End-to-End Optimization for Multimodal Retrieval-Augmented Generation via Reward Backpropagation

Anonymous ACL submission

Abstract

Multimodal Retrieval-Augmented Generation (MM-RAG) has emerged as a promising approach for enhancing the reliability and factuality of large vision-language models (LVLMs). While end-to-end loss backpropagation is infeasible due to non-differentiable operations during the forward process, current methods primarily focus on component-level optimizations, necessitate extensive component-specific training datasets and suffer from a gap between local and global optimization objectives. In this paper, we propose a new paradigm that backpropagates global rewards from the system output to each component and then transforms these rewards into specific local losses, enabling each component to perform gradient descent and thus ensuring end-to-end optimization. Specifically, we first insert two lightweight multimodal components, a query translator and an adaptive reranker, to address the heterogeneity of multimodal knowledge and the varying knowledge demands for different questions, and then tune only these inserted components using our proposed paradigm to integrate the entire system. Our method achieves SOTA performance on multiple knowledge-intensive multimodal benchmarks with high training efficiency, relying exclusively on supervised signals from an external reward model. Experimental results and our detailed analysis of the evolution of components during training collectively reveal the advantages and considerable potential of this paradigm as a promising direction for MM-RAG research.

1 Introduction

Large Vision Language Models (LVLMs) (Lu et al., 2024; Bai et al., 2025) have extended LLMs (Grattafiori et al., 2024; Jiang et al., 2024) with vision encoders (Radford et al., 2021; Oquab et al., 2024), enabling them to process visual inputs and achieve exceptional performance across various vision-language tasks. However, due to

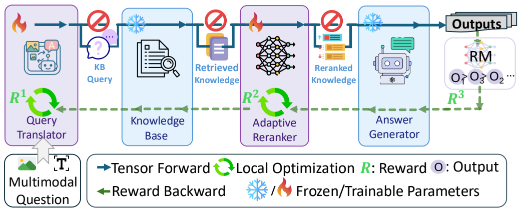

Figure 1: Non-differentiable tensor operations during the forward process render direct loss backpropagation infeasible by disrupting the tensor graph and preventing gradient flow. Our method instead sequentially propagates global rewards backward, converts them to component-specific local losses, and then applies gradient descent for optimization.

parameter capacity constraints and outdated parametric knowledge after pertaining, these models perform poorly on knowledge-intensive tasks, generating hallucinated responses lacking reliability and factuality. The research community has proposed Multimodal Retrieval-Augmented Generation (MM-RAG) (Khandelwal et al., 2020; Lewis et al., 2021; Caffagni et al., 2024) to provide additional contextual knowledge as a supplement to intrinsic parametric knowledge of models.

However, due to the discrete tensor operations between components that disrupt the computational graph, direct optimization through loss backpropagation and gradient descent is infeasible. Several approaches have sought to optimize individual components separately, but they need extensive component-specific training datasets and suffer from a misalignment between local and global objectives, even compromising the generalization of each component.

In this paper, we introduce a novel paradigm for MM-RAG that achieves end-to-end optimization by reward backpropagation, called MM-RewardRAG. As shown in Figure 1, after obtaining the system output from the answer generator, an external re-

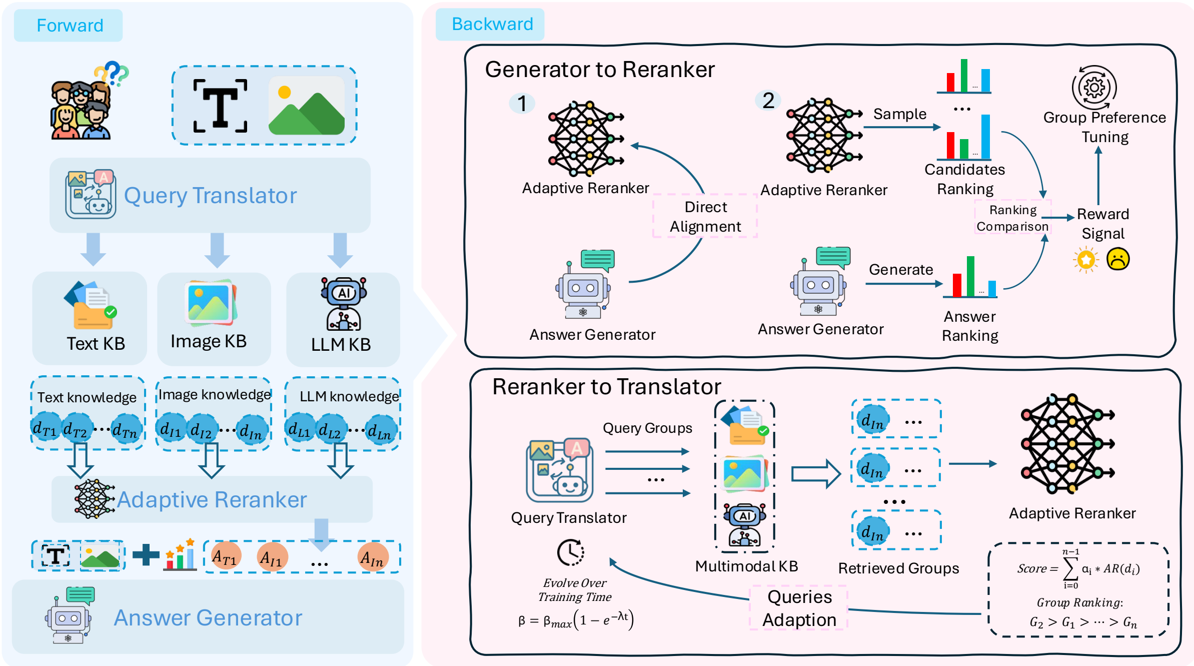

Figure 2: Illustration of MM-RewardRAG. (Left) Forward inference: Multimodal question processing utilizing a query translator to route queries to text, image, and LLM knowledge bases, with adaptive reranking subsequently applied for answer generation. (Right) Reward backpropagation optimization: upper part shows reward propagation via alignment based on direct instance and relative relationship within group ranking respectively, and lower part illustrates the process of query adaptation for integrating knowledge bases using a time-evolving reward signal.

ward model calculates the global reward, which is then backpropagated to each component and converted into a component-specific local loss to guide parameter optimization. To preserve general retrieval and instruction-following capabilities, our method tunes only the inserted lightweight components. As illustrated in more detail in Figure 2, to propagate the reward signal from the answer generator to the adaptive reranker, both traditional direct instance alignment and our proposed novel Group Preference Alignment are employed to model relative relationships within rankings. To guide the query translator, a group-weighted, time-evolving reward signal, derived from the ongoing optimization process, is backpropagated from the adaptive reranker to adapt the translator to heterogeneous knowledge bases.

We evaluate our approach on a diverse set of knowledge-intensive multimodal benchmarks, advancing beyond previous studies that focused solely on benchmarks requiring only textual knowledge, to incorporate those demanding both textual and visual knowledge to perform deep cross-modal reasoning. Our approach achieves SOTA performance on E-VQA, Infoseek, MultimodalQA, WebQA, OKVQA and A-OKVQA, with only 4k total training questions without human-labeled ground truth, depending entirely on the supervised signal from an external verifiable reward model. We provide a detailed analysis to interpret the evolution of components during the training process, demonstrating that our adaptive reranker surpasses three proprietary models using substantially more training data, despite sharing the same model architecture. Our contributions can be summarized as follows:

- We propose MM-RewardRAG, a novel paradigm enabling end-to-end optimization for MM-RAG systems.
- We introduce two multimodal components designed to address inherent challenges within MM-RAG, and offer an interpretative analysis of their learned evolution under our paradigm.
- Experimental results validate the effectiveness of our paradigm and underscore its considerable potential as a promising new direction for MM-RAG research.

2 Related Work

Large Vision Language Models. Recent LVLMs (Bai et al., 2025; Wang et al., 2024) have demonstrated remarkable capabilities by extending LLMs with multimodal alignment modules connected to vision encoders. However, the parametric knowledge of these models is capacity-constrained and outdated after pretraining, and insufficient multimodal alignment further compromises the knowledge already embedded in the language model backbone, which leads to hallucinations (Zhou et al., 2024) in model responses when encountering knowledge-intensive visual questions (Marino et al., 2019; Chen et al., 2023).

