

RGCC: Retriever-Guided Context Compression for Efficient Vision–Language Alignment in Knowledge-Based VQA

Qiankun Li^{1,2}, Feifei Zhai^{1*}, Jianghao Chen^{1,2}, Chong Li^{1,2},
Haoran Gao³, Fanyu Meng³, Junlan Feng³, Yang Zhao^{1,2}, Yu Zhou^{1,4}

¹ State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS),
Institute of Automation, Chinese Academy of Sciences, Beijing, China

² School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

³ JIUTIAN Research, China Mobile ⁴ Fanyu AI Laboratory, Zhongke Fanyu Technology Co., Ltd, Beijing, China
{liqiankun2023, feifei.zhai, chenjianghao2022, lichong2021}@ia.ac.cn

{gaohaoranit, mengfanyuit01, fengjunlanit}@chinamobile.com, {yang.zhao, yzhou}@nlpr.ia.ac.cn

Abstract

Knowledge-Based Visual Question Answering (KB-VQA) requires retrieving external textual knowledge to answer visual queries. However, current methods often suffer from two critical misalignments during retrieval: distributional misalignment, driven by the severe stylistic shift from descriptive pre-training image captions to formal knowledge texts, and granularity misalignment, where lengthy knowledge documents contain excessive noise that drowns out key information, making it difficult to accurately match the visual query. To resolve these challenges, we propose RGCC, a Retriever-Guided Context Compression framework. By mining alignment-friendly data and further training a context compression model, RGCC condenses lengthy documents into compact, retriever-preferred representations that significantly improve cross-modal retrieval. Extensive experiments on the InfoSeek and E-VQA benchmarks demonstrate that RGCC achieves state-of-the-art retrieval performance and boosts downstream question-answering accuracy without additional online inference overhead.¹

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) (Fayyaz et al., 2025; Achiam et al., 2023; Team et al., 2023, 2024) have significantly enhanced their performance in visual reasoning (Tang et al., 2025) and generation tasks, including image captioning (Abdulgalil and Basir, 2025) and visual question answering (VQA) (Antol et al., 2015; Goyal et al., 2017), etc. Nevertheless, these models continue to face challenges in Knowledge-Based Visual Question Answering (KB-VQA) (Lin et al., 2024), due to the absence of long-tail knowledge.

* Corresponding author.

¹Our code is available at: <https://github.com/qiankunlee/RGCC>

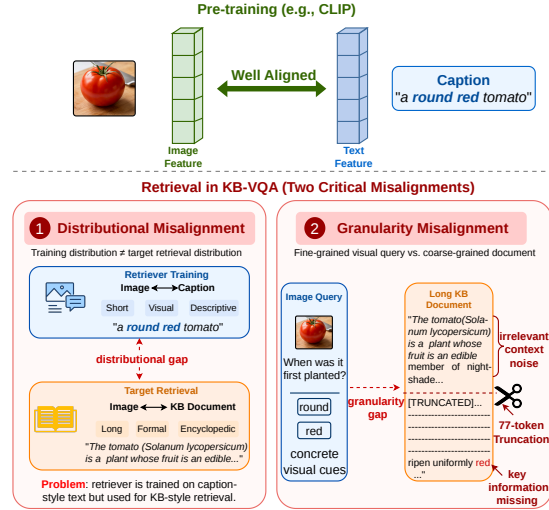


Figure 1: Illustration of the alignment gaps between image-caption pre-training and KB-VQA retrieval, including distributional misalignment and granularity misalignment.

MultiModal Retrieval-Augmented Generation (MMRAG) (Chen et al., 2025) has emerged as an effective approach to enhance MLLMs by incorporating query-relevant external knowledge. MMRAG typically follows a three-stage pipeline: retrieval, reranking, and generation; detailed steps are in Section 3.1. The retrieval stage establishes the informational foundation for subsequent re-ranking and generation. Its quality directly determines the range and relevance of content that can be effectively leveraged. Therefore, the precision of the retrieval component has a significant impact on the performance of MMRAG.

