MDocRAG-RL: Empowering Multi-Modal Document RAG via Complex Visual Reasoning with Reinforcement Learning

Zhongyu Wang[†]
Beihang University
wangzhongyu@buaa.edu.cn

Abstract

While Retrieval-Augmented Generation(RAG) enhances multi-modal large language models(MLLMs) by introducing external knowledge, existing RAG systems still face significant limitations when dealing with complex visual reasoning. On one hand, MLLMs, being generative models, produce suboptimal embeddings for retrieval tasks. On the other hand, existing methods naively insert images into context without adequate visual perception, thereby limiting reasoning capabilities. To address these challenges, we propose MDocRAG-RL, a novel RAG framework for complex visual reasoning. We design specialized pre-training and fine-tuning tasks to enable MLLMs to compress visual document representations and align textual and visual embeddings for improved retrieval efficiency. Additionally, we design a visual perception action space for the generator that allows progressive coarse-to-fine information acquisition from visually-rich documents. Furthermore, we develop a reinforcement learning framework to enhance the complex visual reasoning capability of the RAG system. Extensive experiments on multiple challenging benchmarks demonstrate the significant effectiveness of our approach, achieving state-of-the-art performance across various benchmarks.

1 Introduction

Retrieval-Augmented Generation (RAG) enables large language models (LLMs) to leverage external knowledge bases to mitigate hallucinations and address complex problems (Chen et al., 2025a; Hu et al., 2025). However, traditional text-based RAG approaches fail to effectively capture visual information, limiting their applicability to real-world multi-modal documents (Li et al., 2025a,b). These visually-rich documents typically contain diverse elements including text, tables, and charts, presented in varied spatial layouts. To handle such documents, recent works have introduced multimodal large language models (MLLMs) into RAG systems, thereby extending RAG to the visual domain (Wasserman et al., 2025; Dong et al., 2025).

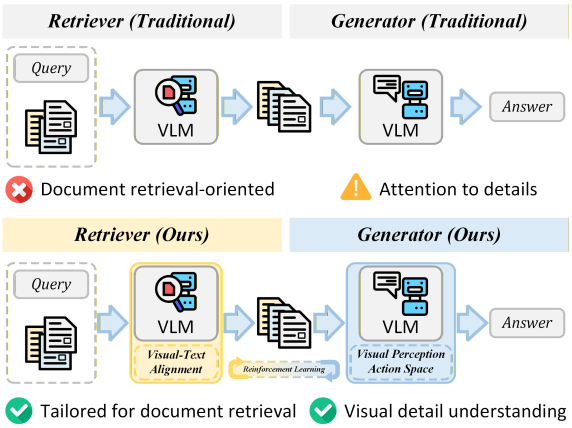

Figure 1: Comparison between our MDocRAG-RL and existing RAG systems.

Despite recent progress, existing visual RAG systems exhibit substantial limitations in retrieval efficiency and complex visual reasoning, manifesting in three key aspects: (1) Suboptimal retrieval efficiency. Current RAG systems employ MLLMs to directly encode visually-rich documents. However, as generative models optimized for next-token prediction, MLLMs produce suboptimal embeddings for retrieval tasks, failing to accurately represent query semantics and visual information, thereby resulting in poor retrieval performance. (2) Insufficient visual perception. Visual RAG systems inadequately consider vision-specific perceptual processes during generation, naively inserting images into the context. This prevents the generator from achieving fine-grained image understanding, leading to the loss of critical visual information. (3) Limited complex reasoning capability. The improper handling of visual information results in

[†]Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 8641–8651
July 2-7, 2026

inadequate reasoning token allocation, preventing models from fully exploiting key visual features in retrieved images, thus constraining their performance on complex visual reasoning tasks.

To address these limitations, we propose MDocRAG-RL, a novel multi-modal RAG framework. We design innovative supervised pre-training and fine-tuning tasks for the MLLM-based retriever, aligning textual and visual information in documents through retrieval and generation tasks. This approach compresses entire images into dense token representations while maintaining alignment with the textual content in the images. Consequently, the retriever can effectively encode both queries and documents, better adapting to retrieval scenarios. Furthermore, we define a tailored action space for the MLLM-based generator to process visually-rich inputs. Through select, crop, and zoom operations, the generator progressively acquires visual information in a coarse-to-fine manner, enhancing its perception of vision-intensive regions. To further strengthen the interactive reasoning capability, we develop a dedicated reinforcement learning framework and employ the GRPO algorithm to train the multi-modal RAG agents, enabling the generator to learn effective interaction strategies with the retriever for acquiring visually-rich information.

Our contributions are summarized as follows:

- We design document retrieval-oriented pre-training tasks for the retriever to compress visual document representations and align textual and visual information.
- We introduce a visual perception action space for the generator, enabling coarse-to-fine understanding of visually-rich documents.
- We develop a reinforcement learning framework to train the RAG agent, enhancing the complex visual reasoning capability.
- We conduct comprehensive experiments on challenging multi-modal benchmarks to validate the effectiveness of our approach, achieving new state-of-the-art performance across multiple benchmarks.