Retrieval Augmented Generation. RAG enhances the factuality and reliability of LLM responses while reducing hallucinations by retrieving relevant information from external knowledge bases. Recent works have extended RAG to the multimodal domain, retrieving text documents or image-text pairs as in-context examples to provide additional commonsense and visual knowledge. EchoSight (Yan and Xie, 2024) and Wiki-LLaVA (Caffagni et al., 2024) retrieve supplementary textual knowledge to improve LVLM performance on knowledge-intensive visual tasks. For domain-specific tasks, RULE (Xia et al., 2024b) and MMed-RAG (Xia et al., 2024a) enhance the factuality of medical LVLMs by retrieving relevant medical reports associated with radiology images. The scope of retrieved content further broadens to diverse visual inputs: V-RAG (Chu et al., 2025) extends retrieval to include similar images, MORE (Cui et al., 2024) leverages images for commonsense reasoning, and VisRAG (Yu et al., 2025) incorporates screenshots as a distinct document type. Regarding optimization strategies, SURf (Sun et al., 2024), RULE (Xia et al., 2024b), and RoRA-VLM (Qi et al., 2024) primarily aim to train the answer generator to selectively utilize retrieved information and avoid being misled by irrelevant or noisy data, while V-RAG (Chu et al., 2025) enables the answer generator to accept multiple interleaved multimodal inputs. Alternatively, VisRAG (Yu et al., 2025) focuses on training the retriever to better adapt to screenshots. Contrary to current works that focus on local component-level optimization, our proposed MM-RewardRAG aims to optimize the whole system end-to-end, which reduces the need for extensive training data and directly aligns local component objectives with the global system objective.

3 Methodology

3.1 Overview

In this section, we first introduce the necessity of heterogeneous knowledge bases for MM-RAG and then detail the proposed native multimodal components designed to integrate the MM-RAG system, followed by an explanation of the reward backpropagation algorithm that enables end-to-end optimization of the entire pipeline.

Notation. We denote the input question by Q; the query for a specific knowledge base by q_m, where $m \in \{T, I, L\}$ indicates the text, image, or language model, respectively; the i-th retrieved document by d_i; and the model's output based on d_i by O_i. Each item can be assigned a reward R_i^n corresponding to different stages n: query translation ($n = 1$), adaptive reranking ($n = 2$), and answer generation ($n = 3$). The ranking r_n for each stage is derived from the rewards set $\{R_i^n\}$; specifically for the adaptive reranking stage ($n = 2$), r_2 can be derived from either usefulness levels $\{l_i\}$, which are unique to this stage, or the rewards $\{R_i^2\}$.

3.2 Heterogeneous Knowledge Bases

Our framework leverages three distinct types of knowledge bases to address the inherent heterogeneity of multimodal information, which stems from the fact that some knowledge is intrinsically modality-specific: *Text KB* contains background content related to entities, including historical context, conceptual definitions, and specific details like times and names. *Image KB* provides information mainly embedded in the visual modality, such as landmarks, the relative sizes of different buildings, and visual attributes. *LLM-as-a-KB* is leveraged to supplement parametric knowledge interruption during the multimodal alignment process. We provide qualitative examples in Appendix H for interested readers.

3.3 Query Translator

The information needed to answer a multimodal question exists across different modalities in a complementary manner. However, using the question directly as a query to retrieve leads to poor performance and incomplete recall due to modality mismatch and semantic decoupling. We thus design a modality-aware query translator that generates queries adapted to different knowledge bases, serving as a soft connector to couple the multimodal question with the appropriate knowledge sources. The generation process can be modeled probabilistically as:

$$P(\{q_T, q_I, q_L\}|I, Q) = \prod_{j \in \{T,I,L\}} P(q_j|I, Q, q_{<j}) \quad (1)$$

where these queries are then used to retrieve from each knowledge base, resulting in the candidate sets $\{d_{T1}, d_{T2}, \ldots, d_{Tk}\}$, $\{d_{I1}, d_{I2}, \ldots, d_{Ik}\}$, and $\{d_{L1}, d_{L2}, \ldots, d_{Lk}\}$ that represent texts, images, and LLM responses, respectively.

3.4 Adaptive Reranker

This component is designed to address the variable knowledge demands across diverse questions. Traditional fixed *Top-K* reranking methods fail to accommodate the fluctuating information needs: they either introduce unnecessary noise when minimal context would suffice, or truncate critical information when more comprehensive knowledge is required. The adaptive reranker, instead, dynamically calibrates the information boundary to match the specific requirements of each question. Specifically, it evaluates each candidate and assigns different levels of usefulness, indicating how helpful the document is anticipated to be for the answer generator. These levels can then be converted to scalars to establish a ranking. Subsequently, all candidates deemed *useless* are discarded. The remaining candidates, along with their assigned usefulness levels, are passed to the answer generator, explicitly informing the generator about the potential utility of each piece of information, highlighting which might be noisy (e.g., *neutral*) and should be utilized selectively.

3.5 Optimization Process

The supervised signal for training is solely provided by an external reward model, which directly reflects the accuracy and factuality of the system output, and can be extended to other preference types that a function or parametric neural networks can model. The optimization objective of every component within the whole MM-RAG system is to collaboratively maximize this global reward. The core idea involves propagating this global reward backward to each component step-by-step. For each neural network-based component, the reward is then converted into a specific local loss, enabling gradient-based optimization of its parameters. To ensure the general retrieval and instruction following capabilities remain unaffected, which is necessary for robustness and transferability across different domains, our approach freezes the parameters of the retriever and answer generator while only adjusting the parameters of the inserted components to enhance coupling and alignment within the system. It is noted that the query translator can be viewed as a pre-retrieval domain shifter. Consequently, even if a distribution gap exists between the target data and the corpus used by the retriever, retrieval performance can still be specifically optimized. Furthermore, the adaptive reranker serves a dual function. Firstly, it acts as a post-retrieval filter to further improve retrieval results. Secondly, it operates as a coupler to ensure that the contextual knowledge supplied to the answer generator is both preferred and complementary to the parametric knowledge of the answer generator.

Generator to Reranker. We describe our algorithm following a backward component-by-component order to enhance clarity. For retrieved documents, the reranker assigns usefulness levels l_i to each item, producing a ranking r_2. These items are then individually paired with the input question Q and fed to the generator to produce corresponding answers O_i, which the reward model evaluates and assigns each score R_i^3 to create another ranking r_3. Ideally, the ranking r_2, provided by the reranker without assessing the final answer directly, should be consistent with r_3, but due to the misalignment between the reranker's local objective and the system's global goal, discrepancies exist between these rankings. We show the detailed analysis results in Appendix A. To address this problem, we propagate the final global reward R_i^3 backward to the reranker to obtain R_i^2, which is then transformed into a specific loss function for reranker optimization. This procedure consists of two sequential stages:

The first stage distills preference from the generator to the reranker directly, transferring awareness of knowledge usefulness for answering questions and aligning the two components. For an input question Q, the generator answers the question both with and without each candidate document d_i separately, then compares the results to determine whether the candidate is helpful. The outcome can be constructed into a restricted-format CoT reasoning sequence, which is then directly used to train the reranker:

$$\mathcal{L}_{distill} = -\sum_i \log P_{AR}(CoT, l_i|(Q, I), d_i) \quad (2)$$

The second stage, which we propose as Group

Preference Tuning, involves a comparative alignment process. The reranker first samples two groups of usefulness levels $\{l_i^a\}$, $\{l_i^b\}$ from n retrieved candidates $\{d_i\}$, generating two distinct rankings r_2^a and r_2^b. These rankings are then compared with the reference r_3 to determine which group demonstrates closer alignment with the desired outcome. Items within the better-aligned group are considered preferred. The reranker, parameterized by θ_{AR}, assigns usefulness levels $l(d;\theta_{AR})$ to each document d. We define the aggregate score for a group of documents G as $U(G;\theta_{AR}) = \sum_{d \in G} l(d;\theta_{AR})$. The Group Preference Tuning then aims to maximize the score difference between a preferred winner group (G_W) and a dispreferred loser group (G_L), where preference ($G_W \succ_{r_3} G_L$) is determined by their relative alignment with the reference ranking r_3. This objective is formalized by minimizing the following loss:

$$\mathcal{L}_{\text{Group}} = -\mathbb{E}_{(G_W,G_L) \text{ s.t. } G_W \succ_{r_3} G_L} \left[\log \sigma \left(U(G_W;\theta_{AR}) - U(G_L;\theta_{AR})\right)\right] \tag{3}$$

This loss guides the reranker to adjust its usefulness levels $l(d;\theta_{AR})$ to favor groups that better align with the global objective reflected in r_3. Unlike traditional methods that focus solely on the quality of individual item outputs, our proposed Group Preference Tuning emphasizes the relative relationships among multiple items within a group. The supervised signal extends beyond the prediction accuracy of single items to encompass the correctness of relative relationships among predictions across multiple items, making it inherently suitable for addressing the challenges posed by questions that require multi-hop reasoning across multiple ground truth documents.

