from 7.2 to just 1.9 points on the TextVQA benchmark.

Our study reveals that the architectural limitations of efficient Visual Language Models—particularly in associative recall—can be effectively mitigated through a combination of feature scaling and data scaling, enabling them to achieve performance on par with much larger Transformer-based models. These results not only close the performance gap with Transformer-based

models but also provide a promising direction for developing scalable and resource-efficient Visual Language Models.

2 Related Works

2.1 Efficient Language Models

Among common efficient language models, Mamba(Gu and Dao, 2023) and RWKV(Peng et al., 2023) each have their unique characteristics and capabilities. Mamba is based on the State Space Model (SSM) and uses a data-dependent dynamic decay mechanism to adaptively adjust the information retention period, thereby optimizing its capability in long sequence processing. It also achieves faster training speeds through CUDA kernel optimization and enhances local feature capture with lightweight Token-shift short convolution operations. Mamba is well-suited for tasks involving long sequences, such as DNA sequence analysis and CRISPR target prediction systems, and performs well on customized AI accelerator cards.

RWKV combines linear attention with RNN characteristics, controlling the weight of historical information via a time decay factor to address the long-term dependency issues of traditional RNNs. It requires only a small number of state variables to maintain long sequence memory.

Both models have a computational complexity of $O(N)$. Mamba is primarily designed for long sequence processing, while RWKV is more suited for ultra-large-scale language models. In the future, the integration of Mamba and RWKV is an emerging trend. For example, RWKV-6 has improved its state update rules by incorporating the data-dependent characteristics of Mamba. This kind of integration is expected to lead to more efficient and powerful model architectures, offering better solutions for future multimodal tasks.

As derivatives of efficient models, the goal of Efficient Language Models is to endow language models with visual capabilities, enabling them to better handle tasks that integrate visual and textual information. In recent years, representative works in this field include VisualRWKV and VMamba(Liu et al., 2024), among others.These models achieve a deep integration of visual and textual information by incorporating visual encoders into the architecture of language models, significantly enhancing performance in tasks such as Visual Question Answering (VQA), image captioning, and document analysis. For example, VisualRWKV enhances the model's ability to understand complex visual information through high-resolution processing and multi-scale feature extraction. VMamba, on the other hand, leverages its dynamic decay mechanism and optimized training strategies to improve inference efficiency in visual-language tasks. The emergence of these models provides new ideas and approaches for the development of multimodal intelligent systems.

2.2 Architectural Limitations of Efficient Language Models

VisualRWKV and VMamba show potential in integrating visual and textual data, but they face challenges in processing long texts and complex dependencies. VisualRWKV, with its high-resolution processing and multi-scale feature extraction, enhances the understanding of visual information. However, it inherits the limitations of the RWKV architecture when dealing with long texts, such as struggling with long-range dependencies and context copying tasks. VMamba, on the other hand, leverages state space models to optimize long sequence processing but is less effective in handling discrete text data and tasks requiring historical recall. Therefore, improving feature quality and optimizing feature fusion mechanisms are key to addressing these issues and enhancing performance in long-text modeling and multimodal tasks.

3 Method

In the improved VisualRWKV, we focused on obtaining high-quality features and optimizing feature management. The key enhancements are as follows:

3.1 Improving feature quality

In the early versions of VisualRWKV, we used SigLIP(Zhai et al., 2023) and DINOv2(Oquab et al., 2023) encoders, focusing on low-resolution image processing. To enhance high-resolution image processing, we introduced a pre-trained SAM(Zou et al., 2023) vision encoder in the improved version, supporting 1024×1024 pixel resolution and significantly boosting the ability to capture key features. This improvement resulted in better performance across multiple benchmarks, making the model more efficient and accurate in complex tasks.

1. **Coarse-grained Feature** Coarse-grained features are obtained using the SigLIP encoder alone, which is designed for low-resolution

images. It can quickly extract global features of the image but lacks the ability to capture details. Although this method is efficient in processing low-resolution images, its neglect of detailed information leads to lower recall rates in tasks that require precise recognition of subtle features, making it prone to missing some key information.

2. **Medium-grained Feature** Medium-grained features are acquired by combining the SigLIP and DINOv2 encoders. SigLIP provides global features, while DINOv2 enhances the diversity and robustness of features through self-supervised learning, thereby compensating for some of SigLIP's shortcomings. This combined approach achieves a better balance between global and local features, capturing more details than coarse-grained features. As a result, recall rates are significantly improved, especially in tasks that require a moderate level of detail. However, when dealing with high-resolution images, the recall rate may still be limited due to the insufficient richness of feature details.