2 Related Work

2.1 MLLMs in Visual Document RAG

MLLMs have experienced rapid development in recent years by integrating visual understanding capabilities through combining vision encoders with language models (Sun et al., 2025; Han et al., 2025; Tanaka et al., 2024). Building upon MLLMs, visual document retrieval and visual RAG systems encode visually-rich documents directly as images for processing (Luo et al., 2025; Lu et al., 2025). While these methods have achieved success in specific scenarios, they still encounter critical challenges when handling diverse real-world documents (Cao et al., 2025; Chen et al., 2025b). A core issue is that MLLMs, as generative models, are trained using the next-token prediction objective, which is suboptimal for embeddings required by retrievers. However, prior work has not designed specialized training strategies to address this gap, directly applying pre-trained MLLMs to retrieval scenarios (Liu et al., 2025; Ravenda et al., 2025). To bridge this gap, our MDocRAG-RL framework designs novel pre-training tasks for the MLLM-based retriever to compress visual documents into dense token representations and align textual and visual embeddings, thereby improving retrieval efficiency.

2.2 Reinforcement Learning for MLLMs

Reinforcement learning has been proven as a foundational approach to enhance the reasoning capabilities of large language models, which is crucial for effectively solving complex problems (Liang et al., 2025; Wang et al., 2025b; Zhang et al., 2025a; Wan et al., 2025). Prior research has applied RL to train LLMs (Hou et al., 2025; Lan et al., 2025; Yue et al., 2025; Ma et al., 2025). Increasingly, studies have attempted to apply RL to improve the visual reasoning capabilities of MLLMs (Yang et al., 2025; Wang et al., 2025a; Shen et al., 2025). Building upon this foundation, recent studies have further applied RL to train MLLM-driven visual RAG systems to address the unique challenges in multi-modal understanding and generation tasks (Zhang et al., 2025b; Jiang et al., 2025; Li et al., 2025c). However, existing visual RAG systems do not fully leverage visual perception of retrieved visual information during generation, simply inserting images into the context (Wang et al., 2025d; Faysse et al., 2024; Yu et al., 2024). This practice leads to insufficient token allocation for visual reasoning, limiting the complex reasoning capability of MLLM-based generators. Our approach designs a visual perception action space for the generator to enable coarse-to-fine understanding of retrieved images, and develops a reinforcement learning framework that enhances the model's capability to utilize visual

perception actions through RAG-specific rewards, thereby improving the complex visual reasoning capability of the RAG system.

3 Method

3.1 Architecture Overview

As illustrated in Figure 2, MDocRAG-RL consists of two main components: a retriever and a generator. The Retriever adopts a dual-encoder architecture based on MLLMs to separately encode queries and document images. The Generator is an MLLM equipped with a vision perception action space, enabling progressive information acquisition from coarse to fine granularity.

3.2 Pre-Training for Retriever

Self-Supervised Pre-training. To enable effective visual document retrieval, we train the retrieval module to encode document images into compact representations. Our training strategy consists of two self-supervised objectives that leverage unlabeled document collections, as illustrated in Figure 3(a). The core idea is to aggregate visual information from the entire document image into a single embedding vector, which is extracted from the final token position of the vision encoder output.

We design two complementary pre-training tasks. The first task employs a contrastive learning framework where we construct training pairs from document images and their automatically extracted text content. For each document image I_i, we obtain its textual counterpart T_i through optical character recognition (OCR). The query encoder processes T_i to produce embedding $\mathbf{e}_q$, while the document encoder processes I_i to yield embedding $\mathbf{e}_d$. Using in-batch negative sampling with batch size B, we optimize the following contrastive objective:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{SIM}(\mathbf{e}_q, \mathbf{e}_d^+)/\tau)}{\sum_{j=1}^{B} \exp(\text{SIM}(\mathbf{e}_q, \mathbf{e}_d^j)/\tau)} \quad (1)$$

where $\text{SIM}(\cdot,\cdot)$ denotes cosine similarity, τ is a temperature parameter, and $\mathbf{e}_d^+$ represents the positive document embedding.

The second task adopts a generative approach to compress visual features into the final embedding. We use a specialized attention mechanism during training: while image tokens attend to all previous tokens as usual, text tokens can only attend to the aggregated embedding and preceding text tokens, thereby forcing the model to consolidate visual information. Given the extracted text sequence $\{t_1, \ldots, t_L\}$ of length L, we minimize:

$$\mathcal{L}_{\text{generate}} = -\frac{1}{L} \sum_{i=1}^{L} \log p(t_i \mid t_{<i}, < \text{EOS} >) \quad (2)$$

where <EOS> denotes the aggregated visual embedding. The complete pre-training loss combines both objectives: $\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{contrast}} + \mathcal{L}_{\text{generate}}$.

Supervised Fine-tuning. After pre-training, we fine-tune the retrieval module on annotated query-document pairs $\{(Q_i, I_i^+)\}$ using contrastive learning, as illustrated in Figure 3(b). For each query Q_i, we treat the paired document I_i^+ as a positive example and other in-batch documents as negatives. The fine-tuning objective follows Eq. (1) but replaces OCR text with actual queries. Once fine-tuning completes, the retriever can identify the top-k relevant documents for any input query, which are subsequently passed to the generator for answer synthesis.

