MM-R5: MultiModal Reasoning-Enhanced ReRanker via Reinforcement Learning for Document Retrieval

Mingjun Xu[1*], **Jinhan Dong**[1*], **Jue Hou**[1*], **Zehui Wang**[1],
Sihang Li[1], **Zhifeng Gao**[1], **Renxin Zhong**[2], **Hengxing Cai**[1†]

[1]DP Technology, Beijing, China
[2]School of Intelligent Systems Engineering, Sun Yat-Sen University, Shenzhen, China

Abstract

Multimodal document retrieval systems enable information access across text, images, and layouts, benefiting various domains like document-based question answering, report analysis, and interactive content summarization. Rerankers improve retrieval precision by reordering retrieved candidates. However, current multimodal reranking methods remain underexplored, with significant room for improvement in both training strategies and overall effectiveness. Moreover, the lack of explicit reasoning makes it difficult to analyze and optimize these methods further. In this paper, We propose MM-R5, a **M**ulti**M**odal **R**easoning-Enhanced **ReR**anker via **R**einforcement Learning for Document **R**etrieval, aiming to provide a more effective and reliable solution for multimodal reranking tasks. MM-R5 is trained in two stages: supervised fine-tuning (SFT) and reinforcement learning (RL). In the SFT stage, we focus on improving instruction-following and guiding the model to generate complete and high-quality reasoning chains. To support this, we introduce a novel data construction strategy that produces rich, high-quality reasoning data. In the RL stage, we design a task-specific reward framework, including a reranking reward tailored for multimodal candidates and a composite template-based reward to further refine reasoning quality. We conduct extensive experiments on MMDocIR, a challenging public benchmark spanning multiple domains. MM-R5 achieves state-of-the-art performance on most metrics and delivers comparable results to much larger models on the remaining ones. Moreover, compared to the best retrieval-only method, MM-R5 improves recall@1 by over 4%. These results validate the effectiveness of our reasoning-enhanced training pipeline. Our code is available at https://github.com/i2vec/MM-R5.

1 Introduction

Multimodal document retrieval enhances traditional retrieval by incorporating visual modalities, (Cui et al. 2021; Sassioui et al. 2023; Lee et al. 2024) offering substantial benefits for tasks such as document-based question answering, report analysis, and interactive content summarization (Hudson and Manning 2019; Marino et al. 2019). The quality of retrieval results serves as a critical determinant

---

[*]These authors contributed equally.
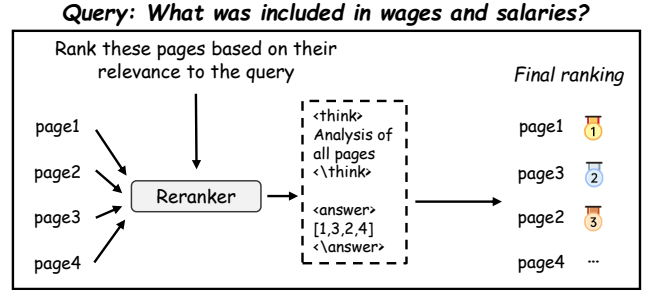
[†]Corresponding authors.

Figure 1: Workflow of MM-R5. It takes all candidate pages at once, analyzes all images, and then outputs the reasoning process along with a relevance ranking.

of the overall performance in downstream tasks (Nogueira and Cho 2019; Zhuang et al. 2025; Moreira et al. 2024), and rerankers further enhance retrieval quality by performing fine-grained sorting of all candidate samples prior to generation. In text-only settings, rerankers leveraging large language models have demonstrated robust performance in document ranking tasks (Ma et al. 2023; Sun et al. 2025; Xu 2024; Zhuang et al. 2023b; Sun et al. 2023). Nevertheless, adapting these approaches to multimodal contexts presents substantial challenges due to the added complexity introduced by visual content.

Recent advances in vision-language models (VLMs) have shown promising potential in multimodal reranking tasks. However, despite some exploratory efforts in this field (Chen et al. 2024; Chaffin and Lac 2024), significant limitations still remain: (i) current training-based approaches are limited both in number and effectiveness, although general-purpose VLMs can be used as rerankers, their performance in instruction-following and task-specific reasoning is often suboptimal without fine-tuning, leaving substantial room for performance improvement and methodological exploration; (ii) to the best of our knowledge, no existing method in multimodal reranking has incorporated Chain-of-Thought (CoT) reasoning. This absence results in a lack of interpretability in how the model assesses the relevance between candidates and the query, thereby hindering further analysis and optimization. To address these gaps, we aim to develop a more powerful multimodal reranker that explicitly integrates

CoT, enabling both improved performance and interpretable reasoning.

To this end, we design a two-stage training framework consisting of supervised fine-tuning (SFT) and reinforcement learning (RL). In the first stage, we construct a dataset containing high-quality reasoning paths and apply SFT to help the model adapt to multi-page reasoning and develop structured reasoning patterns. In the second stage of RL, inspired by prior work (Guo et al. 2025; Zhuang et al. 2025), we leverage Group Relative Policy Optimization (GRPO) to further enhance the model's reasoning ability, enhancing reranking performance while maintaining logical clarity and structural consistency in the reasoning process. Based on this design, we obtain MM-R5, a reasoning-enhanced multimodal reranker designed to overcome the limitations of existing methods. Built upon VLMs and empowered by our two-stage training scheme, MM-R5 demonstrates strong multi-page understanding capabilities and produces interpretable, step-by-step reasoning during the reranking process. As illustrated in Figure 1, our model jointly process all candidate pages, analyze their relevance to the query individually, and produce both an explicit reasoning chain and a final ranking. This approach enabled us to achieve second place in the WWW 2025 Multimodal RAG Challenge. Our main contributions are as follows:

- We design a two-stage training pipeline to create MM-R5, a powerful reranker specifically for multimodal document retrieval. Specifically, we construct a high-quality dataset with well-annotated reasoning paths for supervised fine-tuning, and combine it with a task-specific reinforcement learning phase using GRPO to develop MM-R5.

- We conduct experiments on publicly available benchmark datasets, where our method achieves state-of-the-art (SoTA) performance across multiple evaluation metrics, while remaining metrics are comparable or even better to those of significantly larger models.