Reranker to Translator. Since the adaptive reranker has been aligned with the global rewards of the system output, we continue to propagate signals derived from the reranker's evaluations backward to optimize the query translator. We denote the reinforcement signal for the j-th query group as R_j^1. For each multimodal question (Q, I) or textual question Q, the query translator (parameterized by θ_{QT}) generates n distinct groups of modality-specific queries $\mathbf{G_q} = \{G_{q,j} \mid G_{q,j} = \{q_{T,j}, q_{I,j}, q_{L,j}\}\}_{j=0}^{n-1}$. Each query group $G_{q,j}$ is then used to retrieve a corresponding ranked list of m candidate documents, denoted as $\mathbf{D_j} = (d_{j,1}, d_{j,2}, \ldots, d_{j,m})$. An optimal query translator should formulate queries that maximize the retrieval of pertinent knowledge, with high-utility documents concentrated at the top of each retrieved list $\mathbf{D_j}$. This minimizes the computational load on the adaptive reranker by providing a more focused set of initial candidates. To this end, we define a position-sensitive reward function $\mathcal{R}$ for each list $\mathbf{D_j}$ retrieved by its corresponding query group $G_{q,j}$:

$$\mathcal{R}(\mathbf{D_j}, \beta_t) = \sum_{k=1}^{m} \frac{l(d_{j,k};\theta_{AR}^*)}{(\log_2(k+1))^{\beta_t}} \tag{4}$$

where $l(d_{j,k};\theta_{AR}^*)$ is the usefulness score assigned to document $d_{j,k}$ (at rank k in list $\mathbf{D_j}$) by the previously aligned adaptive reranker, and $\beta_t \geq 0$ is a time-dependent position sensitivity parameter. This computed reward $\mathcal{R}(\mathbf{D_j}, \beta_t)$ serves as the reinforcement signal R_j^1 for optimizing the query translator's parameters θ_{QT} using appropriate policy optimization algorithms (e.g., PPO, GRPO for online settings, or adapting for offline settings like DPO). During the initial training phase of the query translator, we employ a curriculum learning strategy for β_t. We start with β_t close to zero by setting a large retrieval size m and a small initial β_0, which encourages the query translator to retrieve any relevant items, regardless of their position, thus initially optimizing for recall. As training progresses, the objective shifts to prioritize the placement of high-scoring documents at higher ranks. Thus, β_t evolves according to:

$$\beta_t = \beta_{\text{max}} \cdot \left(1 - e^{-\lambda \frac{t}{T_{\text{total}}}}\right) \tag{5}$$

where $\lambda > 0$ controls the rate of convergence to β_{max}. This evolving β_t adapts the reward landscape, guiding the query translator to generate queries that not only retrieve high-quality content but also rank it effectively, thereby enhancing the synergy with the subsequent component.

4 Experiments

Detailed experimental settings are provided in Appendix B.

4.1 Main Results

The experimental results on Infoseek and E-VQA benchmarks are presented in Table 1. Our approach demonstrates superior performance over all existing methods, including those leveraging proprietary search engines and models as well as open-source

Model	Retriever	Feature	E-VQA Single-Hop	E-VQA All	InfoSeek Unseen-Q	InfoSeek Unseen-E	InfoSeek All
Zero-shot MLLMs							
BLIP-2	-	-	12.6	12.4	12.7	12.3	12.5
InstructBLIP	-	-	11.9	12.0	8.9	7.4	8.1
LLaVA-v1.5	-	-	16.3	16.9	9.6	9.4	9.5
Qwen2-VL-Instruct†	-	-	16.4	16.4	17.9	17.8	17.9
Qwen2-VL-Instruct(sft)†	-	-	25.0	23.8	22.7	20.6	21.6
Retrieval-Augmented Models							
Wiki-LLaVA	CLIP ViT-L/14+Contriever	Textual	17.7	20.3	30.1	27.8	28.9
EchoSight	EVA-CLIP-8B	Visual	26.4	24.9	18.0	19.8	18.8
ReflectiVA	EVA-CLIP-8B	Visual	35.5	35.5	28.6	28.1	28.3
MM-RewardRAG (Ours)	Query-Translator	Native multimodality	**41.3**	**43.2**	**39.3**	**40.2**	**39.8**

Table 1: Comparative performance on Encyclopedia-VQA and Infoseek benchmarks. MM-RewardRAG demonstrates the ability to leverage diverse multimodal retrievers without being constrained to dataset-specific retrieval strategies.

Metrics	Text EM	Text F1	Image EM	Image F1	All EM	All F1
Hard Negatives						
Question-Only	15.4	18.4	11.0	15.6	13.8	-
AutoRouting	49.5	56.9	37.8	37.8	46.6	-
ImplicitDecomp	51.6	58.4	44.6	51.2	48.8	55.5
MuRAG	60.8	67.5	58.2	58.2	60.2	-
SKURG	66.1	69.7	52.5	57.2	59.8	64.0
Solar	69.7	74.8	55.5	65.4	69.8	66.1
PERQA	69.7	74.1	54.7	60.3	62.8	67.8
MM-RwardRAG (Ours)	**77.2**	**78.1**	**63.9**	**67.8**	**72.1**	**69.3**
Full Wiki						
AutoRouting	35.6	40.2	32.5	32.5	34.7	-
MuRAG	49.7	56.1	56.5	56.5	51.4	-
MM-RwardRAG (Ours)	**57.6**	**59.8**	**63.2**	**64.7**	**59.5**	**60.3**

Table 2: MultimodalQA evaluation results show that our approach surpasses other methods, including those specifically designed for different settings.

Method	**QA-FL**	**QA-Acc**	**QA**
Baseline	47.6	49.3	27.4
VLP + VinVL	47.6	49.6	27.5
VLP + x101fpn	46.9	44.3	23.8
OFA-Cap + GPT	52.8	55.4	33.5
PROMPTCAP + GPT	53.0	57.2	34.5
ETG	60.1	77.2	47.1
MM-RwardRAG (Ours)	**64.1**	**77.9**	**58.2**

Table 3: Evaluation results on WebQA.

LLMs and LVLMs, achieving new state-of-the-art results. Our method differs from others in three key aspects: 1) We only use 4k data samples for training, which reduces computational resource requirements and is efficient enough to be applied to other domains; 2) Only the query translator and adaptive reranker are fine-tuned to couple each component, thereby preserving the general visual instruction-following capability of the answer generator, which contrasts with previous methods that sacrifice general capabilities to obtain domain-specific performance; 3) The supervised signal is only provided by a reward model from the system output, and then distributed to each component through reward backpropagation, which ensures that the local optimization objectives of each component are aligned with the global goals for improved accuracy and factuality of the system output.

Results on MultimodalQA and WebQA benchmarks are presented in Table 2 and Table 3, respectively. Notably, previous MM-RAG methods mainly focus on scenarios requiring only textual knowledge, while neglecting those demanding joint multimodal reasoning across text and vision information. Additionally, current approaches addressing these two benchmarks typically aim to train task-specific models rather than developing general solutions, due to the challenges in effectively retrieving relevant cross-modality information and leveraging combined multimodal content. Despite these limitations, our method still outperforms all specialized fine-tuned models. Our proposed native multimodal query translator effectively leverages modality-specific retrieval methods by translating the original question into separate queries for different knowledge bases, while the adaptive reranker further enhances the quality of external knowledge for the answer generator, more effectively activating its cross-modality reasoning capability to generate superior answers.

Table 4 presents the results on OK-VQA and A-OKVQA benchmarks. It is observed that current LVLMs already possess sufficient parametric knowledge to answer these relatively outdated

Models	**OK-VQA(%)**	**A-OKVQA(%)**
ViLBERT	30.6	25.8
LXMERT	30.7	26.1
ClipCap	30.9	27.2
KRISP	33.7	29.4
GPV-2	48.6	39.3
REVEAL-Base	50.4	41.7
REVEAL-Large	51.5	42.8
REVEAL	52.2	44.5
MM-RwardRAG (Ours)	**66.1**	**63.8**

Table 4: Performance comparison on OK-VQA and A-OKVQA benchmarks.

questions accurately, and the naïve introduction of external knowledge potentially degrades model performance due to misleading content. However, our adaptive reranker effectively filters out noisy information, selectively retaining only knowledge beneficial to the answer generator, which develops an innovative fusion of contextual and parametric knowledge, thereby further enhancing overall performance.