3. **Fine-grained Feature** Fine-grained features are obtained by integrating the SigLIP, DINOv2, and a pre-trained SAM vision encoder. The SAM encoder, with its powerful feature extraction capabilities, increases the model's supported resolution to 1024×1024 pixels and can capture global and local features with high precision. This high-quality feature acquisition method not only contains rich global information but also precisely captures local details and complex structures in the image, making it suitable for tasks that require rich details and visual complexity. Therefore, fine-grained features excel in recall rates, significantly reducing the occurrence of missed detections and improving the model's recall rate, especially in high-resolution image processing tasks.

3.2 Lossless DownSampler

To enable seamless alignment between high-resolution and low-resolution modules, we designed a lossless downsampler. This downsampler merges 2×2 blocks (each containing four adjacent vectors) into a new channel dimension, allowing high-resolution features to align effectively with low-resolution features without information loss during training. The process of the lossless downsampling can be represented by the following formula:

$$C_{new} = \text{Concat}(C_1, C_2, C_3, C_4) \tag{1}$$

Where:

- C_{new} represents the new channel dimension formed by concatenating the four blocks.
- C_1, C_2, C_3, C_4 are the 2x2 blocks, each containing four adjacent vectors.

This formula illustrates how the new channel dimension is created by combining the lower-resolution representations effectively.

3.3 Image Segmentation and Recombination Strategy

To obtain fine-grained features and enhance the model's ability to understand multi-scale visual information, we adopted an image segmentation and recombination strategy. The input image was divided into four parts, each processed by SigLIP, DINOv2, and SAM encoders for feature extraction. The features were then merged using an average pooling layer and aggregated with global features to create a multi-scale feature representation.This strategy balances coarse- and fine-grained information, significantly improving the model's ability to handle images up to 4096×4096 pixels. It also showed notable improvements in text-image question answering (TQA) tasks, enhancing cross-modal reasoning performance.

3.4 Feature Fusion Projection Layer

During training, we found that excessive feature information could hinder model stability and degrade feature quality. To address this, we redesigned the MLP with Context Gating to optimize feature selection and reduce loss.

Context Gating dynamically adjusts feature representations by applying a learnable gating function. Our improved MLP strengthens this mechanism, allowing finer control over input features. The gating layer uses a sigmoid-activated transformation to filter key information and suppress noise, enhancing feature fusion and representation. This design enables more effective handling of diverse features, improving overall model performance.

$$y = \sigma(W_g x + b_g) \odot x \tag{2}$$

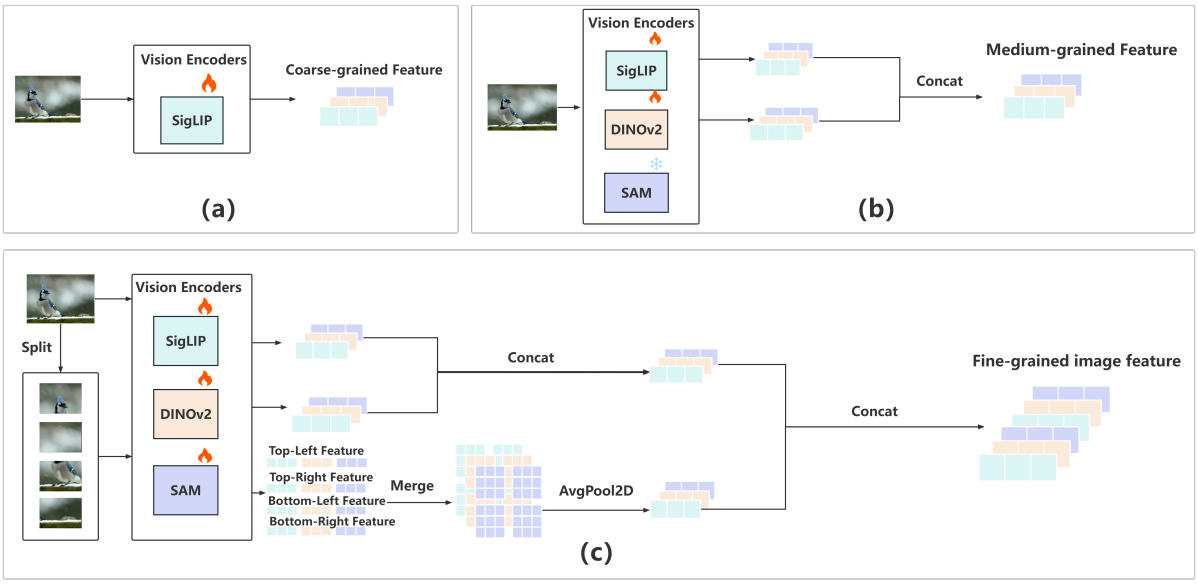

Figure 2: Three levels of feature extraction: coarse-grained (a), medium-grained (b), and fine-grained (c). Figure (a) uses the SigLIP encoder for low-resolution images with limited detail. Figure (b) combines SigLIP and DINOv2 for medium detail. Figure (c) integrates SigLIP, DINOv2, and a pre-trained SAM encoder for high-resolution images, capturing both global and local details. This progression highlights the increasing detail and applicability across different scenarios.