- We integrate MM-R5 with various retrieval models and consistently observe performance improvements, demonstrating the stability, effectiveness, and strong adaptability of our approach.

- Our method generates explicit reasoning chains, enhancing interpretability and offering a new path toward more controllable and trustworthy multimodal retrieval systems.

2 Related Work

2.1 Multimodal Document Retrieval

Multimodal document retrieval focuses on extracting relevant information from documents that incorporate rich visual structures, including images, tables, and complex layouts. Driven by the rapid advancement of VLMs (Alayrac et al. 2022; Bai et al. 2025), retrieval methodologies have transitioned from conventional text-only paradigms to multimodal approaches that integrate textual and visual data. Recent investigations (Wei et al. 2024; Koukounas et al. 2024) delve into retrieval across diverse combinations of textual

and visual inputs, striving to bridge the modality gap and foster more holistic retrieval frameworks. DSE (Ma et al. 2024) leverages large VLMs to encode document screenshots directly into dense representations, facilitating efficient retrieval without OCR dependency. ColPali (Faysse et al. 2024) enhances performance further by constructing high-quality multi-vector embeddings derived from document images and implementing a late interaction mechanism, achieving notable outcomes.

Despite these advantages, existing methods mainly focus on retrieval performance, with the exploration of multimodal rerankers remaining limited. Therefore, in this paper, we aim to contribute an effective approach specifically for multimodal reranking.

2.2 Reranker

The reranker plays a pivotal role in the fine-grained reordering of candidates retrieved by the initial retriever, thereby enhancing the alignment between results and the user's query. For RAG systems, effective high-quality reranking not only significantly boosts reasoning efficiency but also minimizes the risk of introducing noise. In recent years, large language models (LLMs) have been widely adopted for reranking tasks owing to their superior capabilities in language understanding and generation (Zhuang et al. 2023a; Guo et al. 2024; Luo et al. 2024). Likewise, VLMs have begun preliminary explorations into multimodal reranking (Chen et al. 2024; Chaffin and Lac 2024), although current research in this domain is still nascent compared to unimodal scenarios. The complexity of multimodal data amplifies the noise caused by incorrect retrievals, thereby escalating the importance of retrieval accuracy in multimodal RAG systems. Designing a more efficient multimodal reranker has thus become the key to overcoming the performance bottleneck of current multimodal RAG frameworks.

2.3 Multimodal Reasoning

Recent concurrent studies have achieved substantial advancements in multimodal reasoning. For example, R1-OneVision (Yang et al. 2025) developed a cross-modal reasoning framework that converts images into structured visual representations. These representations are subsequently utilized to construct a visual reasoning dataset with the assistance of a language model. The Vision-Language Model (VLM) is initially trained on this dataset and further refined through RL. Similarly, R1-V (Chen et al. 2025) integrates the GRPO method (Shao et al. 2024) Reasoning in Open Language Models] from DeepSeek R1 (Guo et al. 2025) into VLM training, achieving remarkable performance in object-counting tasks—allowing a 3B model to outperform its 72B counterpart.

Vision-R1 (Huang et al. 2025) adopts a data-centric approach by generating a multimodal CoT dataset derived from visual content. This dataset is used for cold-start training and subsequently enhanced via GRPO. VisualThinker-R1-Zero (Zhou et al. 2025) demonstrated that applying R1 techniques to a base VLM rather than an instruction-tuned
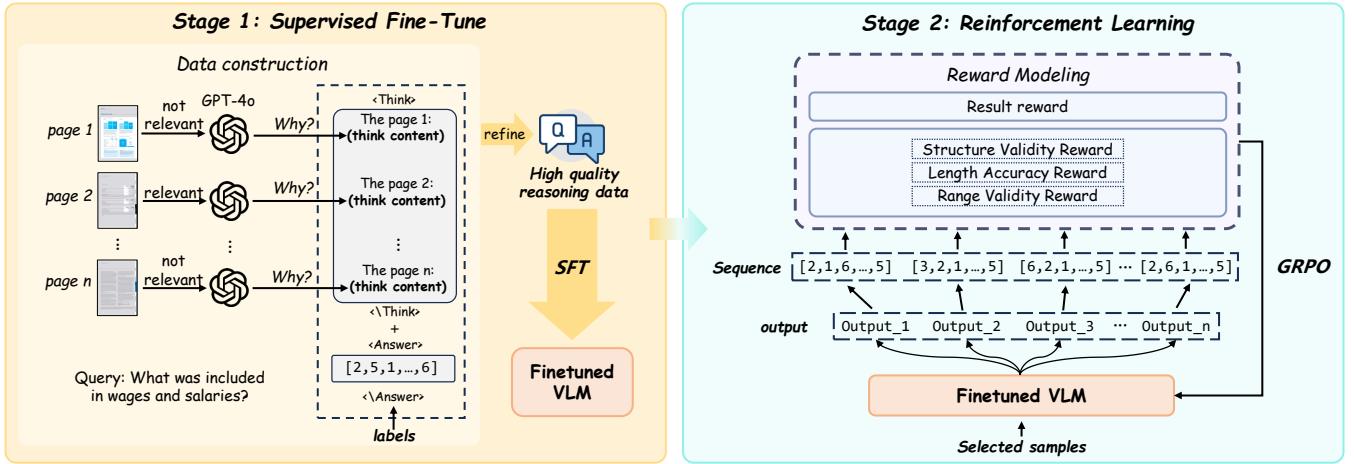
Figure 2: Overview of our two-stage training pipeline.

model yields greater performance improvements and induces the so-called "visual aha moment." A similar phenomenon was observed in MM-Eureka (Meng et al. 2025), which employed RLOO (Ahmadian et al. 2024) to train both an 8B instruction-tuned VLM and a 38B base model, resulting in more sophisticated and extended reasoning outputs.