3.3 Visual Perception Actions for Generator

Visual Perception Action Space. To enable the generator to dynamically interact with visual content, we introduce a structured action space tailored for document understanding tasks. The generator operates through iterative interactions with the environment following a cyclical pattern of reasoning, action execution, and observation acquisition. At each step t, the policy π_θ produces an action $A_t \sim \pi_\theta(\cdot \mid H_{t-1})$ conditioned on the historical trajectory $H_{t-1} = \{T_1, A_1, O_1, \ldots, T_{t-1}, A_{t-1}, O_{t-1}\}$, where T_i, A_i, and O_i denote the thought, action, and observation at step i, respectively.

Our action space encompasses three primary categories. First, the *search* action retrieves relevant document images from the corpus based on the current query context. Second, the *summarization* action synthesizes accumulated information to generate intermediate or final answers. Third, the *visual perception* action enables fine-grained analysis of specific regions within previously retrieved documents. This perception action is expressed through special tokens <region> and </region> that encapsulate spatial coordinates defining a bounding box $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ over the image. These coordinates specify the target region R.

The perception mechanism operates as follows. Given that a document image I_k has been retrieved in an earlier step $k < t$, the perception action at step

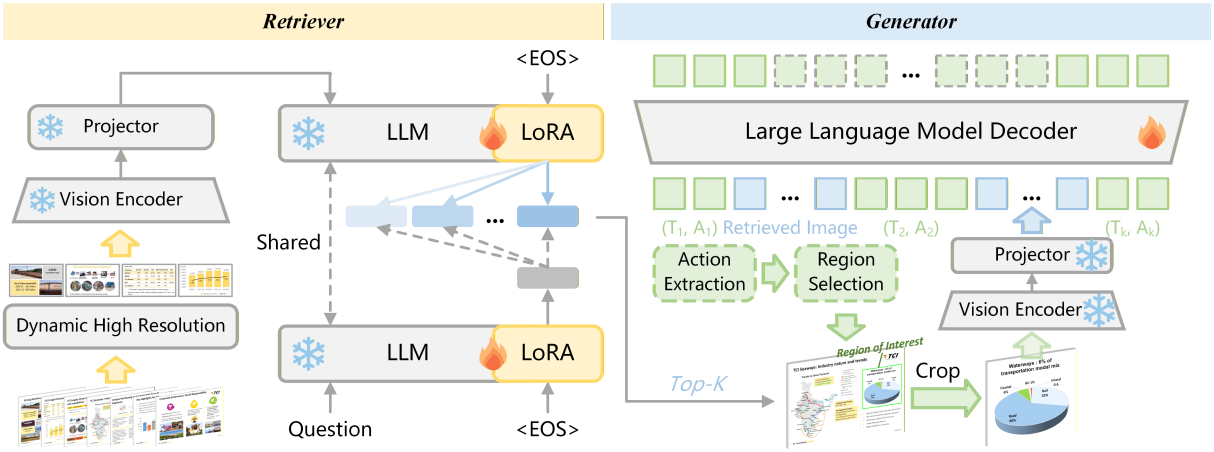

Figure 2: Overview of the MDocRAG-RL framework.

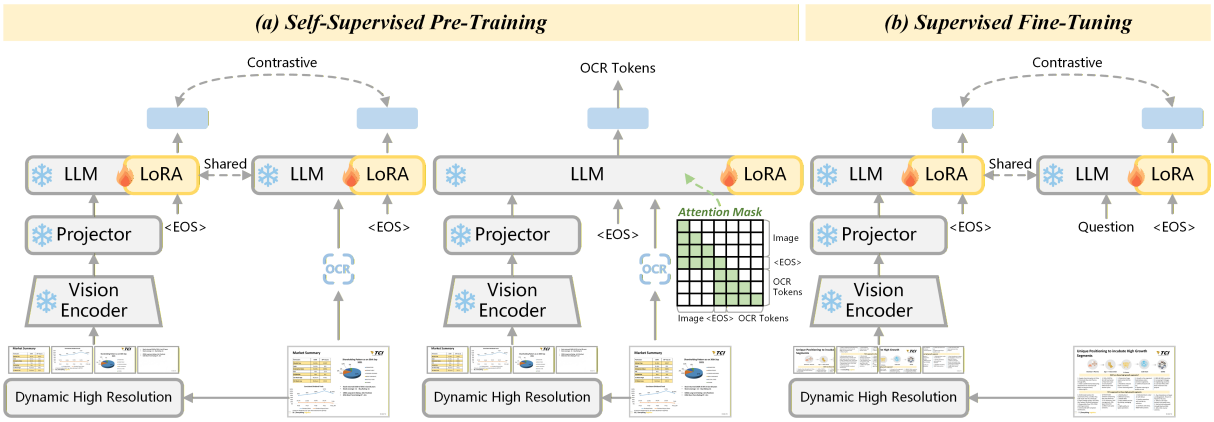

Figure 3: Self-supervised pre-training and supervised Fine-tuning tasks for the retriever.