In this formula, y represents the output, W_g is the gating weight matrix, b_g is the bias term, and x is thenput data. The activation function used is the sigmoid function σ, and the symbol $\odot$ indicates element-wise multiplication. This mechanism optimizes the model's performance by dynamically adjusting the input features.

Compared to traditional MLP designs, our improved approach achieves key optimizations in the following aspects:

1. **Lower Loss**: The newly designed MLP structure optimizes gradient propagation, stabilizing the model and reducing information loss during training.

2. **Higher Feature Quality**: By implementing a refined feature selection mechanism, the model enhances effective feature representation, especially for high-resolution visual tasks.

3. **Improved Training Stability**: Prevents adversarial effects caused by excessive feature information, ensuring more robust model performance across various tasks.

This enhancement not only stabilizes the training process but also significantly improves the model's feature extraction capability, making it more effective in high-resolution vision tasks.

4 Experiments

We refer to this enhanced version of VisualRWKV as VisualRWKV-Boost.In the Experiment section, we evaluate the performance of VisualRWKV-Boost model across a variety of tasks, focusing on its ability to handle high-resolution visual inputs effectively. We conducted experiments on several widely-used visual language model (VLM) benchmarks, with a particular emphasis on text-rich and document analysis tasks that benefit from high-resolution image processing.

4.1 Baselines

In the Baseline Comparisons section, we compare VisualRWKV-Boost not only with the standard VisualRWKV model(Hou et al., 2024) but also with several other relevant models to comprehensively quantify the improvements brought by high-resolution processing. The standard VisualRWKV model serves as the baseline without high-resolution optimization. Through these comparisons, we highlight the significant advantages of high-resolution processing in text-dense tasks and document analysis, which require precise detail recognition. Additionally, the results demonstrate the critical role of high-resolution enhancement in handling complex visual tasks.

Model	Vision Encoder	Resolution	SQA	TextVQA	GQA	VizWiz	MME	POPE	MMB/MMB-CN
VisualRWKV 1.6B	CLIP	336	59.05	43.57	55.23	29.84	1204.90/245.00	0.832	55.75/53.17
+ SigLIP and DINOv2	SigLIP + DINOv2	384	53.35	41.08	56.55	31.44	1273.67/213.92	0.870	57.39/51.72
+ SAM-b-1024	SigLIP + DINOv2 + SAM-b-1024	384	57.02	48.70	58.23	30.46	1250.50/213.21	0.818	58.84/57.13
+ Scale up resolution	SigLIP + DINOv2 + SAM-b-1024	448	58.55	47.75	**60.96**	33.12	1305.38/224.64	0.855	59.45/53.09
+ MLP with Context Gating	SigLIP + DINOv2 + SAM-b-1024	448	54.39	54.71	60.84	**54.97**	**1378.62/266.07**	**0.860**	**60.31/55.41**
+ HD559k dataset	SigLIP + DINOv2 + SAM-b-1024	448	**58.75**	55.62	60.18	51.59	1271.03/230.36	0.857	57.56/51.03
+ HD667k dataset	SigLIP + DINOv2 + SAM-b-1024	448	56.97	**56.31**	59.52	49.88	1321.33/232.14	0.853	58.42/52.84

Table 1: Performance metrics of different VisualRWKV models on academic tasks. Bolded data in the table represents the best performance.

Model	Dataset	DocVQA	InfographicVQA	ChartQA
VisualRWKV 1.6B	mix665k	10.88	-	10.00
VisualRWKV 1.6B + MLP	mix665k	11.00	11.00	8.00
VisualRWKV-Boost 1.6B	HD559k	35.11	16.49	39.32
VisualRWKV-Boost 1.6B	HD667k	35.37	16.82	40.28

Table 2: Performance metrics of VisualRWKV-Boost model on text-rich tasks.

4.2 Benchmarks

We evaluated VisualRWKV-Boost using eight benchmark datasets: SQA(Lu et al., 2022b), TextVQA(Singh et al., 2019), GQA(Hudson and Manning, 2019), VizWiz(Bigham et al., 2010), MME(Fu et al., 2023), POPE(Li et al., 2023), MMB(Liu et al., 2023b), and MMB-CN. These datasets cover various tasks, such as scenario-based question answering (SQA), text extraction from images (TextVQA), and reasoning over image content (GQA). We also included document benchmarks like DocVQA(Mathew et al., 2021), InfographicVQA(Mathew et al., 2022), and ChartQA(Masry et al., 2022) to assess document and chart comprehension. Results show that VisualRWKV-Boost significantly outperforms lower-resolution models in these tasks, demonstrating its effectiveness in handling high-resolution visual data across different modalities and languages.