Curr-ReFT (Deng et al. 2025) proposed a three-stage RL framework featuring progressively challenging reward levels, while LMM-R1 (Peng et al. 2025) introduced a rule-based two-stage RL strategy that first enhances text-based reasoning before addressing complex multimodal tasks. Although many of these works focus on multimodal mathematical reasoning (Lu et al. 2023; Wang et al. 2024; Zhang et al. 2024a), studies such as Visual-RFT (Liu et al. 2025) and VLM-R1 (Shen et al. 2025) extend RL-based reasoning to visual perception tasks.

Building upon these insights, we observe that reasoning has consistently demonstrated its effectiveness across a wide range of multimodal tasks, from visual question answering to complex mathematical reasoning. These advances naturally lead us to explore the incorporation of reasoning into multimodal reranking.

3 Method

In this section, we provide a detailed description of MM-R5. Specifically, we begin with a clear definition of the task, then we describe our two-stage training process, including the data construction and training procedure for each stage. However, our focus differs slightly between the two: in the first stage of SFT, we emphasize the construction of high-quality reasoning data; while in the second stage of reinforcement learning, we focus on the design of the reward function and its significance.

3.1 Problem Formulation

We address the task of multimodal reranking, which aims to reorder a set of retrieved multimodal candidates based on their relevance to a user query. The objective is to produce a ranking that prioritizes the most relevant items to support downstream generation tasks.

Formally, each task instance consists of:

- q, a natural language query representing the user's information need;
- $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$, a set of retrieved multimodal candidates, where each d_i contains visual and textual information.

Given the query q and the candidate set $\mathcal{D}$, the reranker outputs a predicted index list $\hat{\mathcal{I}}$ that specifies a ranking permutation over $\mathcal{D}$, where smaller indices indicate higher predicted relevance to the query q. For evaluation, we denote the golden set of relevant candidate indices as $\mathcal{G}$, where $|\mathcal{G}| \leq |\mathcal{D}|$.

3.2 Stage 1: Supervised Fine-Tune

3.2.1 Data Construction

For general reasoning tasks, a common data construction pipeline involves prompting a strong language model to generate multiple responses containing reasoning paths based on the original question, followed by rule-based or manual rejection sampling to select high-quality samples. However, this approach presents significant limitations in the context of multimodal reranking, as addressed in this work. First, existing VLMs often struggle with multi-image inputs, especially when the number of images increases, resulting in poor reasoning quality. Second, in relevance ranking tasks, these models tend to over-focus on the most obviously relevant candidates while neglecting a thorough analysis of the less relevant ones, ultimately leading to suboptimal ranking accuracy.

To address these issues and construct high-quality SFT data with explicit reasoning paths, we design a novel data construction pipeline based on GPT-4o, as illustrated in Figure 2. Unlike traditional methods that directly ask the model to reason over multiple images simultaneously, our pipeline adopts a single-image reasoning strategy. Specifically, we pair each image with the original question

Please rank the following images according to their relevance to the question.

task description

Provide your response in the format:
<think>your reasoning process here</think>
<answer>[image_id_, image_id_2, ...]</answer>
Where the numbers in the list represent the ranking order of images id from most to least relevant. Before outputting the answer, you need to analyze each image and provide your analysis process.
For example:
<think>Image 1 shows the most relevant content because...</think>
<answer>[id for the most relevant image, id for the second relevant image, ...]</answer>

format requirement

The question is: {Question}
There are {Number of images} images, id from 1 to {Number of images}, Image ID to image mapping:
Image 1: <image>
Image 2: <image>
...
Image N: <image>

task input

Figure 3: Prompt template for Chain-of-Thought based Reranking.

to create multiple single-image sub-tasks. For each sub-task, the model is instructed to explain why the image is or is not relevant to the query, guided by its label, resulting in individual image-level reasoning statements. These statements are then matched with their corresponding image indices and concatenated in order to form a complete reasoning chain, which is placed within <think>...</think> tags to serve as the model's reasoning path. And for the final ranked list of image indices, we sort the images based on their relevance labels—placing relevant images at the front and irrelevant ones at the end—and wrap the list in <answer>...</answer> tags. By combining this structured <think>...</think> <answer>...</answer> response with the original prompt, we obtain a high-quality SFT training sample. We then leverage GPT-4o to further refine the reasoning path of this sample, making it more concise while preserving the completeness and clarity.

Using this pipeline, we constructed 7,200 high-quality training instances for SFT, which played a crucial role in enhancing the model's instruction-following and reasoning capabilities.

3.2.2 Training

After obtaining a dataset with high-quality reasoning paths, we fine-tune a pretrained VLM on it. This stage enables the model to better align with the multimodal reranking task by significantly enhancing its multi-page reasoning abilities and fostering systematic reasoning patterns. In particular, the model learns to explicitly compare each retrieved image with the query, evaluating their relevance individu-

ally rather than relying on shallow or incomplete reasoning. This mitigates the risk of missing relevant candidates due to under-reasoning, a common failure mode in general-purpose VLMs. As a result, this supervised fine-tuning phase lays a solid foundation for the subsequent reinforcement learning stage, where the model's reranking ability is further optimized.

3.3 Stage 2: Reinforcement Learning

3.3.1 Resolution-Balanced Sampling Strategy

Unlike the SFT stage, which requires constructing task-specific data, the training data for the reinforcement learning phase is directly sampled from the original training corpus.

To mitigate the potential bias introduced by image resolution variations during reinforcement learning, we design a resolution-balanced data sampling strategy. Specifically, the original training dataset is first partitioned into 10 subsets, where each subset contains samples with approximately similar image sizes. To ensure uniform representation across different resolution ranges, we perform sampling without replacement from each subset. From the full dataset, we select a total of 3,000 samples, allocating them proportionally across all subsets to form the final resolution-balanced training set.

Since image resolutions within each subset are relatively consistent, this proportional sampling ensures that all resolution segments contribute equally during training, without increasing the total data volume. This strategy effectively reduces gradient estimation bias caused by imbalanced resolution distributions, leading to more stable and reliable reinforcement learning.

3.3.2 Reward Modeling

As illustrated in Fig 2, after obtaining the resolution-balanced training data, we perform reinforcement learning on the model using GRPO. To enhance the model's reranking capability while ensuring that its output remains precise and well-formatted, we design two complementary reward functions: a result reward and a format reward.