t extracts the specified region R from I_k. Formally, we have $A_t \times O_k \rightarrow O_t$, where O_k represents the encoded features of image I_k with resolution $w \times h$. The coordinates are mapped back to the original high-resolution image I_{raw} with dimensions $w_{\text{raw}} \times h_{\text{raw}}$ through the transformation:

$$\hat{R} = \text{Crop}\left(I_{\text{raw}}, \left[\frac{x_{\min} \cdot w_{\text{raw}}}{w_{\text{enc}}}, \frac{y_{\min} \cdot h_{\text{raw}}}{h_{\text{enc}}}, \frac{x_{\max} \cdot w_{\text{raw}}}{w_{\text{enc}}}, \frac{y_{\max} \cdot h_{\text{raw}}}{h_{\text{enc}}}\right]\right) \tag{3}$$

where $w_{\text{enc}} \times h_{\text{enc}}$ represents the resolution used by the vision encoder, typically constrained by a maximum pixel limit $P_{\max}$. The cropped region $\hat{R}$ is then re-encoded and integrated into the context as observation O_t. This strategy effectively increases the density of visual tokens for the selected region, enabling finer-grained perception despite the encoder's resolution constraints.

Trajectory Data Scaling-Up. Training the generator to effectively utilize visual perception actions requires high-quality trajectory data demonstrating proper action sequences. We propose a collaborative annotation strategy that leverages both large-scale foundation models and specialized expert models.

We employ a large vision-language model π_{LM} to determine the overall reasoning flow and action types for each trajectory. At step t, the large model generates both the reasoning thought T_t and the preliminary action A_t based on the accumulated history:

$$\{T_t, A_t\} = \pi_{\text{LM}}(\cdot \mid H_{t-1}) \tag{4}$$

This provides a high-level strategic plan for solving the query, encompassing decisions about document retrieval, visual perception, and answer generation.

Whenever the large model proposes a visual perception action, we invoke a specialized grounding model π_{EM} to determine precise bounding box coordinates. The expert model receives guidance from the large model's reasoning thought T_t, which provides contextual understanding about what visual

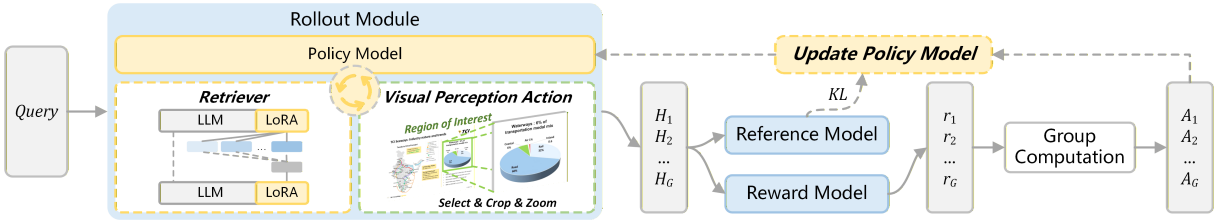

Figure 4: Reinforcement Learning Framework.

information is needed:

$$\hat{A}_t = \pi_{\text{EM}}(\cdot \mid H_{t-1}; T_t) \tag{5}$$

The refined coordinates $\hat{A}_t$ replace the initial coordinates in A_t, and the corresponding cropped region is encoded to form the observation $\hat{O}_t$:

$$\hat{O}_t = f_{\text{vis}}(O_{t-1}, \hat{A}_t) \tag{6}$$

where f_{vis} denotes the visual processing function that performs cropping, scaling, and encoding. This process ensures that trajectories exhibit both coherent high-level reasoning and accurate low-level visual grounding, providing diverse and precise training data for supervised fine-tuning(SFT) before RL.

3.4 Reinforcement Learning Framework with Iterative Reasoning

Reward Function. Effective RL for visual retrieval-augmented generation requires a comprehensive reward signal that captures multiple aspects of system performance. We design a comprehensive reward function comprising three complementary components that jointly guide the model toward efficient retrieval and accurate answer generation.

The primary objective of our framework is to encourage the model to retrieve relevant documents early in the interaction sequence through a retrieval quality reward. Retrieving relevant information promptly allows the model to build a coherent context without being distracted by excessive irrelevant content. We adapt the concept of Discounted Cumulative Gain(DCG) to measure retrieval quality. For a trajectory containing retrieved documents $D_{\text{trj}} = \{d_1, d_2, \ldots, d_n\}$ and a ground-truth set of relevant documents D_{gt}, we compute:

$$\begin{aligned} \text{DCG}(D_{\text{trj}}) &= \sum_{i=1}^{|D_{\text{trj}}|} \frac{2^{r_i} - 1}{\log_2(i+1)}, \\ r_i &= \begin{cases} 1, & d_i \in D_{\text{gt}} \\ 0, & d_i \notin D_{\text{gt}} \end{cases} \end{aligned} \tag{7}$$

where r_i indicates whether document d_i is relevant. The logarithmic discount factor emphasizes early retrieval of relevant documents. To normalize this metric, we define the ideal DCG as the score obtained when all relevant documents are retrieved first:

$$\text{IDCG}(D_{\text{gt}}) = \sum_{i=1}^{|D_{\text{gt}}|} \frac{1}{\log_2(i+1)} \tag{8}$$

The retrieval quality reward is then computed as the normalized ratio:

$$r_{\text{ret}} = \frac{\text{DCG}(D_{\text{trj}})}{\text{IDCG}(D_{\text{gt}})} \tag{9}$$

This formulation encourages the model to prioritize retrieving relevant documents while minimizing the retrieval of irrelevant ones.