	E-VQA	InfoSeek		
Model	Single-Hop	Un-Q	Un-E	All
KB Article				
Vanilla (Vicuna-7B)	34.1	5.3	4.3	4.7
Vanilla (LLaMA-3-8B)	72.9	10.0	7.9	8.8
Vanilla (LLaMA-3.1-8B)	73.6	15.2	13.9	14.5
LLaVA-v1.5 (Vicuna-7B)	42.9	14.2	13.4	13.8
LLaVA-v1.5 (LLaMA-3.1-8B)	54.1	20.1	17.7	18.8
Ours	**83.2**	**59.8**	**59.7**	**59.8**
KB Passages				
Wiki-LLaVA	38.5	52.7	50.3	51.5
Wiki-LLaVA ◇	46.8	51.2	50.6	50.9
ReflectiVA	75.2	57.8	57.4	57.6
Ours	**89.2**	**62.6**	**62.3**	**62.4**

Table 5: Oracle evaluation demonstrates our approach achieves a superior upperbound across different backbones and granularities, given identical ideal retrieval results.

We also evaluate our framework under oracle settings, with results presented in Table 5, demonstrating consistent superior performance with a substantial margin of improvement. This suggests that our approach can continuously benefit from advances in multimodal retrieval techniques. Our framework exhibits robust model transferability, which is detailed in Appendix C, enabling straightforward integration of emerging models. We will track developments in the field and report updated results as our method incorporates these advancements.

The transfer results on M2KR benchmarks using PreFLMR are presented in Table 6. Our method consistently outperforms all previous approaches, further demonstrating its robustness and generalization effectiveness.

Model	OKVQA	Infoseek	E-VQA
Zero-shot MLLMs			
RA-VQAv2	55.44	21.78	19.80
Qwen2-VL-Instruct	60.45	21.75	19.01
Retrieval-Augmented Models			
RA-VQAv2 w/ FLMR	60.75	-	-
RA-VQAv2 w/ PreFLMR	61.88	30.65	54.45
Qwen2-VL-Instruct w/ PreFLMR	46.99	24.68	51.81
Qwen2.5-VL-Instruct w/PreFLMR	65.07	30.74	53.89
Ours w/ PreFLMR	**66.02**	**44.44**	**63.28**

Table 6: Evaluation results on M2KR filtered benchmarks using PreFLMR as a retriever.

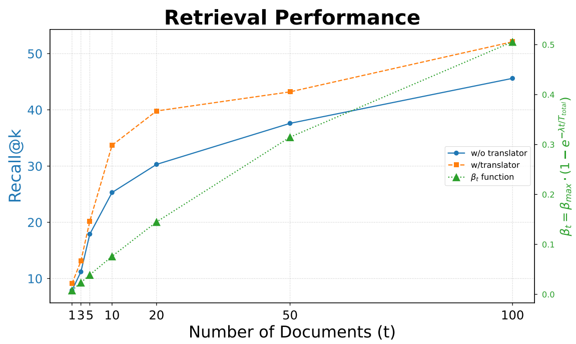

Figure 3: The impact of the query translator on retrieval performance.

4.2 Query Translator Analysis

In this section, we provide the answer to the question: *What has changed during the optimization process of the query translator?* As shown in Figure 3, the overall recall across heterogeneous knowledge bases increases as training progresses, with the position of pseudo documents gradually advancing toward the front of retrieval results until reaching a threshold. Notably, the optimized query translator enables top-50 retrieval results to achieve performance comparable to the original top-100, which allows us to set a lower value of n for the retriever, passing significantly fewer candidates to the reranker, thereby reducing computational resources and latency while maintaining comparable performance.

4.3 Reranker Comparison

We compared our adaptive reranker with other multimodal rerankers based on the `Qwen-VL` architecture, which includes (1) `Jina-reranker-m0`, including an additional post-trained MLP head to generate ranking scores measuring query-document

Models	Webqa				MMQA				Infoseek				E-VQA			
	NDCG	MAP	MRR	P@1	NDCG	MAP	MRR	P@1	NDCG	MAP	MRR	P@1	NDCG	MAP	MRR	P@1
Mono	73.51	70.99	74.69	61.47	82.81	79.48	80.23	71.73	-	-	-	-	-	-	-	-
Des	68.74	65.45	69.71	55.28	80.33	77.42	78.60	68.47	47.36	27.42	27.42	14.80	60.29	48.61	47.99	41.90
Jina	79.02	78.07	81.16	70.45	87.21	85.04	86.14	80.16	77.20	74.76	74.76	63.10	65.02	59.39	58.95	41.90
Ours	**90.87**	**91.22**	**91.79**	**85.62**	**92.00**	**92.11**	**92.50**	**86.68**	**100.0**	**100.0**	**100.0**	**100.0**	**97.34**	**97.03**	**100.0**	**100.0**

Table 7: Performance comparison of various reranker models across different benchmark hard-negative datasets.

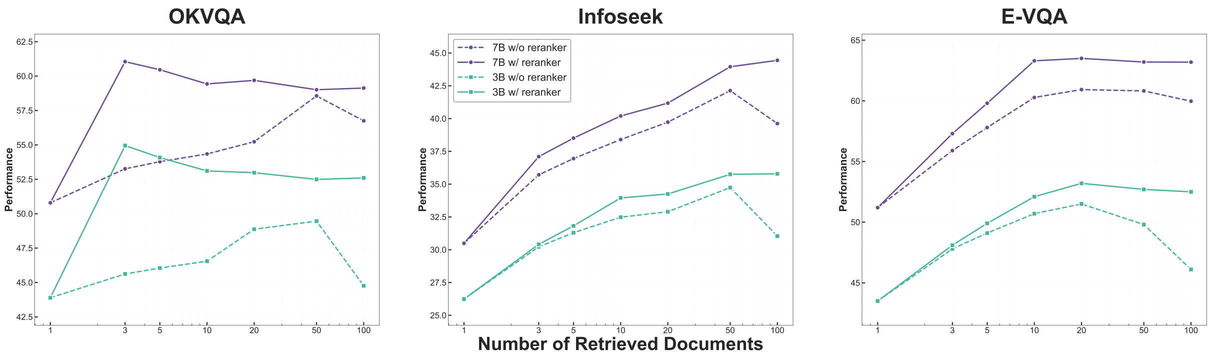

Figure 4: Scaling Law of retrieved documents on three datasets across models with different parameters.

relevance; (2) `Mono-reranker`, which compares the logits of two tokens (True and False) to obtain a relevancy score that can be used to rerank candidates; (3) `Dse-reranker`, which generates embeddings for the query and document separately and then calculates the relevance score. Unlike our approach that dynamically selects candidates based on usefulness, these models all use fixed top-k selection after ranking. Additionally, our model features versatile any-to-any modality support, contrasting with existing models constrained to fixed text-to-image or image-to-text pathways. As shown in Table 7, our adaptive reranker outperforms other competitors on four benchmarks across all metrics, with much less computing resources for training. For inference, our model achieves a throughput of 21.5k tokens per second on a single GPU, which significantly outperforms others that incorporate non-autoregressive structures, resulting in slower processing speeds despite higher GPU power utilization. Specifically, competing models consume substantially more power (e.g. `Jina`: 453 W, `Des`: 315 W, `Mono`: 426 W) compared to our model's 271 W.

4.4 Scaling Law

Figure 4 shows the scaling behavior of our proposed method on three benchmarks within M2KR as the number of retrieved documents increases, using the PreFLMR retriever. Despite minor fluctuations across different benchmarks, the trends remain consistent. Performance increases to reach an upper bound before subsequently declining, which demonstrates that indiscriminately retrieving more documents is not optimal due to the uncertainty in determining the ideal quantity for each dataset. However, when incorporating our adaptive reranker, which dynamically determines the optimal number of external knowledge sources for the answer generator regardless of the total retrieved documents, MM-RAG consistently achieves superior performance. Even when the retrieval count reaches high values, the knowledge sources passed to the answer generator remain effectively filtered, eliminating potentially misleading content. We provide additional ablation studies in Appendix C.

5 Conclusion

In this paper, we propose MM-RewardRAG, a novel end-to-end paradigm that optimizes MM-RAG systems by using reward backpropagation, achieving superior performance, robust transferability across diverse benchmarks and backbones, and high training efficiency with reduced resource needs. Our detailed analysis interprets the learning dynamics and component evolution during training, offering clear interpretability for the effectiveness of our paradigm. Looking ahead, we aim to continuously integrate advancements in multimodal retrieval to further enhance this framework, inspiring continued research in this promising direction.

Limitations

A primary factor limiting the upper-bound performance of MM-RewardRAG is the inherent capability of the employed retriever. While our method effectively trains lightweight components to couple with the retriever at both pre-retrieval (e.g., query translation) and post-retrieval (e.g., adaptive reranking) stages, the overall system's ability to surface relevant knowledge is ultimately constrained by the retriever's own performance. Consequently, future work incorporating more powerful pretrained retrievers will be essential to further boost the performance of our MM-RewardRAG system.