4.3 Quantitative Evaluation

In the quantitative evaluation, Table 3 highlights the significant advantages of VisualRWKV-Boost across multiple academic tasks. Specifically, VisualRWKV-Boost achieves an accuracy of 56.97% in the SQA task, surpassing MobileVLM 1.7B (54.7%)(Chu et al., 2023) and VisualRWKV (59.1%). In the TextQA task, its accuracy is 56.31%, a significant improvement over VisualRWKV's 43.6%. For the GQA task, VisualRWKV-Boost reaches an accuracy of 60.84%, outperforming other models. Additionally, it achieves a competitive score of 54.97% in the VizWiz task. Overall, the increased resolution enables VisualRWKV-Boost to better process high-resolution image details, leading to higher accuracy and superior performance across various tasks.

4.4 Ablation Study

In the Experiment section, we evaluate VisualRWKV-Boost on several VLM benchmarks, focusing on text-rich and document analysis tasks that demand high-resolution visual understanding. The results highlight the model's effectiveness in real-world scenarios where detail and clarity are critical.

4.4.1 Ablation on Vision Encoder

In this section, we compared visual encoders, specifically SigLIP and SigLIP + DINOv2, based on a resolution of 384, as shown in Table 4. The results showed a comprehensive performance improvement. We further enhanced the model by integrating SAM into the SigLIP + DINOv2 framework, leading to additional performance gains on the SQA, TQA, and MMB/MMB_{CN} datasets.We assessed the impact of using DINOv2 and SAM on training stability and computational efficiency. The key metrics evaluated included training stability and overall computational cost, as well as the performance results across various datasets. After introducing DINOv2 and SAM, the model exhibited enhanced stability during training and improved performance across all datasets. This highlights the significant role that the SAM and DINOv2 visual encoders play in effectively processing high-resolution inputs.

4.4.2 Ablation on Resolution

In this experiment, we introduced SigLIP, DINOv2, and SAM and increased the resolution from 384

Method	LLM	Resolution	SQA	TextQA	GQA	VizWiz	MME	POPE	MMB/MMB-CN
MobileVLM 1.7B	MobileLLaMA-1.4B	336	54.7	-	56.1	-	1196.2/-	84.5	53.2/-
Mini-Gemini	Gemma-2B	336	-	-	-	-	1341.0/312.0	-	59.8/-
TinyLLaVa-v1	TinyLlama-1.1B	-	59.4	-	57.5	-	-	-	-
LLaVa-1.5	Vicuna 7B	336	66.8	58.2	62.0	50.0	-	-	-
FastViT	Vicuna 7B	256	-	51.6	60.2	-	-	82.9	-
FastViTHD	Vicuna 7B	-	-	53.1	60.6	-	-	82.3	-
VisualRWKV	VisualRWKV6-1.6B	336	59.1	43.6	55.2	-	1204.9/-	83.2	55.8/53.2
VisualRWKV	VisualRWKV6-7B	336	68.2	51.0	64.3	-	1387.8/-	84.7	65.8/63.7
VisualRWKV-Boost	VisualRWKV6-1.6B	448	57.0	56.3	60.8	55.0	1378.6/266.1	86.0	60.3/55.4

Table 3: Performance comparison of different visual language models across various academic tasks.

Model	Vision Encoder	Resolution	SQA	TextVQA	GQA	VizWiz	MME	POPE	MMB/MMB-CN
VisualRWKV 1.6B	CLIP	336	59.05	43.57	55.23	29.84	1204.90/245.00	0.832	55.75/53.17
VisualRWKV 1.6B	SigLIP + DINOv2	384	53.35	41.08	56.55	31.44	1273.67/213.92	0.870	57.39/51.72
VisualRWKV 1.6B	SigLIP + DINOv2 + SAM-b-1024	384	57.02	48.70	58.23	30.46	1250.50/213.21	0.818	58.84/57.13

Table 4: Ablation study on Vision Encoder

to 448, as shown in Table 5. This adjustment improved the model's performance on datasets such as SQA, GQA, and VizWiz. We compared the performance of VisualRWKV-Boost under different resolution settings to explore the impact of increased resolution on accuracy and processing time.

Feature Quality: Increasing the input resolution to 448 pixels significantly improved feature quality, enabling the model to capture finer details—especially important in text-heavy tasks like TextVQA. The enhanced features improved the model's ability to recognize and extract textual information, leading to higher answer accuracy. While inference time increased, techniques like segmentation and downsampling effectively balanced efficiency and accuracy.