To ensure that the model follows the predefined action structure during interaction, we introduce an action compliance reward as a pattern-based reward component. This reward evaluates whether the generated trajectory adheres to the expected format of alternating thoughts and actions. We employ a parsing function that extracts action tokens from the trajectory:

$$r_{\text{act}} = \text{Eval}(H) \tag{10}$$

where H represents the complete trajectory and $\text{Eval}(\cdot)$ checks for proper use of action tokens such as <search> and </search>. This component is particularly important during the initial stages of training to establish proper interaction patterns.

Rather than relying solely on rule-based evaluation, which may be limited in capturing semantic correctness, we employ a model-based answer quality reward evaluator. Given the input query Q, the ground-truth answer A_{gt}, and the model-generated answer A_{pred}, we use a reward model π_{RM} to assess answer quality:

$$r_{\text{ans}} = \pi_{\text{RM}}(Q, A_{\text{gt}}, A_{\text{pred}}) \tag{11}$$

This reward model is trained to evaluate semantic similarity and factual correctness, providing a more nuanced assessment than exact string matching.

The final reward function integrates all three components through a weighted combination:

$$r_\phi = \alpha \cdot r_{\text{ret}} + \beta \cdot r_{\text{ans}} + \gamma \cdot r_{\text{act}} \tag{12}$$

where $\alpha + \beta + \gamma = 1$. In practice, we set $\gamma = 0$ after the model has learned proper action patterns through SFT, focusing the reward on retrieval quality and answer correctness. During cold-start scenarios, we use $\gamma = 0.1$ to guide initial pattern learning.

Reinforcement Learning Framework. As shown in Figure 4, we formulate the training of the visual RAG agent as a RL problem where the policy model learns to interact with the retrieval environment through multi-step reasoning. Our framework builds upon recent advances in policy optimization for language models while introducing specialized mechanisms for handling multi-modal trajectories.

The learning objective maximizes the expected reward while constraining the policy model to remain close to a reference model, preventing catastrophic forgetting and maintaining stability:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x;\mathcal{V})} \Big[r_\phi(x, y) - \beta \cdot D_{\text{KL}}[\pi_\theta(y|x;\mathcal{V}) \,||\, \pi_{\text{ref}}(y|x;\mathcal{V})] \Big] \tag{13}$$

where $\mathcal{D}$ denotes the training dataset, π_θ is the policy model, π_{ref} is the reference model, β controls the strength of the KL constraint, x denotes the input query, and $\mathcal{V}$ represents the external environment. The notation $y \sim \pi_\theta(\cdot|x;\mathcal{V})$ indicates that trajectories are generated through the interaction between the policy model and the environment.

We implement a group relative policy optimization(GRPO) (Guo et al., 2025) strategy that generates multiple trajectory samples for each query and uses their relative performance to compute policy gradients.

4 Experiments

4.1 Datasets and Evaluation Metrics

We curate 500k unlabeled documents from the DocStruct4M (Hu et al., 2024) dataset to train the retriever, where each sample consists of a document image and its corresponding OCR text pair. We then fine-tune the retriever on the OpenDocVQA (Tanaka et al., 2025) dataset. We construct a dataset based on approximately 70k visual documents for RL training of the generator. We evaluate MDocRAG-RL on three challenging and visually-rich benchmark datasets: SlideVQA (Tanaka et al., 2023), ViDoSeek (Wang et al., 2025c), and MMLongBench (Ma et al., 2024).

4.2 Implementation Details

We use Phi3V (Abdin et al., 2024) as the base model for the retriever. During training, we apply LoRA (Hu et al., 2022) fine-tuning to the LLM component while keeping other parameters frozen, training for 1 epoch with the AdamW (Loshchilov and Hutter, 2017) optimizer and FlashAttention (Dao et al., 2022) acceleration. The batch size is set to 16 for pre-training and 64 for fine-tuning. The temperature parameter τ is set to 0.01. We employ Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct (Bai et al., 2025) as the base models for the generator, respectively. We utilize Qwen2.5-7B-Instruct (Yang et al., 2024) as the reward model π_{RM}, while Qwen2.5-VL-72B-Instruct and Qwen2.5-VL-32B-Instruct serve as π_{LM} and π_{EM}, respectively. We conduct SFT and RL on the llama-factory (Zheng et al., 2024) and verl (Sheng et al., 2025) frameworks, respectively. During the SFT stage, we adopt full-parameter fine-tuning with a cosine learning rate scheduler and a warmup ratio of 0.1. For GRPO algorithm training, the group size is set to 5, and the KL loss coefficient is typically set to 0.01; for cold start scenarios, it is set to 0 to disable the KL constraint on the model. All experiments are conducted on 8 H100 GPUs.

4.3 Comparison with State-of-the-Art Methods

We compare MDocRAG-RL against a comprehensive suite of baselines spanning several categories to validate its effectiveness. These include traditional RAG and reasoning frameworks (ReAct (Yao et al., 2023)), vision-based document retrieval systems (Vanilla RAG (Faysse et al., 2024), VDocRAG (Tanaka et al., 2025), ViDoRAG (Wang et al., 2025c)) and RL-enhanced RAG systems (Search-R1-VL (Guo et al., 2025), VRAG-RL (Wang et al., 2025d)). To demonstrate the scalability of our approach, all comparisons are performed on two vision-language models of different scales, Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct (Bai et al., 2025), as backbones.