Although existing methods have achieved advancements on related tasks, several significant gaps still remain:

(1) **Distributional Misalignment.** Discrepancies in language distribution between the retrieval training data (e.g., concise image-caption pairs) and the target retrieval corpus (e.g., long-form en-

cyclopedic texts covering diverse long-tail entities) create a severe domain shift. This shift impedes effective alignment in the shared cross-modal embedding space, causing key visual cues to fail to find corresponding textual descriptions (Qiu et al., 2022), as illustrated in Figure 1.

(2) **Granularity Misalignment.** Retrieval content often contains lengthy encyclopedic descriptions, introducing abundant abstract semantic information that misaligns with the concrete visual cues of the query object. For example, a passage describing a tomato might read: “The tomato (*Solanum lycopersicum*) is a plant whose fruit is an edible berry... member of the nightshade family...”, which can dilute the salient local visual features such as color, shape, and texture. Moreover, text input limits in retrievers (e.g., 77 tokens in CLIP) may truncate visually relevant details while retaining irrelevant content, further exacerbating this misalignment.

To address the challenges outlined above, we propose RGCC (Retriever-Guided Context Compression), a novel framework that explicitly enhances vision–language alignment. First, we introduce a retrievability-driven data screening strategy to mine alignment-friendly yet challenging image–entity pairs from large-scale datasets. By selecting samples that are not trivially retrieved but remain accessible to the target retriever, this strategy constructs informative supervision that better matches the retriever’s preferences, improving training efficiency and robustness under real-world retrieval scenarios. Second, RGCC employs Group Relative Policy Optimization (GRPO) with a visual-similarity reward to train the context compression model. The reward encourages the model to preserve image-relevant visual evidence while filtering out abstract encyclopedic noise, producing compact contexts that are better aligned with the retriever’s preferences.

We evaluate our method on two KB-VQA benchmarks, InfoSeek (Hu et al., 2023) and Encyclopedic VQA (E-VQA) (Van Horn et al., 2021; Weyand et al., 2020), where extensive retrieval and VQA experiments demonstrate its effectiveness. In summary, the main contributions of our work are as follows:

- We propose a retrievability-driven data screening strategy to mine alignment-friendly yet challenging image-entity pairs from large-scale datasets. This enables more efficient retriever-guided compression training and im-

proves robustness under distributional shifts.

- We introduce RGCC, a Retriever-Guided Context Compression framework that integrates GRPO with a visual-similarity reward to alleviate distributional and granularity misalignment, producing compact contexts that better align with the retriever’s preferences.
- We achieve state-of-the-art retrieval and reranking performance on KB-VQA benchmarks, and show that these gains can translate into improved downstream VQA accuracy under controlled generator settings. Importantly, our method does not introduce any additional online inference overhead.

2 Related Work

2.1 KB-VQA

Prior datasets such as OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022) focus on visual question answering tasks that require external knowledge. However, the knowledge involved in these benchmarks primarily consists of generic commonsense information, which lacks sufficient specificity and distinctiveness. The emergence of InfoSeek (Hu et al., 2023) and E-VQA (Van Horn et al., 2021; Weyand et al., 2020) addresses this limitation by requiring models to accurately identify specific species, industrial products, and other fine-grained categories. Recent works such as Wiki-LLaVA (Caffagni et al., 2024), EchoSight (Yan and Xie, 2024a), and OMGM (Yang et al., 2025) employ hierarchical retrieval strategies to facilitate efficient multimodal retrieval. LLaVA-mR²AG (Zhang et al., 2024) and ReflectiVA (Cocchi et al., 2025) focus on reflective retrieval, where large language models are fine-tuned to make decisions on whether retrieval is necessary through token-level generation. VLM-PRF (Hong et al., 2025) processes complex multimodal queries to extract precise multimodal information for retrieval, ensuring highly relevant knowledge is obtained.