Ethics Statements

A core ethical benefit of the proposed MM-RewardRAG lies in its targeted approach to mitigating hallucinations in LVLMs. Addressing model-generated fabrications is critical, as hallucinations can lead to the spread of misinformation and erode user trust. By reducing such outputs, MM-RewardRAG enhances the factuality and reliability of LVLMs, rendering their outputs more trustworthy. Another ethical aspect of MM-RewardRAG is its efficiency. Requiring only a final reward signal for supervision makes the system particularly applicable to resource-limited scenarios, thereby promoting broader and more equitable access to these more reliable and safer LVLM capabilities.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. *Preprint*, arXiv:2404.15406.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. *Preprint*, arXiv:2109.00590.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? *Preprint*, arXiv:2302.11713.

Yun-Wei Chu, Kai Zhang, Christopher Malon, and Martin Renqiang Min. 2025. Reducing hallucinations of medical multimodal large language models with visual retrieval-augmented generation. *Preprint*, arXiv:2502.15040.

Wanqing Cui, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. More: Multi-modal retrieval augmented generative commonsense reasoning. *Preprint*, arXiv:2402.13625.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. *Preprint*, arXiv:2302.11154.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021. Weakly-supervised visual-retriever-reader for knowledge-based question answering. *Preprint*, arXiv:2109.04014.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. *Preprint*, arXiv:1906.00067.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. 2023. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. *Preprint*, arXiv:2306.09224.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. Dinov2: Learning robust visual features without supervision. *Preprint*, arXiv:2304.07193.

Jingyuan Qi, Zhiyang Xu, Rulin Shao, Yang Chen, Jin Di, Yu Cheng, Qifan Wang, and Lifu Huang. 2024. Rora-vlm: Robust retrieval-augmented vision language models. *Preprint*, arXiv:2410.08876.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *Preprint*, arXiv:2206.01718.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449. ACM.

Jiashuo Sun, Jihai Zhang, Yucheng Zhou, Zhaochen Su, Xiaoye Qu, and Yu Cheng. 2024. SURf: Teaching large vision-language models to selectively utilize retrieved information. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7611–7629, Miami, Florida, USA. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *Preprint*, arXiv:2104.06039.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, and 6 others. 2024. Emu3: Next-token prediction is all you need. *Preprint*, arXiv:2409.18869.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024a. Mmed-rag: Versatile multimodal rag system for medical vision language models. *Preprint*, arXiv:2410.13085.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024b. Rule: Reliable multimodal rag for factuality in medical vision language models. *Preprint*, arXiv:2407.05131.

Yibin Yan and Weidi Xie. 2024. Echosight: Advancing visual-language models with wiki knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 1538–1551. Association for Computational Linguistics.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. *Preprint*, arXiv:2310.00754.

A Retrieval Performance

We provide a detailed analysis of retrieval performance in this section, explaining why the query translator and adaptive reranker are necessary for MM-RAG, which operate in the pre-retrieval and post-retrieval stages, respectively. As shown in Tables 15, 16, 17, and 18, the optimal model and retrieval strategies (e.g., img2img or text2img) vary across different benchmarks, likely due to distribution shifts and differences in data architecture. Additionally, directly using a single input question or image as a query leads to poor performance, as this approach ignores the complementary nature of multimodal questions. Despite advancements in unsupervised learning methods for retrieval, some neural network-based approaches still fall behind the sparse BM25 algorithm. Therefore, it is necessary to fuse multimodal information before the retrieval phase and utilize heterogeneous knowledge bases equipped with different retrieval models to obtain targeted retrieval results. After retrieval, documents that can lead to the correct final answer are not limited to just the annotated ground truth, as human annotations are often imperfect and insufficient. In practice, useful documents frequently extend beyond labeled ground truth, as shown in Table 13. The suboptimal retriever performance makes it inadvisable to provide all retrieved content directly to the answer generator. Therefore, the adaptive reranker plays a crucial post-retrieval role in refining retrieved documents. Even in the worst-case scenario where all retrieved documents are filtered out, the system would simply revert to a vanilla multimodal QA task, which is still superior to a MM-RAG system contaminated with misleading noisy content.

B Experimental Settings

B.1 Datasets

Evaluation Benchmarks. Previous MM-RAG methods have primarily focused on benchmarks requiring only textual knowledge, which overlooks realistic scenarios where essential knowledge exists across both textual and visual modalities, necessitating cross-modality joint understanding and reasoning. To bridge this gap, we evaluate our approach using six datasets comprehensively covering scenarios that better reflect real-world multimodal information needs. (1) Infoseek (Chen et al., 2023), which focuses on information-seeking visual questions that cannot be answered directly through common sense knowledge; (2) Encyclopedia-VQA (Mensink et al., 2023), containing visual questions about detailed properties of fine-grained categories and instances requiring Wikipedia knowledge (hereafter referred to as E-VQA); (3) MultimodalQA (Talmor et al., 2021) and (4) WebQA (Chang et al., 2022), which include questions necessitating reasoning across visual and textual knowledge; (5) OK-VQA (Marino et al., 2019) along with its augmented successor (6) A-OKVQA (Schwenk et al., 2022), both containing visual questions requiring outside knowledge to answer.

Metrics. We evaluate our system using the benchmark-specific metrics: Accuracy, F1 score, Fluency, Exact Match, and BARTScore for the answer generator; recall@$\{1, 3, 5, 10, 20, 50, 100\}$ for the retriever to comprehensively assess result distributions; and standard ranking metrics NDCG, MAP, MRR, and P@1 for the reranker performance.

Knowledge Base. We utilize dataset-provided multimodal knowledge sources for WebQA and MultimodalQA, including both distractor and full-wiki settings. For E-VQA, we employ WIT (Srinivasan et al., 2021), which contains 2M Wikipedia pages consisting of free-form text and images. For Infoseek, we use OVEN (Hu et al., 2023), which includes 6M Wikipedia information entries. We also use the filtered knowledge corpus provided by Echosight and Reflectiva for fair comparison. Since OK-VQA and A-OKVQA do not provide dedicated knowledge sources, we employ the same knowledge base as used for Infoseek, and use GS112k (Luo et al., 2021) to study the transfer capabilities of optimized MM-RAG systems across different knowledge bases, following previous works.

B.2 Implementation Details

Retrieval. `CLIP-ViT-Large`, `EVA-CLIP-8B`, and `Jina-CLIP-v2` are employed for cross-modality retrieval. While the first two models are constrained by a context window of 77 tokens, `Jina-CLIP-v2` extends this capacity to 1024 tokens and incorporates additional optimizations for text-to-text retrieval. For image-to-image retrieval, we utilize the vision encoders from these models to extract image features. `BGE` dense embedding model and `BM25` sparse algorithms are used for text-to-text retrieval.

Index. All embeddings are precomputed in advance to enhance computational efficiency. FAISS-GPU is leveraged for index construction, specifically implementing IndexFlatIP for exhaustive vector search operations.

Backbone. The backbone of the query translator and adaptive reranker is initialized with Qwen-2-VL 3B, allowing for fair comparisons against other models with comparable parameter counts. Qwen-2-VL 7B/3B is utilized for the answer generator without any finetuning.

C Ablation Study

Backbone. We find that using alternative LVLMs as backbones for answer generators, including those not involved in the training process, also yields improvements across these benchmarks, highlighting the practical value of our approach in being compatible with existing models for inference without requiring additional training for adaptation.

Cross-KB Transferability. Knowledge bases in real-world scenarios inevitably require temporal updates to maintain currency and timeliness, which requires MM-RAG systems to seamlessly incorporate newly available information. Our learned query translator and adaptive reranker can effectively operate with different embedded corpora, simulating the extreme scenario of complete knowledge replacement during practical updates. This confirms that our framework ensures knowledge bases can be continuously updated to maintain current information, which is critical for long-term deployment in environments where knowledge rapidly evolves.

Computational Efficiency Our approach requires significantly fewer training data and computational resources due to its efficient design. During the training or adapting to new domains, the embedding models do not require retraining, thus eliminating the need for costly frequent reindexing, a critical and resource-intensive phase in traditional RAG systems. Instead, we insert the tunable query translator and adaptive reranker at the pre-retrieval and post-retrieval stages, respectively, allowing for targeted optimization without disturbing the core indexing and retrieval infrastructure.

D Evaluation Details.

D.1 Inference Parameters.

The query translator, adaptive reranker, and answer generator are all served in an OpenAI-compatible format, using vllm. To make sure the results can be reproduced, the temperature is set to 0, and top-p is set to 1, max length is 2048. To accelerate the evaluation process, eight instances are served at the same time across multiple servers. Power consumption is calculated through nvml. The models that support flash-attn are all enabled to accelerate inference speed.