Method	SlideVQA		ViDoSeek		MMLongBench					Overall
	Single-hop	Multi-hop	Extraction	Logic	Text	Table	Chart	Figure	Layout	
Qwen2.5-VL-3B-Instruct										
ReAct	16.2	11.3	7.1	13.8	2.9	3.4	3.6	2.9	5.3	11.2
Vanilla RAG	18.9	12.6	9.8	16.9	2.4	3.9	5.4	4.5	4.6	13.0
Search-R1-VL	25.8	19.7	20.6	30.2	8.2	8.1	7.6	9.6	7.3	21.5
VDocRAG	32.5	24.3	28.7	36.2	11.2	9.5	10.6	11.8	9.4	26.8
ViDoRAG	41.8	28.9	38.4	45.7	14.6	11.3	13.8	14.2	11.7	33.4
VRAG-RL	64.8	39.1	62.7	74.3	23.1	15.8	22.3	21.7	19.2	53.8
MDocRAG-RL	**71.5**	**44.7**	**69.3**	**78.9**	**26.1**	**19.4**	**25.2**	**24.6**	**22.3**	**58.4**
Qwen2.5-VL-7B-Instruct										
ReAct	35.3	20.8	27.9	41.6	10.4	12.1	10.5	6.5	6.9	27.2
Vanilla RAG	28.7	17.9	26.1	40.8	12.8	15.1	16.2	4.6	7.3	24.5
Search-R1-VL	47.9	41.8	40.9	50.8	19.5	13.7	12.6	11.7	10.5	37.2
VDocRAG	54.2	47.8	46.9	56.7	21.8	17.3	16.5	15.2	13.6	42.5
ViDoRAG	59.6	51.4	52.3	62.1	23.5	20.6	19.7	18.9	16.4	47.2
VRAG-RL	68.9	43.6	60.2	75.3	25.8	26.7	24.5	26.2	21.5	57.4
MDocRAG-RL	**75.9**	**50.3**	**66.2**	**80.1**	**29.4**	**30.2**	**28.3**	**29.5**	**24.7**	**62.5**

Table 1: Performance comparison across different benchmarks.

Retriever Pre-training	Reward		Generator	Acc
	Vanilla	RAG-Specific	Visual-Perception	
	✓			46.9
✓	✓			50.8
✓	✓		✓	52.4
		✓		55.2
✓		✓		58.6
✓		✓	✓	**62.5**

Table 2: Ablation study across three benchmarks.

As shown in Table 1, MDocRAG-RL achieves substantial improvements over all baselines across different model scales and benchmarks. With Qwen2.5-VL-3B-Instruct, our method achieves 58.4% overall accuracy, outperforming the strongest baseline by 4.6 points, with the advantage even more pronounced using the 7B model. On reasoning-intensive benchmarks like SlideVQA and ViDoSeek, our method demonstrates particularly strong performance, highlighting the effectiveness of our visual perception action space for complex multi-modal reasoning. On MMLongBench, our method consistently outperforms baselines across diverse visual content types including tables, charts, and figures. Compared to other RL-based methods, MDocRAG-RL shows consistent advantages, validating the effectiveness of our specialized retriever pre-training, visual perception action space, and RAG-specific reward design.

4.4 Ablation Study

To validate the contribution of each component, we conduct ablation studies across all three benchmarks using Qwen2.5-VL-7B-Instruct. As shown in Table 2, retriever pre-training brings a 3.9-point improvement, demonstrating that specialized training helps MLLMs compress textual and visual information into embeddings while maintaining semantic alignment for retrieval scenarios. Adding visual perception actions enables progressive information acquisition from coarse to fine granularity. The RAG-specific reward design yields substantial improvements over vanilla RL rewards, highlighting the importance of jointly optimizing retrieval quality and generation accuracy. The complete MDocRAG-RL system achieves 62.5%, delivering a 15.6-point improvement over the vanilla RL baseline. Notably, visual perception actions yield larger gains when combined with pre-trained retriever and RAG-specific rewards compared to their isolated effect, indicating that accurate retrieval provides better candidates for fine-grained visual understanding while RAG-specific rewards guide more strategic use of perception actions.

5 Analysis and Discussion

5.1 Pre-Training Improves Retrieval Efficiency

Pre-training significantly enhances the retriever's ability to identify relevant documents across all benchmark datasets. As illustrated in Figure 5, the self-supervised pre-training strategy in MDocRAG-RL enables MLLMs to compress textual and visual information into compact embeddings while maintaining semantic alignment. This approach is particularly valuable for document retrieval scenarios where labeled query-document pairs are scarce. The consistent improvements across diverse bench-

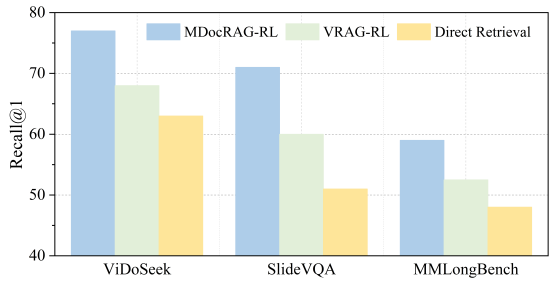

Figure 5: Analysis of retrieval performance.