Result reward. We propose a novel reward formulation specifically designed to optimize relevance-based reranking performance. Let the candidate set be denoted by $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$, and let $\hat{\mathcal{I}} = [\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_k]$ represent the predicted indices over $\mathcal{D}$ output by the reranker. The golden set of relevant item indices is denoted as $G \subseteq \{1, \ldots, n\}$. For each predicted index $\hat{i}_j \in \hat{\mathcal{I}}$, we define a binary indicator s_j, where $s_j = 1$ if $\hat{i}_j \in G$, and $s_j = 0$ otherwise.

$$R_{result} = \frac{\sum_{j=1}^{|\hat{\mathcal{I}}|} \frac{s_j}{j^3}}{\sum_{j=1}^{|\mathcal{G}|} \frac{1}{j^3}}, \quad s_j = \begin{cases} 1, & \text{if } \hat{i}_j \in \mathcal{G} \\ 0, & \text{if } \hat{i}_j \notin \mathcal{G} \end{cases} \quad (1)$$

The numerator sums the inverse cubes of the predicted ranks for all correctly identified items, which gives much higher weight to items ranked at the top, especially the first position. The denominator represents the maximum possible reward when all ground-truth items are ranked in the top K positions. This normalization ensures the final reward R falls within the range $[0, 1]$ and reduces the effect of varying numbers of ground-truth items across samples. Compared to standard NDCG with logarithmic discounting, using the cube of the rank provides stronger emphasis on top-ranked predictions.

Format reward. In addition to relevance-oriented optimization, we define a format reward R_{format} to ensure that the model produces structurally correct outputs. This reward is composed of three independent and differentiable components that jointly evaluate the legality of output formatting, the deviation in list length, and the validity of index values. Let o denote the raw model output, and let $\hat{\mathcal{I}}$ denote the predicted index list derived from o after parsing and deduplication. Let $G \subseteq \{1, \ldots, n\}$ be the golden set of relevant indices, where N is the total number of reference candidates. The list $\hat{\mathcal{I}}$ is treated as a set-valued list, in which any duplicate indices are merged to ensure uniqueness. The reward components are defined as follows:

- **Structure Validity Reward**: If the output sequence o is properly enclosed within the required tags `<think>...</think>` and `<answer>...</answer>`, we assign a structure validity reward $R_{\text{valid}} = 1$; otherwise, $R_{\text{valid}} = 0$. This hard constraint ensures that any formatting error leads to a zero format reward, regardless of content correctness.

- **Length Accuracy Reward**: Let $abs(\cdot)$ donate the absolute value function. The length accuracy reward is de-

fined as:

$$R_{\text{len}} = 1 - \frac{abs(|\hat{\mathcal{I}}| - n)}{n} \quad (2)$$

This component encourages the model to generate answers of equal length to the reference.

- **Range Validity Reward**: To ensure that predicted indices remain within the valid range, we define the reward as follows:

$$R_{\text{range}} = \frac{\sum_{i \in \hat{\mathcal{I}}} \delta(i)}{|\hat{\mathcal{I}}|}, \quad \delta(i) = \begin{cases} 1, & \text{if } 1 \leq i \leq n, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

This reward encourages the model to generate indices strictly within the valid range and penalizes any out-of-bound predictions.

The final format reward is computed as:

$$R_{\text{format}} = R_{\text{valid}} \times R_{\text{len}} \times R_{\text{range}}. \quad (4)$$

This formulation provides the following benefits: (i) R_{valid} enforces strict adherence to the output structure through a hard-stop penalty, (ii) R_{len} and R_{range} offer smooth gradient signals when outputs are nearly correct, (iii) the three factors are independently adjustable, allowing fine-grained control over the penalty strength for different error types during training.

3.3.3 Prompt Design

To ensure fairness and consistency across all CoT based rerankers, we adopt a unified prompting strategy that standardizes the reasoning process by explicitly separating reasoning and final answers using structured markers, as illustrated in Figure 3. The prompt consists of three components: (i) a target description, which instructs the model to rank images based on their relevance to the query, (ii) a format requirement, which clearly defines the expected output structure, and (iii) a structured task input containing the query and a list of images, notably, each image is preceded by a textual identifier indicating its image ID, which helps the model better associate the visual content with its corresponding ID. This carefully designed prompt not only enforces output consistency across models but also improves interpretability and grounding by explicitly linking image content with their IDs.

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset

We conduct our experiments on MMDocIR (Dong et al. 2025), a large-scale benchmark specifically designed for multimodal document retrieval and reasoning. MMDocIR comprises both a training set and an evaluation set, providing rich modality diversity and reflecting the complexity of real-world documents.

The training set includes 6,878 documents and 73,843 QA pairs sourced from seven existing DocVQA-related datasets, e.g., MP-DocVQA (Tito, Karatzas, and Valveny 2023), SlideVQA (Tanaka et al. 2023), TAT-DQA (Zhu et al. 2022),

Algorithm	Method	Macro			Micro		
		Recall@1	Recall@3	Recall@5	Recall@1	Recall@3	Recall@5
Retriever-only	CLIP (Radford et al. 2021)	0.3334	0.5428	0.6452	0.3227	0.5337	0.6350
	E5-V (Jiang et al. 2024)	0.4201	0.6470	0.7389	0.4249	0.6578	0.7519
	DSE (Ma et al. 2024)	0.5109	0.7194	0.7925	0.5023	0.7234	0.7971
	GME (Zhang et al. 2024b)	0.5400	0.7603	0.8308	0.5421	0.7603	0.8377
	ColQwen (Faysse et al. 2024)	0.6481	0.8331	0.8766	0.6354	0.8213	0.8667
Reranking from ColQwen	RagVL	0.3814	0.6462	0.7667	0.3411	0.6206	0.7497
	Qwen2.5-VL-7B (Bai et al. 2025)	0.6230	0.8243	0.8675	0.6012	0.8069	0.8555
	Qwen2.5-VL-7B-cot	0.6479	0.8179	0.8670	0.6336	0.8103	0.8604
	Qwen2.5-VL-32B (Bai et al. 2025)	0.6422	0.8161	0.8669	0.6309	0.8118	0.8636
	Qwen2.5-VL-32B-cot	<u>0.6768</u>	<u>0.8500</u>	<u>0.8842</u>	<u>0.6609</u>	**0.8448**	**0.8800**
	Gemma3-12B (Team et al. 2025)	0.4743	0.6675	0.7278	0.4456	0.6353	0.7025
	Gemma3-12B-cot	0.5403	0.7729	0.8292	0.5166	0.7536	0.8153
	Qwen2.5-VL-7B-sft	0.6673	0.8475	0.8828	0.6498	0.8366	0.8746
	Qwen2.5-VL-7B-rl	0.6586	0.8370	0.8827	0.6454	0.8309	0.8744
	MM-R5(ours)	**0.6951**	**0.8520**	**0.8842**	**0.6759**	<u>0.8401</u>	<u>0.8755</u>