D.2 Hard Negatives Construction.

Hard negatives for the MMQA and WebQA datasets are sourced directly from their respective original datasets. For Infoseek and EVQA, hard negatives are derived from the *top-10* results of retrieval results that separately use questions, images, and ground truth documents as query inputs. To simulate demanding real-world conditions where strong retrievers are employed, and thus generate negatives that are genuinely difficult to discriminate, BGE, Eva-CLIP, and CLIP-Large are leveraged as our retriever models to construct these hard-negative samples.

E Training Details

E.1 Component Initialization.

We leverage the visual instruction following capabilities to initialize the query translator and adaptive reranker components by designing prompt templates that guide the model to produce responses in the expected format. For the query translator, we prompt the model to generate queries in a strict JSON format, which can be easily parsed and tailored to different knowledge bases. For the Adaptive Reranker, we instruct the model to perform Chain-of-Thought (CoT) (Wei et al., 2023) reasoning first to improve interpretation, then output a usefulness level from predefined options. Note that the initialization phase only sets up the output style of these components, while their critical policies still need to be activated and aligned through the following algorithms.

E.2 Hyperparameters

We use LoRA to train these models to preserve their original general visual instruction capabilities. The hyperparameters for training these models are

Parameter	Value
SFT Stage Parameters	
Learning Rate	2×10^{-5}
Batch Size (per device)	16
Number of Epochs	3
Optimizer	AdamW
Weight Decay	0.01
Warmup Steps	500
LR Scheduler Type	Cosine
Max Sequence Length	2048
Max Gradient Norm	1.0
Dataset Size	4k
LoRA Rank (r)	8
LoRA Alpha (α)	16
LoRA Trainable Modules	q_proj, v_proj
DPO Stage Parameters	
Learning Rate	1×10^{-6}
Batch Size (per device)	8
Number of Epochs	1
Optimizer	AdamW
Weight Decay	0.01
Warmup Steps	100
LR Scheduler Type	Constant
Max Gradient Norm	1.0
Max Sequence Length	2048
Dataset Size	4k
DPO β	0.1
Label Smoothing	0.0
Loss Type	Sigmoid

Table 8: Hyperparameters for SFT and RL Training Stages of MM-RewardRAG Components.

presented in Table 8. All models are trained on a server with 8 H200 GPUs.

E.3 Training Datasets

To train our inserted components, we construct a dataset by sampling 1,000 question-answer pairs (along with any associated images) from each of the E-VQA, Infoseek, MultimodalQA, and WebQA benchmarks, yielding a total of 4,000 examples. The supervision for optimizing these components is provided exclusively by an external verifiable reward model. This reward model generates a scalar reward for each system output by comparing it against the corresponding ground-truth answer, reflecting aspects such as accuracy and factuality. This reward then serves as the basis for the learning signals propagated throughout our MM-RewardRAG framework.

F SFT Limitations

This section provides the answer to the question: *Why can SFT not be solely used to optimize the query translator and adaptive reranker?*, to discriminate the contributions of our proposed end-to-end optimization paradigm and the two lightweight components designed.

The query translator aims to generate queries to retrieve relevant knowledge from diverse knowledge bases. A primary challenge for SFT here is the absence of available ground truth for what constitutes an *optimal* translated query. While human labeling could theoretically produce training datasets (i.e., pairs of original queries and ideal translated queries), determining the "best" translation is non-trivial, even for humans. The effectiveness of a translated query can often only be assessed after executing a search with the retriever and evaluating the results again and again. This inherent characteristic suggests that the optimization of the query translator is more aptly modeled as a reinforcement learning problem. In such a framework, the model iteratively refines its query generation strategy (action) by interacting with the retriever and corpus (environment) and analyzing the retrieval performance (reward).

For the adaptive reranker, SFT faces limitations even though a ground truth document is typically provided for each question. As demonstrated in Table 13, documents capable of leading to the correct answer are often not restricted to this single ground truth instance. Consequently, an SFT approach that narrowly defines only the provided ground truth as positive and other retrieved candidates as negative can introduce significant training noise. This occurs when other genuinely useful documents, which could also lead to the correct answer, are incorrectly labeled and penalized as negatives during the SFT process.

G Detailed Discussion with Related Works

Before the era of large pre-trained models, early MM-RAG systems aimed to jointly train a generation module for final answers, a knowledge encoder, and an embedding model. Common methods included using MIPS for optimizing knowledge retrieval and employing momentum encoders for updating embeddings. However, these early systems faced several limitations. They were typically pre-trained from scratch, resulting in models with significantly smaller parameter counts than contemporary LVLMs. Furthermore, the joint training of the generation and knowledge encoding components often led to tight coupling between them, making it difficult to independently upgrade or replace modules, such as integrating more advanced, separately developed retriever models. Additionally, these approaches required loading all model parameters and knowledge base embeddings into memory for training, necessitating substantial computing resources. In contrast, our MM-RewardRAG paradigm, while aiming for end-to-end optimization, is designed to leverage the power of LVLMs. It enables end-to-end optimization by strategically tuning only lightweight, newly inserted components via reward backpropagation, which preserves the general capabilities of the foundational models, reduces the need for extensive resources for pre-training from scratch, mitigates hardware demands for fine-tuning, and offers greater modularity for component updates.

H Qualitative Examples

Figure 5 presents qualitative results underscoring the necessity of a heterogeneous knowledge base for MM-RAG. Information pertinent to queries often spans multiple modalities and is frequently modality-specific. Consequently, reliance on a unimodal knowledge base inherently leads to an incomplete representation of the required knowledge.

Illustrative outputs from the RL process are presented in Table 9 and Table 10. These examples demonstrate that queries achieving the highest and lowest reward scores often exhibit only minor lexical differences, yet yield substantially divergent retrieval results. The sensitivity to subtle query variations underscores the impracticality of manually annotating translated queries for these knowledge bases, given the prevalence of nuanced, system-specific biases. Consequently, employing RL to train a model capable of adapting to these knowledge bases by discerning their implicit biases through exploration presents a resource-efficient and, indeed, necessary strategy.

Tables 11 and 12 show that, during the RL process, the output of the adaptive reranker gradually aligns with the preferences of the answer generator, exhibiting subtle refinements in semantics and logical flow. The subtle nature of these refinements makes them difficult for humans to directly identify and distinguish. This, in turn, complicates the direct collection of labeled datasets required for supervised tuning. Consequently, after training, the adaptive reranker becomes closely coupled with the answer generator to maximize this implicit reward.

I Prompt Tuning vs. Parameter Tuning

While directly fine-tuning the parameters of the answer generator might be an intuitive solution, we found this approach suboptimal for two main reasons: (i) intensive parameter tuning on specific datasets can degrade its broad, pre-trained visual instruction-following capabilities, potentially harming its generalization. (ii) More importantly, with our proposed query translator and adaptive reranker, which effectively filter noisy data and dynamically determine the optimal amount of contextual information passed to the generator, we find that lightweight prompt engineering, primarily just to define input and output formats, is sufficient for achieving high performance. Consequently, the additional advantages of full parameter fine-tuning for the generator become minimal in our MM-RewardRAG framework.

J Prompt Templates

We provide the prompts used in this paper in Figure 6, 8, 7, 9.

Example: Dirt Track Racing Driver

Highest Scoring Output Details

Chain of Thought: The image shows a dirt track racing car. The question asks about a New Zealander who raced on this track in 2008. We need to identify the track and find out if any New Zealanders raced there in 2008.

Generated Queries:
- "New Zealand dirt track racing 2008"
- "2008 New Zealand dirt track racing driver"

Keywords String: New Zealand, dirt track, racing, 2008, driver

Lowest Scoring Output Details

Chain of Thought: The image shows a race car on a dirt track, suggesting it might be related to motorsport, possibly involving New Zealand racing history. The question asks about a New Zealander who raced at this track in 2008. The first step is to identify the race track, which could be inferred from the context or the track's distinctive features. If the track is not recognizable, the focus shifts to identifying any specific races or events in 2008 where New Zealand drivers participated. This could involve searching for notable New Zealand racing events from that year.

Generated Queries:
- "New Zealand motorsport history 2008"
- "New Zealand race cars 2008"

Keywords String: New Zealand motorsport 2008 race track

Table 9: Qualitative example for *Dirt Track Racing Driver* case: generated queries and reasoning during RL training.

Example: Plant Synonym from Seed Pod

Highest Scoring Output Details

Chain of Thought: The question asks for a synonym of the plant shown in the image. To find this information, I need to identify the plant species first. The image shows a dry seed pod with a distinctive shape. This could be a type of plant like a milkweed or a similar species known for its seed pods. Once the plant species is identified, I can search for synonyms related to that species.