4.4.3 Ablation on Projection

In the experiment, the introduction of MLPWithContextGating significantly enhanced the recall performance in the TextVQA task, increasing it from 47.75% to 54.71%. This improvement is attributed to the mechanism's ability to dynamically filter key features and suppress noise, thereby optimizing feature quality. Additionally, it enhanced training stability and reduced memory consumption, particularly when processing high-resolution inputs. This highlights the crucial role of MLPWithContextGating in improving performance for complex visual-textual tasks.

4.4.4 Ablation on Data Scaling up

In the ablation study on data expansion, we analyzed VisualRWKV-Boost's feature utilization across different datasets (mix665k, HD559k, HD667k). The results showed that as dataset size increased, feature utilization improved, enhancing performance in tasks like SQA, TextVQA, and MME. For instance, in the SQA task, accuracy rose from 54.39% with mix665k to 58.75% with HD559k, indicating that larger datasets with high-resolution images improve learning and reasoning. However, further increasing the dataset size to HD667k led to a slight performance decline in some tasks, suggesting a need to balance dataset size and computational efficiency. The improvement in feature quality also boosted recall, allowing better extraction of key information from high-resolution images.

5 Conclusion

In the conclusion, we propose an enhanced VisualRWKV that integrates lossless downsampling and high-/low-resolution visual encoders to improve feature quality while maintaining computational efficiency. Advanced encoders like SigLIP, DINOv2, and SAM are incorporated to balance coarse- and fine-grained information representation. Additionally, an image segmentation and reassembly mechanism is introduced to strengthen multi-scale feature representation, supporting input resolutions up to 4096×4096 pixels. These improvements significantly enhance the model's visual understanding and recall in high-resolution tasks, offering a new direction for developing efficient and scalable VLMs.

Limitations Despite significant improvements in feature quality and high-resolution processing capabilities, the proposed method still has several limitations and potential risks. First, while integrating multiple visual encoders enhances feature extraction, it may also introduce additional computational overhead, potentially affecting inference

Model	Vision Encoder	Resolution	SQA	TextVQA	GQA	VizWiz	MME	POPE	MMB/MMB-CN
VisualRWKV-Boost	SigLIP + DINOv2 + SAM-b-1024	384	57.02	48.70	58.23	30.46	1250.50/213.21	0.818	58.84/57.13
VisualRWKV-Boost	SigLIP + DINOv2 + SAM-b-1024	448	58.55	47.75	60.96	33.12	1305.38/224.64	0.855	59.45/53.09

Table 5: Ablation study of VisualRWKV-Boost on different resolutions.

Model	Vision Encoder	Resolution	SQA	TextVQA	GQA	VizWiz	MME	POPE	MMB/MMB-CN
VisualRWKV + Linear Projection	SigLIP + DINOv2 + SAM-b-1024	448	58.55	47.75	60.96	33.12	1305.38/224.64	0.855	59.45/53.09
VisualRWKV + MLP	SigLIP + DINOv2 + SAM-b-1024	448	54.39	54.71	60.84	54.97	1378.62/266.07	0.860	60.31/55.41

Table 6: Ablation study on projection layer.

Model	Dataset	Vision Encoder	Resolution	SQA	TextVQA	GQA	VizWiz	MME	POPE	MMB/MMB-CN
VisualRWKV-Boost	mix665k	SigLIP + DINOv2 + SAM-b-1024	448	54.39	54.71	60.84	54.97	1378.62/266.07	0.860	60.31/55.41
VisualRWKV-Boost	HD559k	SigLIP + DINOv2 + SAM-b-1024	448	58.75	55.62	60.18	51.59	1271.03/230.36	0.857	57.56/51.03
VisualRWKV-Boost	HD667k	SigLIP + DINOv2 + SAM-b-1024	448	56.97	56.31	59.52	49.88	1321.33/232.14	0.853	58.42/52.84

Table 7: Ablation study of VisualRWKV-Boost on different datasets.

efficiency, especially in resource-constrained environments. Second, although the model supports input resolutions up to 4096×4096 pixels, processing high-resolution images is computationally intensive, leading to increased training time and resource consumption. Additionally, the method relies on specific visual encoders (such as SigLIP, DINOv2, and SAM), which may limit its adaptability to different visual tasks or datasets. Applying the model to new tasks may require additional fine-tuning or retraining. Moreover, while the approach demonstrates strong recall performance in text-rich scenes and document analysis, its generalization capability across more diverse visual-language tasks remains to be further validated. Furthermore, the introduction of complex multi-scale representations and advanced encoders may increase model complexity, potentially bringing risks such as overfitting to specific datasets or performing poorly in unseen scenarios. Finally, for real-time applications, optimizing computational efficiency while maintaining high-resolution visual understanding is still a key challenge. Future work will focus on addressing these issues, exploring more efficient and flexible architectural designs to enhance the model's adaptability and practicality.