Table 1: Performance on MMDocIR. "Retriever-only" indicates that we only used the results from the Retriever as the final output, and "Reranking from ColQwen" means that we used the reranker model to reorder the candidates output by ColQwen. Methods suffixed with "cot" require the model to generate a chain-of-thought reasoning process. All three of our methods incorporate reasoning chains. "MM-R5" refers to the model trained with the full two-stage pipeline, while the variants marked with "sft" and "rl" indicate models trained only with supervised fine-tuning or reinforcement learning, respectively. The bolded metrics indicate the best performance, while the underlined metrics represent the second best.

SciQAG (Wan et al. 2024), DUDE (Van Landeghem et al. 2023), CUAD (Hendrycks et al. 2021), spanning diverse domains and document lengths. Meanwhile, the evaluation set comprises 313 lengthy documents, averaging 65.1 pages each, spanning ten distinct domains such as academic, legal, and financial texts. Each domain is treated as an individual subset for evaluation. These documents integrate multiple modalities, including text, images, tables, and layout/meta elements, and are accompanied by 1,658 expert-annotated questions. Each question is annotated with both page-level and layout-level ground-truth labels, enabling fine-grained and precise assessment. Importantly, many questions demand complex reasoning that involves cross-modal understanding, multi-page synthesis, and layout-aware interpretation, making this benchmark particularly challenging.

We adopt MMDocIR for both training and evaluation due to its comprehensive supervision signals and its ability to assess a model's capacity to retrieve and reason over complex multimodal content at the page level.

4.1.2 Evaluation Metrics
We follow (Dong et al. 2025) and adopt Recall@k as the primary evaluation metric. MM-R5 scores each page in the document based on its relevance to the input question and returns the top-k pages with the highest scores. Recall@k is defined as the proportion of ground-truth pages that appear within the top-k retrieved results. This metric reflects the model's ability to correctly identify and prioritize the most relevant pages containing evidence required to answer the question. In addition, We report both micro and macro Re-

call@k: micro averages recall across all individual samples, while macro first computes the average recall within each subset and then averages across subsets, providing a more balanced view across varying data distributions.

4.1.3 Model and Training Configuration
We adopt Qwen2.5-VL-7B (Bai et al. 2025) as the baseline model for our two-stage training. In the first stage, we fine-tune the baseline model using the swift (Zhao et al. 2024), which allows efficient SFT on our carefully constructed dataset. This process yields a finetuned model with enhanced instruction-following and reasoning capabilities. In the second stage, we further train the model using the GRPO algorithm within the open-source VLM-R1 (Shen et al. 2025) framework, where Low-Rank Adaptation (LoRA) is applied to improve training efficiency.

The training is conducted with the following hyperparameter settings:

- In the stage of supervised fine-tune, the training runs for 1 epoch with a learning rate of 1×10^{-4}, a batch size of 1, and gradient accumulation over 4 steps. LoRA is applied for efficient parameter tuning, with the rank lora_r set to 8 and the scaling factor lora_alpha set to 32.

- In the stage of reinforcement learning, the rollout number is set to 4, and the training runs for 1 epoch. We use a learning rate of 1×10^{-5}, a batch size of 1, and apply gradient accumulation over 2 steps. LoRA is used to improve parameter efficiency during fine-tuning, with the rank lora_r set to 64 and the scaling factor lora_alpha set

Retriever	MM-R5	Macro			Micro		
		Recall@1	Recall@3	Recall@5	Recall@1	Recall@3	Recall@5
CLIP (Radford et al. 2021)	w/o	0.3334	0.5428	0.6452	0.3227	0.5337	0.6350
	w	0.5942 (+26.08%)	0.6749 (+13.21%)	0.7173 (+7.21%)	0.5803 (+25.76%)	0.6598 (+12.61%)	0.7018 (+6.68%)
E5-V (Jiang et al. 2024)	w/o	0.4201	0.6470	0.7389	0.4249	0.6578	0.7519
	w	0.6236 (+20.35%)	0.7508 (+10.38%)	0.7966 (+5.77%)	0.6321 (+20.72%)	0.7516 (+9.38%)	0.8009 (+4.90%)
DSE (Ma et al. 2024)	w/o	0.5109	0.7194	0.7925	0.5023	0.7234	0.7971
	w	0.6487 (+13.78%)	0.7777 (+5.83%)	0.8213 (+2.88%)	0.6429 (+14.06%)	0.7817 (+5.83%)	0.8254 (+2.83%)
GME (Zhang et al. 2024b)	w/o	0.5400	0.7603	0.8308	0.5421	0.7603	0.8377
	w	0.6612 (+12.12%)	0.8095 (+4.92%)	0.8578 (+2.70%)	0.6551 (+11.30%)	0.8037 (+4.34%)	0.8566 (+1.89%)
ColQwen (Faysse et al. 2024)	w/o	0.6481	0.8331	0.8766	0.6354	0.8213	0.8667
	w	0.6951 (+4.70%)	0.8520 (+1.89%)	0.8842 (+0.76%)	0.6759 (+4.05%)	0.8401 (+1.88%)	0.8755 (+0.88%)

Table 2: Extended experiments validating the effectiveness of MM-R5 across different retrievers. "w" means reranking is applied using MM-R5, while "w/o" refers to using original retrieval results without reranking.

to 128. All experiments are carried out on 4 NVIDIA A100 GPUs, each with 80GB of memory.