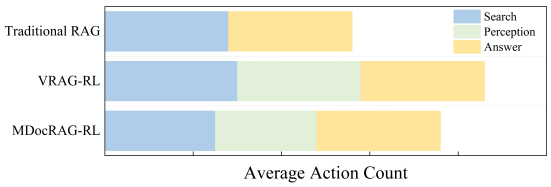

Figure 6: Analysis of latency in generation.

marks validate that pre-training helps the model learn robust visual-semantic representations that generalize well to different types of multi-modal documents and query patterns.

5.2 Visual Perception Action Space Enables Fine-Grained Understanding

The visual perception action space in MDocRAG-RL allows the generator to progressively refine its understanding by perceiving specific regions of interest within retrieved documents. As demonstrated in Table 2, visual perception actions contribute notable performance improvements, with gains being more pronounced when combined with RAG-specific rewards. Analysis reveals that the model predominantly focuses on information-dense areas such as tables, charts, and detailed text blocks, aligning well with human reasoning patterns. This coarse-to-fine information acquisition strategy proves particularly beneficial for multi-hop reasoning, where the ability to dynamically adjust visual focus enables more efficient token allocation and enhances the model's capability to handle complex reasoning scenarios.

5.3 Reinforcement Learning Enhances complex Reasoning

Reinforcement learning plays a crucial role in teaching the RAG agents effective interaction strategies, particularly for multi-hop reasoning tasks. The RAG-specific reward function proves essential by explicitly rewarding early retrieval of relevant documents, leading to more strategic behavior with reduced retrieval attempts while simultaneously improving retrieval precision. Moreover, RL training enhances the model's use of visual perception actions, demonstrating that the model learns when fine-grained analysis is necessary. This adaptive decision-making capability enables MDocRAG-RL to balance between efficiency and thoroughness depending on query complexity, resulting in superior performance on complex reasoning benchmarks.

5.4 Time Efficiency

Despite introducing visual perception actions and multi-step reasoning, MDocRAG-RL achieves superior accuracy-latency trade-offs compared to baseline methods, as shown in Figure 6. The moderate increase in total latency represents a worthwhile investment given the substantial accuracy improvements delivered. Notably, our RL framework helps mitigate latency through more efficient retrieval strategies. The search phase becomes more streamlined as improved retrieval quality reduces the need for extensive search iterations. This efficiency gain demonstrates that learning more effective retrieval strategies through reinforcement learning can partially offset the computational overhead introduced by additional reasoning steps.

6 Conclusion

In this paper, we introduce MDocRAG-RL, a novel multi-modal RAG framework tailored for complex visual reasoning tasks over document collections. Our approach integrates three key innovations: specialized pre-training tasks that enable MLLMs to effectively compress visual documents into retrieval-oriented embeddings, a visual perception action space that facilitates progressive coarse-to-fine information acquisition, and a reinforcement learning framework with RAG-specific rewards that enhances the complex visual reasoning capability of the RAG system. Extensive experiments across diverse benchmarks demonstrate the significant effectiveness of our method, achieving state-of-the-art performance with notable improvements over existing methods. Future work will explore extending the visual perception action space to support richer interaction patterns and investigating scalability to larger architectures and broader document domains.

Limitations

While our framework enhances the retrieval and complex visual reasoning capabilities of RAG systems, several limitations remain. First, although our method achieves a favorable accuracy-latency trade-off, it still introduces additional latency compared to traditional RAG approaches. Second, despite incorporating a visual perception action space, the richness of our actions remains limited compared to the diverse actions employed by humans when processing complex information. Therefore, optimizing the interaction logic between the generator and retriever, as well as exploring richer visual perception actions, constitute key directions for future work.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv:2404.14219*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ruisheng Cao, Hanchong Zhang, Tiancheng Huang, Zhangyi Kang, Yuxin Zhang, Liangtai Sun, Hanqi Li, Yuxun Miao, Shuai Fan, Lu Chen, and 1 others. 2025. Neusym-rag: Hybrid neural symbolic retrieval with multiview structuring for pdf question answering. *arXiv preprint arXiv:2505.19754*.

Jennifer Chen, Aidar Myrzakhan, Yaxin Luo, Hassaan Muhammad Khan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. 2025a. Drag: Distilling rag for slms from llms to transfer knowledge and mitigate hallucination via evidence and graph-based distillation. *arXiv preprint arXiv:2506.01954*.

Zhe Chen, Yusheng Liao, Shuyang Jiang, Pingjie Wang, Yiqiu Guo, Yanfeng Wang, and Yu Wang. 2025b. Towards omni-rag: Comprehensive retrieval-augmented generation for large language models in medical applications. *arXiv preprint arXiv:2501.02460*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.

Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. 2025. Mmdocir: Benchmarking multi-modal retrieval for long documents. *arXiv preprint arXiv:2501.08828*.

Manuel Faysse, Hugues Sibille, Tony Wu, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. ColPali: Efficient document retrieval with vision language models. *arXiv:2407.01449*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. 2025. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*.