References

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

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. 2023. Zoology: Measuring and improving recall in efficient language models. *arXiv preprint arXiv:2312.04927*.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. 2023. Mobilevlm: A fast, strong and open vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Haowen Hou, Peigen Zeng, Fei Ma, and Fei Richard Yu. 2024. Visualrwkv: Exploring recurrent neural networks for visual language models. *arXiv preprint arXiv:2406.13362*.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023b. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. 2024. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and A. Kalyan. 2022b. Learn to explain: Multimodal reasoning via thought chains for science question answering. *ArXiv*, abs/2209.09513.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Jiaju Lin, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Bolun Wang, Johan S. Wind, Stanislaw Wozniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. Rwkv: Reinventing rnns for the transformer era. *Preprint*, arXiv:2305.13048.

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, et al. 2024. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*, 3.

Bo Peng, Ruichong Zhang, Daniel Goldstein, Eric Alcaide, Xingjian Du, Haowen Hou, Jiaju Lin, Jiaxing Liu, Janna Lu, William Merrill, et al. 2025. Rwkv-7" goose" with expressive dynamic state evolution. *arXiv preprint arXiv:2503.14456*.

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.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. 2023. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*.

A Model Architecture and Computing

Model Architecture: The VisualRWKV models used in our experiments are visual extensions of the Recurrent Weighted Key-Value (RWKV) architecture, designed to handle both visual and textual data. We experimented with the following configurations:

- **VisualRWKV 1.6B:** A baseline model using 1.6 billion parameters.
- **VisualRWKV 1.6B + MLP:** Enhanced with a Multi-Layer Perceptron (MLP) to improve feature extraction.
- **VisualRWKV 1.6B + MLP (VisualRWKV-Boost):** A model that adopts the VisualRWKV-Boost strategy to extract more fine-grained features.

Computing Infrastructure : Infrastructure A range of computational resources were employed in the study. The standard training and benchmark evaluation were conducted using 8 NVIDIA A100-80GB GPUs. The VisualRWKV 7B model is trained with 6 A100 GPUs due to insufficient memory capacity with 8 GPUs. For the efficiency analysis, we employed an NVIDIA RTX 3090 GPU.

Computing Budget: Training an epoch of VisualRWKV 1.6B with 8 A100 GPUs takes 6.7 hours, equivalent to 53.6 GPU hours; Training an epoch of VisualRWKV 3B with 8 A100 GPUs takes 11.3 hours, equivalent to 90.4 GPU hours; Training an epoch of VisualRWKV 7B with 6 A100 GPUs takes 26.5 hours, equivalent to 159 GPU hours

In all cases, the RWKV backbone was adapted for visual tasks by incorporating Vision Encoders and using Context Gating. These models were fine-tuned for visual question-answering tasks on various datasets.

B Datasets

We trained and evaluated the models on the following datasets:

- **mix665k:** This is the dataset used by LLaVA for instruction tuning, comprising 665,000 diverse images aimed at enhancing the model's adaptability to various visual tasks and instructions, thereby improving its overall performance and usability.
- **HD559k:** This dataset is our custom high-resolution dataset consisting of 559,000 high-quality images. It focuses on testing the model's performance when processing high-quality visual content, particularly in terms of detail, color, and clarity, ensuring that the model can accurately capture complex visual information. Table 8 and Figure 3 provide an overview of the data proportions in the HD559k dataset.
- **HD667k:** As another significant contribution from our team, HD667k is a larger high-resolution dataset containing 667,000 images. This dataset not only enriches the training data for the model but also provides additional support for its performance in diverse and complex visual scenarios, helping to improve the model's generalization ability and robustness in practical applications. Table 8 and Figure 3 provide an overview of the data proportions in the HD667k dataset.

C Experimental Setup

Preprocessing: Input images were divided into four sections, each encoded by three vision encoders (SigLIP, DINOv2, SAM). Features were merged and passed through an MLP with Context Gating.

Training: The models were trained using the AdamW optimizer with a learning rate of X, using NVIDIA GPUs and mixed precision. Training continued for 100 epochs, with early stopping applied after 10 epochs of no improvement.

Evaluation: Models were evaluated on the DocVQA, InfographicVQA, and ChartQA datasets. These datasets represent different challenges, from document understanding to infographics and chart analysis.