4.1.4 Comparison Methods

To demonstrate the effectiveness of our proposed method, we conduct a comprehensive evaluation across both retrieval and reranking settings. For retrieval baselines, we compare against several strong models, including DSE (Ma et al. 2024), GME (Zhang et al. 2024b), and ColQwen (Faysse et al. 2024). While for reranking, we compare our method with the prior multimodal reranking model RAG-VL (Chen et al. 2024), as well as representative vision-language models such as Qwen2.5-VL-32B (Bai et al. 2025) and Gemma3-12B (Team et al. 2025), covering a wide range of capabilities in vision-language understanding and instruction following. To ensure fair comparison in reranking, we fix the retrieval backbone to generate the top-10 candidate pages. This consistent candidate set allows us to isolate the impact of different reranking strategies under identical input conditions.

4.2 Experimental Results

4.2.1 Main Results

As illustrated in Table 1, Our method, MM-R5, consistently achieves SoTA performance across most evaluation metrics, including both macro and micro recall at different thresholds. Compared with all retriever-only methods, MM-R5 brings significant improvements, for instance, compared with the previous SoTA retriever ColQwen, macro recall@1 improves from 0.6481 to 0.6951, and micro recall@1 improves from 0.6354 (ColQwen) to 0.6759, demonstrating the effectiveness of our reranking strategy in refining the initial retrieval results.

Furthermore, when compared to strong reranking baselines, including large-scale general-purpose models such as Qwen2.5-VL-32B and Gemma3-12B, MM-R5 still achieves the best performance, despite being based on a much smaller baseline, Qwen2.5-VL-7B. Specifically, MM-R5 improves macro Recall@1 from 0.5403 to 0.6951 and micro Recall@1 from 0.5166 to 0.6759 compared to Gemma3-12B-cot, achieving remarkable gains of 0.1548 and 0.1593, re-

spectively. It also surpasses the much larger Qwen2.5-VL-32B-cot by a notable margin in both metrics. These results collectively validate the design of our method: with effective supervision and well-designed training strategies, our reranker not only enhances retrieval performance but also outperforms much larger vision-language models.

4.2.2 Ablation Studies

In this part, we investigate the impact of incorporating CoT reasoning into multimodal reranking tasks, and examine the necessity of our proposed two-stage training framework.

Impact of CoT. As shown in Table 1, incorporating CoT reasoning consistently improves reranking performance. For instance, Qwen2.5-VL-7B with CoT improves macro recall@1 by 2.49 and micro recall@1 by 3.24, while Qwen2.5-VL-32B achieves gains of 3.46 and 3.00 in macro and micro recall@1 respectively. These results suggest that CoT reasoning can effectively enhance the model's reranking capability. In particular, our introduced reasoning paradigm, which guides the model to explicitly evaluate the relevance between each image and the query before producing the final ranking—not only leads to improved reranking performance, but also enhances the interpretability of the model's reasoning process. The discussion of the interpretability will be detailed in the case study section.

Impact of two-stage training. Table 1 also includes two ablated variants of our model: Qwen2.5-VL-7B-sft, trained only with supervised fine-tuning, and Qwen2.5-VL-7B-rl, trained only with reinforcement learning. Both fall short of the full model on every metric. For example, in macro Recall@1 the full model reaches 0.6951, whereas Qwen2.5-VL-7B-sft and Qwen2.5-VL-7B-rl obtain 0.6673 (-2.78) and 0.6586 (-3.65), respectively. A similar pattern appears in micro Recall@1: MM-R5 achieves 0.6759, surpassing the SFT-only and RL-only variants by 2.61 and 3.05. Although these single-stage variants substantially outperform the baseline Qwen2.5-VL-7B-cot (macro Recall@1: 0.6479, and micro Recall@1: 0.6336), the full two-stage model further amplifies these gains, confirming that SFT and RL are complementary, validating the necessity and synergy of our two-stage training pipeline.

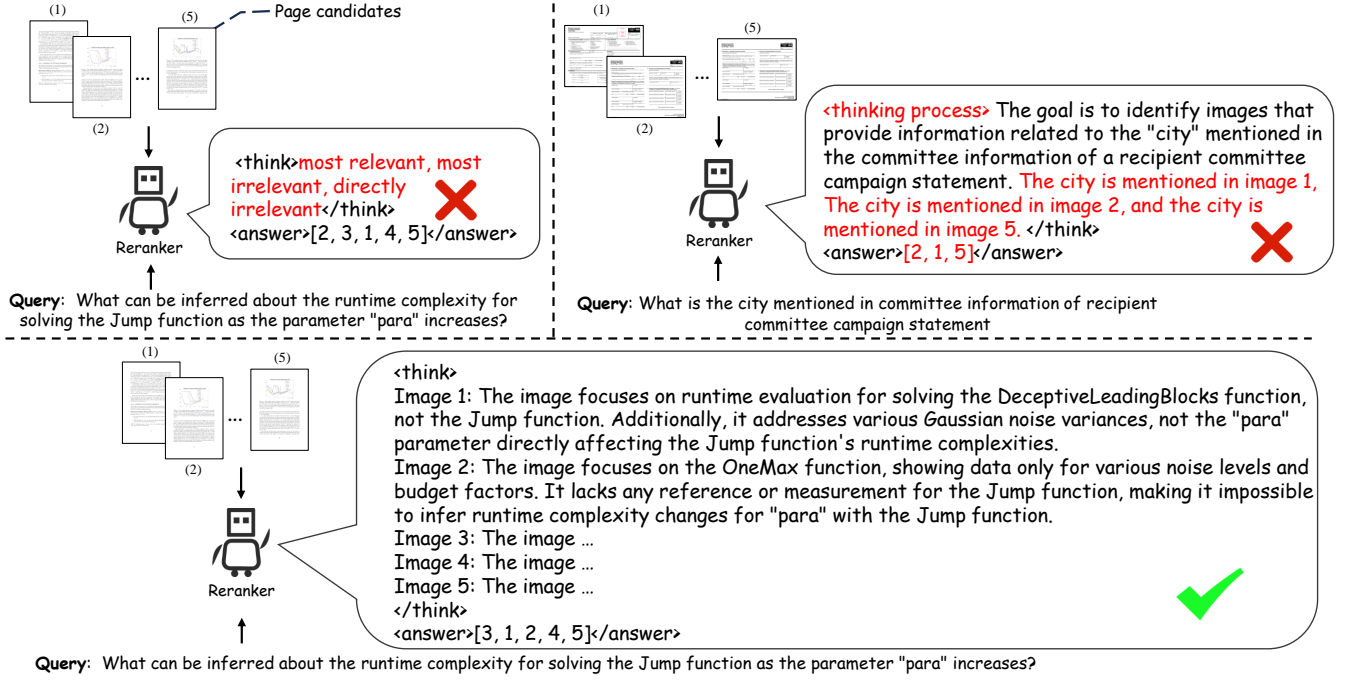**Task: Output a sequence based on the relevance between images and the query.**