Zhenyu Hou, Ziniu Hu, Yujiang Li, Rui Lu, Jie Tang, and Yuxiao Dong. 2025. Treerl: Llm reinforcement learning with on-policy tree search. *arXiv preprint arXiv:2506.11902*.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and 1 others. 2024. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Qing Li. 2025. Removal of hallucination on hallucination: Debate-augmented rag. *arXiv preprint arXiv:2505.18581*.

Pengcheng Jiang, Jiacheng Lin, Lang Cao, Runchu Tian, SeongKu Kang, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2025. Deepretrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. *arXiv preprint arXiv:2503.00223*.

Guangchen Lan, Huseyin A Inan, Sahar Abdelnabi, Janardhan Kulkarni, Lukas Wutschitz, Reza Shokri, Christopher G Brinton, and Robert Sim. 2025. Contextual integrity in llms via reasoning and reinforcement learning. *arXiv preprint arXiv:2506.04245*.

Mingzhe Li, Jing Xiang, Qishen Zhang, Kaiyang Wan, and Xiuying Chen. 2025a. Flipping knowledge distillation: Leveraging small models' expertise to enhance llms in text matching. *arXiv preprint arXiv:2507.05617*.

Qiwei Li, Teng Xiao, Zuchao Li, Ping Wang, Mengjia Shen, and Hai Zhao. 2025b. Dialogue-rag: Enhancing retrieval for llms via node-linking utterance rewriting. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24423–24438.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025c. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*.

Xiao Liang, Zhong-Zhi Li, Yeyun Gong, Yang Wang, Hengyuan Zhang, Yelong Shen, Ying Nian Wu, and Weizhu Chen. 2025. Sws: Self-aware weakness-driven problem synthesis in reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.08989*.

Pei Liu, Xin Liu, Ruoyu Yao, Junming Liu, Siyuan Meng, Ding Wang, and Jun Ma. 2025. Hm-rag: Hierarchical multi-agent multimodal retrieval augmented generation. *arXiv preprint arXiv:2504.12330*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Yuxing Lu, Gecheng Fu, Wei Wu, Xukai Zhao, Sin Yee Goi, and Jinzhuo Wang. 2025. Doctorrag: Medical rag fusing knowledge with patient analogy through textual gradients. *arXiv preprint arXiv:2505.19538*.

Haoran Luo, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, and 1 others. 2025. Hypergraphrag: Retrieval-augmented generation via hypergraph-structured knowledge representation. *arXiv preprint arXiv:2503.21322*.

Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, Nan Du, and Jia Li. 2025. S^2r: Teaching llms to self-verify and self-correct via reinforcement learning. *arXiv preprint arXiv:2502.12853*.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, and 1 others. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010.

Federico Ravenda, Seyed Ali Bahrainian, Andrea Raballo, Antonietta Mira, and Noriko Kando. 2025. Are llms effective psychological assessors? leveraging adaptive rag for interpretable mental health screening through psychometric practice. *arXiv preprint arXiv:2501.00982*.

Junhao Shen, Haiteng Zhao, Yuzhe Gu, Songyang Gao, Kuikun Liu, Haian Huang, Jianfei Gao, Dahua Lin, Wenwei Zhang, and Kai Chen. 2025. Semi-off-policy reinforcement learning for vision-language slow-thinking reasoning. *arXiv preprint arXiv:2507.16814*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

Jiashuo Sun, Xianrui Zhong, Sizhe Zhou, and Jiawei Han. 2025. Dynamicrag: Leveraging outputs of large language model as feedback for dynamic reranking in retrieval-augmented generation. *arXiv preprint arXiv:2505.07233*.

Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2025. Vdocrag: Retrieval-augmented generation over visually-rich documents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24827–24837.

Ryota Tanaka, Taichi Iki, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2024. Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19071–19079.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645.

Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, and 1 others. 2025. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. *arXiv preprint arXiv:2506.01713*.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025a. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*.

Jiayu Wang, Yifei Ming, Zixuan Ke, Caiming Xiong, Shafiq Joty, Aws Albarghouthi, and Frederic Sala. 2025b. Beyond accuracy: Dissecting mathematical reasoning for llms under reinforcement learning. *arXiv preprint arXiv:2506.04723*.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025c. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. 2025d. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. *arXiv preprint arXiv:2505.22019*.

Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. 2025. Real-mm-rag: A real-world multi-modal retrieval benchmark. *arXiv preprint arXiv:2502.12342*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Senqiao Yang, Junyi Li, Xin Lai, Bei Yu, Hengshuang Zhao, and Jiaya Jia. 2025. Visionthink: Smart and efficient vision language model via reinforcement learning. *arXiv preprint arXiv:2507.13348*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Kongcheng Zhang, Qi Yao, Shunyu Liu, Yingjie Wang, Baisheng Lai, Jieping Ye, Mingli Song, and Dacheng Tao. 2025a. Consistent paths lead to truth: Self-rewarding reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.08745*.

Wenlin Zhang, Xiangyang Li, Kuicai Dong, Yichao Wang, Pengyue Jia, Xiaopeng Li, Yingyi Zhang, Derong Xu, Zhaocheng Du, Huifeng Guo, and 1 others. 2025b. Process vs. outcome reward: Which is better for agentic rag reinforcement learning. *arXiv preprint arXiv:2505.14069*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.