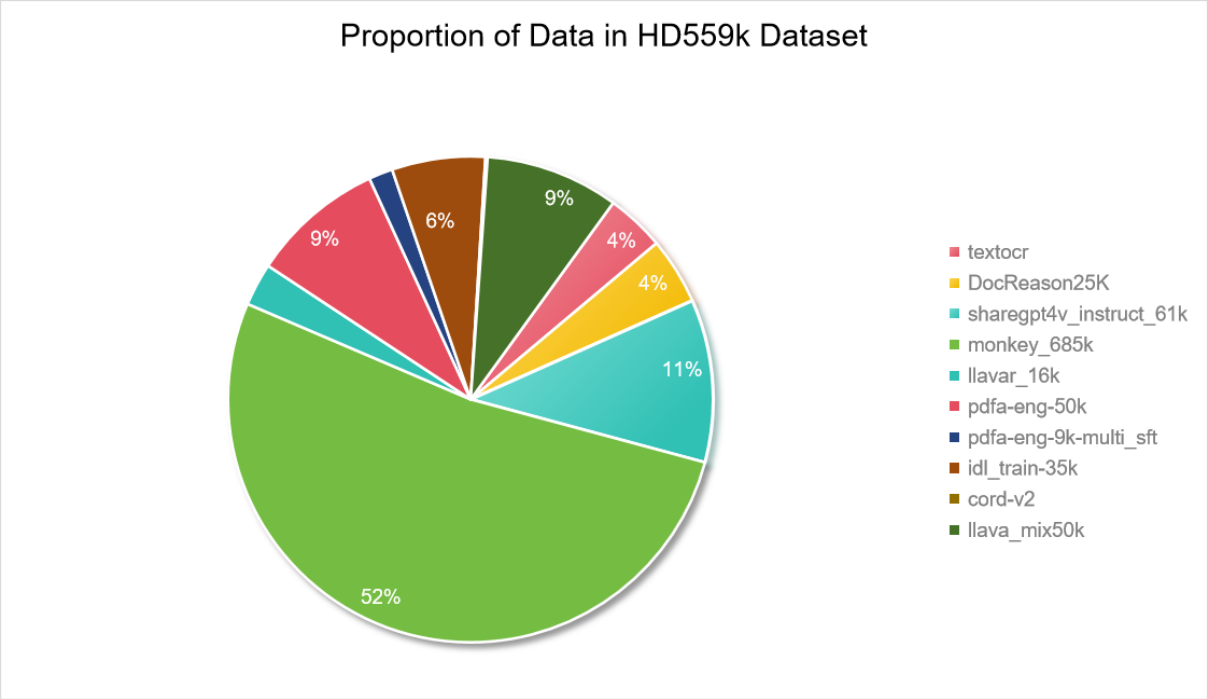

Figure 3: Distribution of the HD559k dataset, showcasing the various datasets and their respective quantities. This comprehensive dataset includes a diverse range of sources, contributing to a total of 559,494 images utilized for training and evaluation purposes.