Figure 4: Comparison between our model and multimodal rerankers without training. All page candidates are retrieved by ColQwen. The red-highlighted parts in the model's output indicate segments that are of poor quality. The example in the top-right illustrates that the reranker model not only fails to follow the instruction to produce a sequence as required but also exhibits insufficient reasoning in its response, while example in the top-left corner reveals that its reasoning process totally lacks interpretability. In contrast, our model demonstrates significantly improved instruction-following ability, and exhibits more coherent and interpretable reasoning steps.

4.2.3 Generalization Across Retrievers

To further evaluate the generalization ability of our reranker, we conduct extended experiments by applying MM-R5 on top of five representative multimodal retrievers: CLIP (Radford et al. 2021), E5-V (Jiang et al. 2024), DSE, GME, and ColQwen.

As shown in Table 2, across all retrievers and evaluation metrics, applying MM-R5 consistently improves both macro and micro recall. For instance, on strong retrievers like GME, recall@1 improves from 0.5400 to 0.6612 (macro) and from 0.5421 to 0.6551 (micro); on ColQwen, recall@1 improves from 0.6481 to 0.6951 (macro) and from 0.6354 to 0.6759 (micro). In addition to recall@1, MM-R5 also brings consistent improvements on recall@3 and recall@5. These results demonstrate that MM-R5 is highly effective and robust across different retrieval backbones, confirming its general applicability as a plug-and-play reranking component. Importantly, the relative gains are particularly significant for weaker retrievers like CLIP and E5-V, indicating that MM-R5 not only refines strong baselines but also substantially enhances simpler retrievers.

4.2.4 Case Study For Interpretability

In previous experiments, we demonstrated that incorporating CoT reasoning significantly boosts model performance.

However, compared to MM-R5, directly prompting existing general untrained VLMs to produce CoT often leads to insufficient and unclear reasoning, resulting in poor interpretability. As illustrated in Figure 4, the upper examples use our baseline model, Qwen2.5-VL-7B, as the reranker, while the lower example shows results from our MM-R5. In the top-right case, the baseline model already struggles with format adherence, moreover, its reasoning merely notes that pages 1, 2, and 5 mention the word "city" from the query, failing to consider other relevant candidates and providing unclear, superficial explanations. The top-left example underscores the problem of vague and insufficient justifications, where the model provides no reasoning at all. In contrast, as shown in the lower part, when presented with the same sample as in the top-left example, MM-R5 demonstrates strong adherence to the required output format and delivers well-structured analysis for each image before presenting the final answer. This highlights that our trained reranker is capable of generating interpretable CoT reasoning, which facilitates deeper analysis and supports future improvements.

5 Conclusion

In this paper, we propose MM-R5, a reasoning-enhanced multimodal reranker trained via a two-stage pipeline. In

the first stage, supervised fine-tuning, we introduce a novel reasoning-oriented data construction strategy that provides high-quality instruction-following and reasoning examples to effectively guide the model's initial learning. In the second stage, we design a reinforcement learning framework tailored to the multimodal reranking task, incorporating a customized reward function that promotes reranking effectiveness while preserving structured and well-formed outputs. Extensive experiments demonstrate the effectiveness of MM-R5 in improving reranking quality and highlight its generality across various retriever backbones.

References

Ahmadian, A.; Cremer, C.; Gallé, M.; Fadaee, M.; Kreutzer, J.; Pietquin, O.; Üstün, A.; and Hooker, S. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chaffin, A.; and Lac, A. 2024. MonoQwen: Visual Document Reranking.

Chen, L.; Li, L.; Zhao, H.; Song, Y.; and Vinci. 2025. R1-V: Reinforcing Super Generalization Ability in Vision-Language Models with Less Than $3. https://github.com/Deep-Agent/R1-V. Accessed: 2025-02-02.

Chen, Z.; Xu, C.; Qi, Y.; and Guo, J. 2024. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439*.

Cui, L.; Xu, Y.; Lv, T.; and Wei, F. 2021. Document ai: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*.

Deng, H.; Zou, D.; Ma, R.; Luo, H.; Cao, Y.; and Kang, Y. 2025. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint arXiv:2503.07065*.

Dong, K.; Chang, Y.; Goh, X. D.; Li, D.; Tang, R.; and Liu, Y. 2025. MMDocIR: Benchmarking Multi-Modal Retrieval for Long Documents. *arXiv preprint arXiv:2501.08828*.

Faysse, M.; Sibille, H.; Wu, T.; Omrani, B.; Viaud, G.; Hudelot, C.; and Colombo, P. 2024. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Guo, F.; Li, W.; Zhuang, H.; Luo, Y.; Li, Y.; Yan, L.; and Zhang, Y. 2024. Generating diverse criteria on-the-fly to improve point-wise LLM rankers. *CoRR*.

Hendrycks, D.; Burns, C.; Chen, A.; and Ball, S. 2021. CUAD: an expert-annotated NLP dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Xu, Z.; Hu, Y.; and Lin, S. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Jiang, T.; Song, M.; Zhang, Z.; Huang, H.; Deng, W.; Sun, F.; Zhang, Q.; Wang, D.; and Zhuang, F. 2024. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*.