Generated Queries:
- "dry seed pod plant"
- "plant with dry seed pod synonyms"
- "common names for plants with dry seed pods"

Keywords String: dry seed pod, plant, synonyms, common names

Lowest Scoring Output Details

Chain of Thought: The question asks for a synonym of the plant shown in the image. The first step is to identify the plant in the image. Once identified, synonyms can be searched.

Generated Queries:
- "What is a synonym for the plant in the image?"
- "Identify the plant in the image and find its synonyms."

Keywords String: plant, synonyms

Table 10: Qualitative example for *Plant Synonym from Seed Pod* case: generated queries and reasoning during RL training.

Transition: useful → useless

Positive Case (Predicted Label: useful)

Chain of Thought: The knowledge snippet provides information about Monarda didyma, commonly known as crimson beebalm. It mentions that this plant grows in dense clusters along stream banks, moist thickets, and ditches. This directly answers the VQA question about what kind of clusters the plant grows in.

Negative Case (Predicted Label: useless)

Chain of Thought: The knowledge snippet provides information about Monarda didyma, commonly known as crimson beebalm or scarlet monarda. It describes various aspects of the plant, such as its family, habitat, and characteristics. However, it does not explicitly mention the type of clusters in which the plant grows. Therefore, the knowledge snippet is not directly relevant to answering the VQA question about the plant's cluster growth pattern.

Transition: useful → neutral

Positive Case (Predicted Label: useful)

Chain of Thought: The knowledge snippet provides information about Monarda didyma, commonly known as crimson beebalm. It mentions that the plant grows in dense clusters along stream banks, moist thickets, and ditches. This matches the VQA question which asks about the type of clusters the plant grows in. Therefore, the knowledge snippet is relevant and useful for answering the question.

Negative Case (Predicted Label: neutral)

Chain of Thought: The knowledge snippet provides information about the plant Monarda didyma, commonly known as crimson beebalm. It describes various aspects of the plant such as its appearance, growth habits, and historical uses. However, it does not specifically mention how the plant grows in clusters or provide details about its cluster formation.

Table 11: Qualitative examples of textual knowledge for adaptive reranker.

Transition: useful → neutral

Positive Case (Predicted Label: useful)

Chain of Thought: The image confirms that the city in question is Tokyo, which hosted events in the BWF Super Series in 2017 and has a tall, narrow tower (Tokyo Skytree) illuminated in blue light. This visual evidence supports the question's requirements, making the image useful for answering the population-related part.

Negative Case (Predicted Label: neutral)

Chain of Thought: The question asks about the population of a city that hosted a scheduled event in 2017's BWF Super Series and has a tall, narrow tower illuminated in blue light. The image shows Tokyo, which is known to host such events and has a distinctive tall, narrow tower (the Tokyo Skytree) that is illuminated in blue at night. However, the image does not provide information about the population of Tokyo's metropolitan area.

Transition: neutral → useless

Positive Case (Predicted Label: neutral)

Chain of Thought: The image confirms that the city in question is Tokyo, which fits the description of having a tall, narrow tower (Tokyo Skytree) lit up in blue light. However, the image does not provide any data or indicators regarding the population of the metropolitan area.

Negative Case (Predicted Label: useless)

Chain of Thought: The image shows Tokyo and the Tokyo Skytree illuminated in blue, but it does not confirm that Tokyo hosted the 2017 BWF Super Series events or provide data on the population of its metropolitan area.

Table 12: Qualitative examples of visual knowledge for adaptive rerankers.

Dataset	Pseudo Recall						Recall					
	@1	@5	@10	@20	@50	@100	@1	@5	@10	@20	@50	@100
OKVQA	0.421	0.676	0.781	0.866	0.941	0.969	0.136	0.303	0.392	0.493	0.635	0.718
EVQA	0.511	0.719	0.785	0.833	0.883	0.909	0.402	0.624	0.702	0.780	0.861	0.912
Infoseek	0.311	0.575	0.686	0.780	0.874	0.916	0.190	0.392	0.487	0.584	0.713	0.797

Table 13: PReFLMR Recall Performance on Different Datasets.

Dataset	Model	NDCG				Precision				MAP	MRR
		Overall	@3	@5	@10	@1	@3	@5	@10		
EVQA	Mono	–	–	–	–	–	–	–	–	–	–
	DES	60.29	43.42	47.03	57.54	34.10	18.33	13.36	10.51	48.61	47.99
	Jina$_{img}$	65.02	56.90	62.50	65.36	41.90	25.30	10.00	11.46	59.39	58.95
	Jina$_{text}$	86.16	85.99	87.41	86.42	76.80	34.53	22.36	11.54	85.56	85.33
	Ours	97.34	99.69	98.83	97.51	100.0	34.23	21.42	11.46	97.03	100.0
Infoseek	Mono	–	–	–	–	–	–	–	–	–	–
	DES	47.36	20.34	21.96	35.98	14.80	8.20	5.90	6.86	27.42	27.42
	Jina$_{img}$	77.20	74.52	76.39	76.64	63.10	27.76	18.08	9.75	74.76	74.76
	Jina$_{text}$	71.76	68.26	70.56	71.89	54.30	26.33	17.54	9.90	68.97	68.97
	Ours	100.0	100.0	100.0	100.0	100.0	33.33	20.00	10.00	100.0	100.0
MMQA	Text-to-Image										
	Mono	82.81	79.89	81.28	81.72	71.73	30.15	21.01	11.34	79.48	80.23
	DES	80.33	77.87	80.18	80.13	68.47	31.32	21.28	11.59	77.42	78.60
	Jina	87.21	85.70	86.33	85.79	80.16	33.33	22.01	11.51	85.04	86.14
	Ours	92.00	93.05	93.36	91.87	86.68	37.24	23.75	12.12	92.11	92.50
	Text-to-Text										
	Mono	-	-	-	-	-	-	-	-	-	-
	DES	75.47	76.99	78.38	75.46	68.27	37.36	26.03	15.32	71.46	78.60
	Jina	90.66	91.93	91.82	90.67	87.62	47.70	30.01	15.39	90.47	92.22
	Ours	95.70	96.24	96.37	95.70	93.74	49.54	30.64	15.39	95.71	96.20
WebQA	Text-to-Image										
	Mono	73.51	74.74	76.40	75.37	61.47	33.26	22.63	13.08	70.99	74.69
	DES	68.74	68.83	71.01	71.03	55.28	30.80	21.87	13.10	65.45	69.71
	Jina	79.02	81.54	81.81	80.96	70.45	36.92	24.59	13.40	78.07	81.16
	Ours	90.87	93.01	92.89	91.66	85.62	44.17	27.46	13.88	91.22	91.79
	Text-to-Text										
	Mono	-	-	-	-	-	-	-	-	-	-
	DES	76.05	84.33	82.26	78.35	77.55	47.09	33.06	18.94	72.83	85.44
	Jina	89.65	30.00	30.00	21.54	30.00	16.66	20.00	5.00	88.09	30.00
	Ours	100.0	100.0	100.0	100.0	100.0	66.66	40.00	20.00	100.0	100.0

Table 14: Performance Comparison of Reranker Models Across Multiple Datasets.

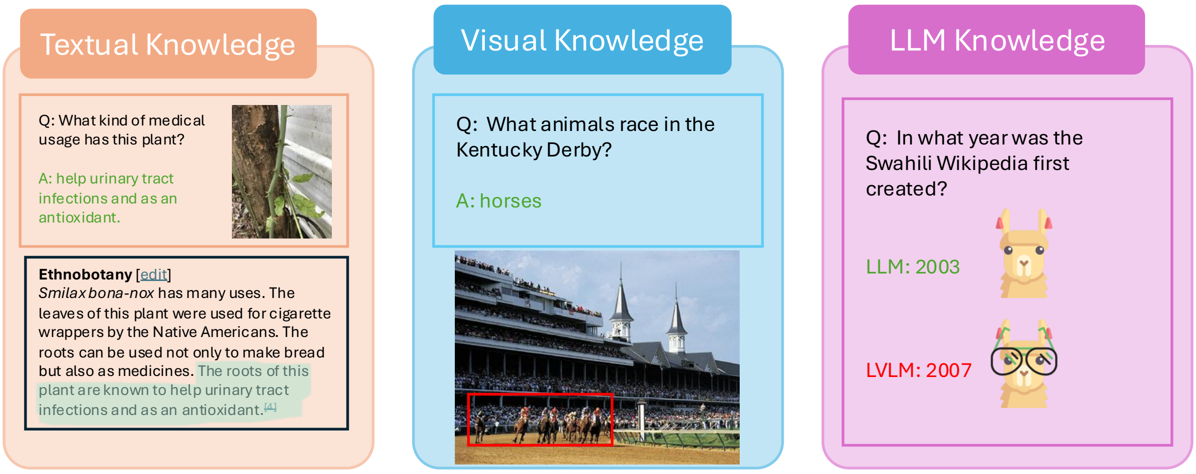

Figure 5: Different types of knowledge that are required to answer questions.