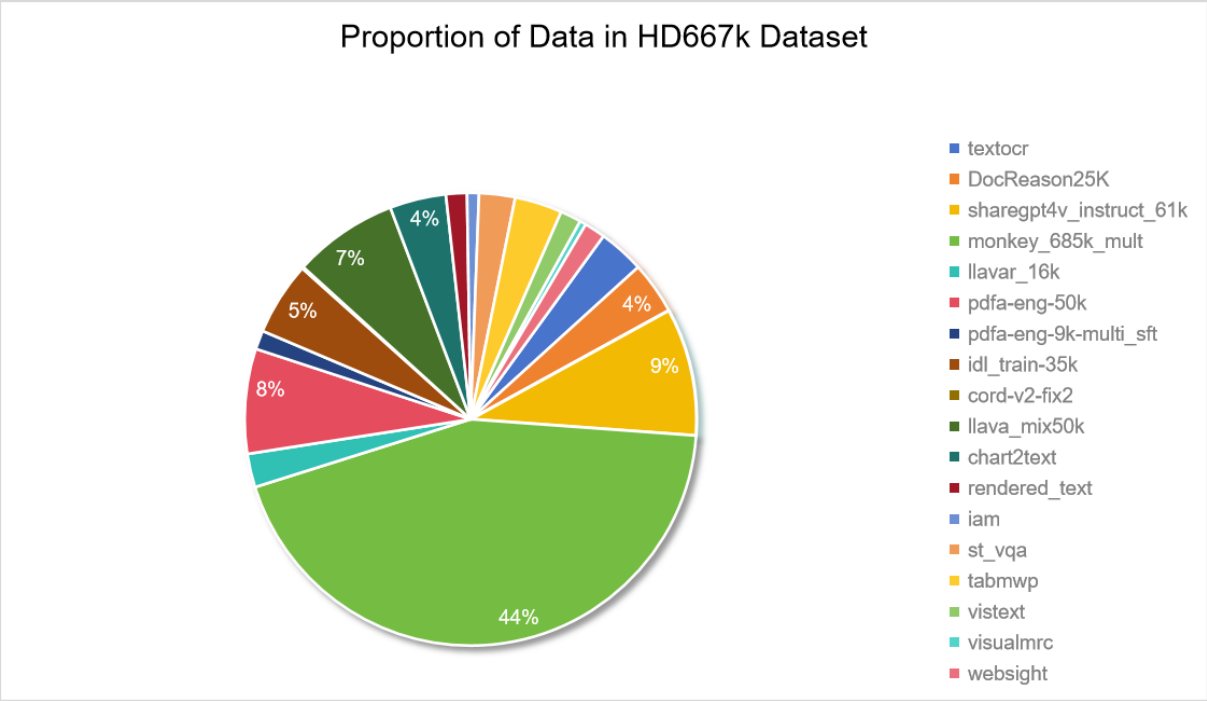

Figure 4: Distribution of the HD667k dataset, illustrating the composition and quantity of various datasets included. With a total of 667,000 images, this dataset encompasses a wide array of visual tasks and sources, aimed at enhancing the training and evaluation of model performance.

Dataset Name	Quantity
textocr	21.9k
DocReason25K	25k
sharegpt4v_instruct_61k	61k
monkey_685k_multi_round	294k
llavar_16k	16k
pdfa-eng-50k	50k
pdfa-eng-9k-multi_sft	9k
idl_train-35k	35k
cord-v2-fix2	0.8k
llava_mix50k	50k

Table 8: Overview of Datasets Used of HD559k

Dataset Name	Quantity
textocr	21.9k
DocReason25K	25k
sharegpt4v_instruct_61k	61k
monkey_685k_multi_round	294k
llavar_16k	16k
pdfa-eng-50k	50k
pdfa-eng-9k-multi_sft	9k
idl_train-35k	35k
cord-v2-fix2	0.8k
llava_mix50k	50k
chart2text	26.9k
rendered_text	10k
iam	5.66k
st_vqa	17.2k
tabmwp	22.7k
vistext	9.97k
visualmrc	3k
websight	10k
infographic_vqa	2.1k

Table 9: Overview of Datasets Used of HD667k

D Data and Hyperparameters

- **A. Training Data**

  We used a two-phase training process for VisualRWKV. In the *Feature Alignment Phase*, 558K images from LAION-CC-SBU were utilized to connect a frozen vision encoder with a frozen LLM. This phase establishes the foundation for robust image-text alignment. In the *Visual Instruction Tuning Phase*, an expanded dataset of 150K multimodal examples generated by GPT and 515K VQA datasets were used to enhance the model's capacity for multimodal tasks.All the data used in this paper are consistent with their intended use.

  Ethical guidelines were strictly followed in data preparation, focusing on identifying and handling PII and sensitive content via automated tools and manual reviews. Anonymization techniques, such as data masking, were applied to ensure data integrity and privacy.

- **B. Evaluation Benchmarks**

  We employed various benchmarks to evaluate the model. VQA-v2 and GQA metrics are based on

the test-dev split, while TextVQA is evaluated on its validation set. ScienceQA and POPE metrics are from their respective test sets. MMBench metrics are based on the development set, and MME is evaluated on a specific test set.

- **C. Data Language**

  Our training data spans multiple datasets, with most Visual Question Answering (VQA) datasets being in English. The ShareGPT data is multilingual, covering multiple user-contributed languages. Among the evaluation benchmarks, MMBench-cn is in Chinese, while the rest are in English.

- **D. Hyperparameters**

  The models used 1.6B parameters for experiments. Detailed hyperparameters for both the vision-language alignment pretraining and the visual instruction tuning phases are listed in Table 10. These include settings optimized for diverse tasks across different datasets, ensuring robust model performance.

Hyperparameter	1.6B-Pretrain	1.6B-Finetune
batch size	256	128
lr init	1e-3	6e-5
lr end	1e-5	1.5e-5
lr schedule	cosine decay	cosine decay
lr warmup ratio	0	0
weight decay	0	0
epoch	2	2
optimizer	AdamW	AdamW
DeepSpeed stage	1	1

Table 10: Hyperparameters for 1.6B model pretraining and finetuning.

E Limitations and Future Work

Although this strategy significantly improves model performance, especially on ChartQA and TextQA, challenges remain in document understanding tasks. Future work will explore improved feature extraction methods and further optimize the model for multimodal tasks.

F Use of AI Assistants

In this research, an AI writing assistant is solely employed for the purposes of paraphrasing, spell-checking, and enhancing the author's original content, and it does not introduce any novel content.