Koukounas, A.; Mastrapas, G.; Günther, M.; Wang, B.; Martens, S.; Mohr, I.; Sturua, S.; Akram, M. K.; Martínez, J. F.; Ognawala, S.; et al. 2024. Jina clip: Your clip model is also your text retriever. *arXiv preprint arXiv:2405.20204*.

Lee, J.; Ko, J.; Baek, J.; Jeong, S.; and Hwang, S. J. 2024. Unified Multi-Modal Interleaved Document Representation for Information Retrieval. *arXiv preprint arXiv:2410.02729*.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Luo, J.; Chen, X.; He, B.; and Sun, L. 2024. Prp-graph: Pairwise ranking prompting to llms with graph aggregation for effective text re-ranking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5766–5776.

Ma, X.; Lin, S.-C.; Li, M.; Chen, W.; and Lin, J. 2024. Unifying multimodal retrieval via document screenshot embedding. *arXiv preprint arXiv:2406.11251*.

Ma, X.; Zhang, X.; Pradeep, R.; and Lin, J. 2023. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 3195–3204.

Meng, F.; Du, L.; Liu, Z.; Zhou, Z.; Lu, Q.; Fu, D.; Han, T.; Shi, B.; Wang, W.; He, J.; et al. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*.

Moreira, G. d. S. P.; Ak, R.; Schifferer, B.; Xu, M.; Osmulski, R.; and Oldridge, E. 2024. Enhancing Q&A

Text Retrieval with Ranking Models: Benchmarking, fine-tuning and deploying Rerankers for RAG. *arXiv preprint arXiv:2409.07691*.

Nogueira, R.; and Cho, K. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Peng, Y.; Zhang, G.; Zhang, M.; You, Z.; Liu, J.; Zhu, Q.; Yang, K.; Xu, X.; Geng, X.; and Yang, X. 2025. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Sassioui, A.; Benouini, R.; El Ouargui, Y.; El Kamili, M.; Chergui, M.; and Ouzzif, M. 2023. Visually-Rich Document Understanding: Concepts, Taxonomy and Challenges. In *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, 1–7.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shen, H.; Liu, P.; Li, J.; Fang, C.; Ma, Y.; Liao, J.; Shen, Q.; Zhang, Z.; Zhao, K.; Zhang, Q.; Xu, R.; and Zhao, T. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

Sun, S.; Zhuang, S.; Wang, S.; and Zuccon, G. 2025. An investigation of prompt variations for zero-shot llm-based rankers. In *European Conference on Information Retrieval*, 185–201. Springer.

Sun, W.; Yan, L.; Ma, X.; Wang, S.; Ren, P.; Chen, Z.; Yin, D.; and Ren, Z. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542*.

Tanaka, R.; Nishida, K.; Nishida, K.; Hasegawa, T.; Saito, I.; and Saito, K. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 13636–13645.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Tito, R.; Karatzas, D.; and Valveny, E. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144: 109834.

Van Landeghem, J.; Tito, R.; Borchmann, Ł.; Pietruszka, M.; Joziak, P.; Powalski, R.; Jurkiewicz, D.; Coustaty, M.; Anckaert, B.; Valveny, E.; et al. 2023. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19528–19540.

Wan, Y.; Liu, Y.; Ajith, A.; Grazian, C.; Hoex, B.; Zhang, W.; Kit, C.; Xie, T.; and Foster, I. 2024. SciQAG: A Framework for Auto-Generated Science Question Answering Dataset with Fine-grained Evaluation. *arXiv preprint arXiv:2405.09939*.

Wang, K.; Pan, J.; Shi, W.; Lu, Z.; Ren, H.; Zhou, A.; Zhan, M.; and Li, H. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37: 95095–95169.

Wei, C.; Chen, Y.; Chen, H.; Hu, H.; Zhang, G.; Fu, J.; Ritter, A.; and Chen, W. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, 387–404. Springer.

Xu, Z. 2024. RankMamba: Benchmarking Mamba's Document Ranking Performance in the Era of Transformers. *arXiv preprint arXiv:2403.18276*.

Yang, Y.; He, X.; Pan, H.; Jiang, X.; Deng, Y.; Yang, X.; Lu, H.; Yin, D.; Rao, F.; Zhu, M.; et al. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Zhang, R.; Jiang, D.; Zhang, Y.; Lin, H.; Guo, Z.; Qiu, P.; Zhou, A.; Lu, P.; Chang, K.-W.; Qiao, Y.; et al. 2024a. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, 169–186. Springer.

Zhang, X.; Zhang, Y.; Xie, W.; Li, M.; Dai, Z.; Long, D.; Xie, P.; Zhang, M.; Li, W.; and Zhang, M. 2024b. GME: Improving Universal Multimodal Retrieval by Multimodal LLMs. *arXiv preprint arXiv:2412.16855*.

Zhao, Y.; Huang, J.; Hu, J.; Wang, X.; Mao, Y.; Zhang, D.; Jiang, Z.; Wu, Z.; Ai, B.; Wang, A.; Zhou, W.; and Chen, Y. 2024. SWIFT:A Scalable lightWeight Infrastructure for Fine-Tuning. arXiv:2408.05517.

Zhou, H.; Li, X.; Wang, R.; Cheng, M.; Zhou, T.; and Hsieh, C.-J. 2025. R1-Zero's" Aha Moment" in Visual Reasoning on a 2B Non-SFT Model. *arXiv preprint arXiv:2503.05132*.

Zhu, F.; Lei, W.; Feng, F.; Wang, C.; Zhang, H.; and Chua, T.-S. 2022. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4857–4866.

Zhuang, H.; Qin, Z.; Hui, K.; Wu, J.; Yan, L.; Wang, X.; and Bendersky, M. 2023a. Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels. *arXiv preprint arXiv:2310.14122*.

Zhuang, S.; Liu, B.; Koopman, B.; and Zuccon, G. 2023b. Open-source large language models are strong zero-shot query likelihood models for document ranking. *arXiv preprint arXiv:2310.13243*.

Zhuang, S.; Ma, X.; Koopman, B.; Lin, J.; and Zuccon, G. 2025. Rank-R1: Enhancing Reasoning in LLM-based Document Rerankers via Reinforcement Learning. arXiv:2503.06034.