Prompt For Verifiable Reward Model

```
Given a question, a ground truth answer, and a prediction answer, please
evaluate the prediction answer.
If the prediction answer is correct, please return "True".
If the prediction answer is wrong, please return "False".
The question is: {{ {question}} }}
The ground truth answer is: {{ {ground_truth_answer} }}
The prediction answer is: {{ {prediction_answer} }}
Please answer with "True" or "False".
```

Figure 6: Prompt for Verifiable Reward Model.

Model	**Modality**	**Recall**						
		@1	@3	@5	@10	@20	@50	@100
Jina-CLIP-V2	$\mathcal{I} \to \mathcal{I}$	10.99	17.46	20.62	24.81	31.35	38.66	44.01
	$\mathcal{I} \to \mathcal{T}_{\text{part}}$	2.31	5.09	6.45	10.43	15.18	22.50	28.13
	$\mathcal{I} \to \mathcal{T}_{\text{whole}}$	4.43	8.33	10.06	14.67	19.49	27.96	34.88
	$\mathcal{T} \to \mathcal{I}$	0.27	0.43	0.60	1.06	1.53	2.85	4.10
	$\mathcal{T} \to \mathcal{T}_{\text{part}}$	3.75	5.42	6.71	8.31	9.98	13.00	15.84
	$\mathcal{T} \to \mathcal{T}_{\text{whole}}$	1.89	2.64	3.28	4.26	5.49	7.93	10.38
CLIP-Large	$\mathcal{I} \to \mathcal{I}$	10.34	17.72	21.32	26.69	33.37	40.60	47.66
	$\mathcal{I} \to \mathcal{I}_{\text{title}}$	21.85	23.89	32.45	35.84	39.02	43.23	45.09
	$\mathcal{I} \to \mathcal{I}_{\text{summary}}$	15.23	19.96	31.89	35.55	39.08	42.12	45.79
EVA-CLIP-8B	$\mathcal{I} \to \mathcal{I}$	15.28	25.09	30.20	37.28	43.86	50.97	55.66
	$\mathcal{I} \to \mathcal{I}_{\text{title}}$	17.29	30.01	33.22	35.13	41.12	45.09	49.82
	$\mathcal{I} \to \mathcal{I}_{\text{summary}}$	20.83	33.72	37.56	42.30	43.58	47.26	50.39

Table 15: Recall results on E-VQA dataset

Prompt For Answer Generator

```
Given a Visual Question Answering (VQA) question and a knowledge snippet,
please generate the answer to the question.
Here is the VQA question:
<img_start><img><img_end>
Question: {question}
Here is the knowledge snippet: {document}
Please output the answer to the question.
```

(a) Prompt For Answer Generator (VQA with Knowledge Snippet).

Prompt For Answer Generator

```
Given a question and a knowledge snippet, please generate the answer to the
question.
Here is the question:
Question: {question}
Here is the knowledge snippet: {document}
Please output the answer to the question.
```

(b) Prompt For Answer Generator (Text Question with Knowledge Snippet).

Prompt For Answer Generator

```
Given a question and an image with a caption, please generate the answer to
the question.
Here is the question:
Question: {question}
Here is the image with caption:
<img_start><img><img_end>
Caption: {caption}
Please output the answer to the question.
```

(c) Prompt For Answer Generator (Question with Image and Caption).

Figure 7: Examples of different prompts for the Answer Generator module.

Prompt For Query Translator

```
Given a Visual Question Answering (VQA) question, please generate some
possible search engine queries and keywords that can be used to retrieve
knowledge from an external knowledge base that can answer the question
without referring to the input image.
Please output strict JSON that can be directly parsed by Python, in the
following format: {"Chain_of_Thought": your analysis here, "queries": search
engine queries here, "image_queries": image search engine queries here, "
key_words_string": keywords for BM25 here.}.
Here is the VQA question:
<img_start><img_token><img_end>
Question: {question}
Please output the Chain of Thought reasoning and the queries in strict JSON
format.
```

Figure 8: Prompt For Query Translator.

Prompt For Adaptive Reranker

```
Given a Visual Question Answering (VQA) question and a knowledge snippet,
please determine whether this knowledge snippet is useful for answering the
VQA question. Please output one of the following levels: useful, neutral,
useless. First, perform Chain of Thought (CoT) reasoning and then output the
label. Please output strict JSON that can be directly parsed by Python, in
the following format: {"CoT": your analysis here, "level": the level obtained
 after analysis}.
Here is the VQA question:
<img_start><img_token><img_end>
Question: {question}
Here is the knowledge snippet: {document}
Please output the Chain of Thought reasoning and the label in strict JSON
format.
```

Figure 9: Prompt For Adaptive Reranker.

Model	**Modality**	**Recall**						
		@1	@3	@5	@10	@20	@50	@100
Jina-CLIP-V2	$\mathcal{I} \rightarrow \mathcal{I}$	2.76	4.57	5.47	6.93	8.79	11.86	14.87
	$\mathcal{I} \rightarrow \mathcal{T}$	9.9	16.82	20.67	26.34	32.64	41.49	48.00
	$\mathcal{T} \rightarrow \mathcal{I}$	0	0	0	0	0.01	0.01	0.09
	$\mathcal{T} \rightarrow \mathcal{T}$	0	0	0	0.04	0.15	0.30	0.67
CLIP-Large	$\mathcal{I} \rightarrow \mathcal{I}$	10.81	16.22	19.78	23.89	27.96	33.52	37.36
	$\mathcal{I} \rightarrow \mathcal{I}_{\text{title}}$	9.72	13.22	17.08	21.09	22.78	25.32	29.85
	$\mathcal{I} \rightarrow \mathcal{I}_{\text{summary}}$	9.83	15.72	17.99	25.37	28.90	31.85	33.05
Eva-CLIP-8b	$\mathcal{I} \rightarrow \mathcal{I}$	16.00	23.21	26.66	30.29	33.60	38.17	41.42
	$\mathcal{I} \rightarrow \mathcal{I}_{\text{title}}$	21.23	26.79	33.95	35.23	40.26	43.23	47.58
	$\mathcal{I} \rightarrow \mathcal{I}_{\text{summary}}$	20.01	27.83	35.23	37.21	39.89	42.08	43.72

Table 16: Recall results on Infoseek dataset

Model	Modality	Recall						
		@1	@3	@5	@10	@20	@50	@100
Jina-CLIP-V2	$\mathcal{T} \to \mathcal{I}$	12.40	24.80	29.60	33.20	37.60	44.00	50.40
	$\mathcal{T} \to \mathcal{I}_{\text{title}}$	82.80	88.00	89.20	90.40	92.00	93.60	94.40
	$\mathcal{T} \to \mathcal{T}$	39.67	58.29	63.80	71.08	77.66	83.99	87.22
BM25	$\mathcal{T} \to \mathcal{I}_{title}$	87.60	90.40	90.80	91.20	91.20	92.00	92.00
	$\mathcal{T} \to \mathcal{T}$	36.08	60.06	67.47	76.01	81.52	86.84	90.06
BGE	$\mathcal{T} \to \mathcal{I}_{title}$	86.13	90.53	91.13	91.93	92.93	93.53	94.33
	$\mathcal{T} \to \mathcal{T}$	45.76	70.76	77.28	83.19	86.37	89.96	91.43

Table 17: Recall results on MMQA dataset

Model	Modality	Recall						
		@1	@3	@5	@10	@20	@50	@100
Jina-CLIP-V2	$\mathcal{T} \to \mathcal{I}$	65.20	80.00	91.60	99.20	1	-	-
	$\mathcal{T} \to \mathcal{I}_{\text{title}}$	93.60	97.20	98.80	1	-	-	-
	$\mathcal{T} \to \mathcal{T}$	53.73	77.15	86.84	1	-	-	-
BM25	$\mathcal{T} \to \mathcal{I}_{\text{title}}$	96.80	99.60	1	-	-	-	-
	$\mathcal{T} \to \mathcal{T}$	48.42	85.06	92.59	1	-	-	-
BGE	$\mathcal{T} \to \mathcal{I}_{\text{title}}$	96.40	99.20	99.60	1	-	-	-
	$\mathcal{T} \to \mathcal{T}$	61.65	89.56	95.13	1	-	-	-

Table 18: Recall results on WebQA dataset