Unlike previous methods, our approach improves retrieval performance by compressing long contexts while simultaneously mitigating granularity and distributional misalignment in cross-modal retrieval, without introducing any online overhead. This method not only enhances re-ranking efficiency but also improves answer generation quality. Furthermore, it has broad applicability and can be

extended to other re-ranking or knowledge selection scenarios.

2.2 Reinforcement Learning for Text Compression

Reinforcement learning (RL) (Mnih et al., 2013; Arulkumaran et al., 2017; François-Lavet et al., 2018) is a machine learning paradigm in which an agent improves its decision-making through interaction with an environment and reward feedback, and has recently been widely adopted in large language models (LLMs) (Wang et al., 2024).

Weng et al. (Weng et al., 2024) improve inference efficiency by compressing retrieved contexts at generation time. Ghalandari et al. (Ghalandari et al., 2022) employ reinforcement learning for coarse-grained demonstration selection and fine-grained token pruning, thereby compressing few-shot chain-of-thought contexts while preserving mathematical reasoning performance. Ting et al. (Ting et al., 2026) introduce an outcome-based reward mechanism to improve the utilization of compressed representations for downstream reasoning, addressing the gap between context retention and effective inference.

In contrast, our method performs compression at the retrieval stage. Specifically, we use Group Relative Policy Optimization (GRPO) to dynamically compress knowledge-base entries and directly optimize the compression policy with retrieval-based rewards. This formulation encourages compressed representations to preserve information that is most relevant to the visual query, thereby improving image-text alignment and retrieval recall. Our approach therefore differs from prior work in terms of its optimization objective, training strategy, and the stage at which compression is performed.

3 Methodology

This section begins by formalizing the task definition of MMRAG (§3.1), and subsequently introduces a retrievability-driven data screening strategy (§3.2) and retriever-guided context compression training for efficient vision-language alignment (§3.3).

3.1 Task Setup

A typical MMRAG system consists of:

- Datastore $\mathcal{D}$ containing a large collection of knowledge entities $\{E_i\}_{i=1}^{|\mathcal{D}|}$, where each en-

tity E_i is composed of a visual component V_i and an associated textual article T_i ;

- Retriever $\mathcal{R}_\theta$, parameterized by θ , which encodes the query image Q^v and candidate visual or textual contents $\{V_i, T_i\}$ into a shared embedding space. Relevance is measured via cosine similarity, $\text{Cos}(\text{En}(Q^v), \text{En}(T_i))$. $\text{En}(\cdot)$ denotes the vector encoded by the retriever, such as CLIP and Eva-CLIP.
- An optional reranker $\mathcal{M}_\theta$ is introduced to further reorder the retrieved entities by capturing fine-grained semantic representations to the query;
- A multimodal query $Q = (Q^v, Q^t)$, together with the retrieved knowledge, is fed into LLMs or MLLMs to generate the final answer.

As described above, the retrieval stage establishes the informational foundation for subsequent reranking and question answering. Its quality directly constrains the relevance and diversity of the knowledge accessible to downstream models. Therefore, the precision of the retrieval component has a significant impact on the performance of MMRAG. Knowledge search and answer generation can be found in the Multi-modal Retrieval-Augmented Generation module in Figure 2.

3.2 Retrievability-driven Data Screening Strategy

To enhance training efficiency and strengthen cross-modal retrieval, we propose a retrievability-driven data screening strategy that mines alignment-friendly yet challenging image-entity pairs from large-scale datasets. By selecting samples that are not trivially retrieved but remain accessible to the target retriever, this strategy constructs informative supervision that better matches the retriever’s preferences. Such data exposes the compression model to non-trivial retrieval cases while maintaining retriever compatibility, thereby improving the efficiency of retriever-guided compression training and enhancing robustness under real-world retrieval scenarios. Specifically, the strategy proceeds in three stages:

- **Step 1: Vector Database Construction.** For each entity, its corresponding article is first condensed into a concise summary using a large language model. These summaries are then embedded using a text encoder such as

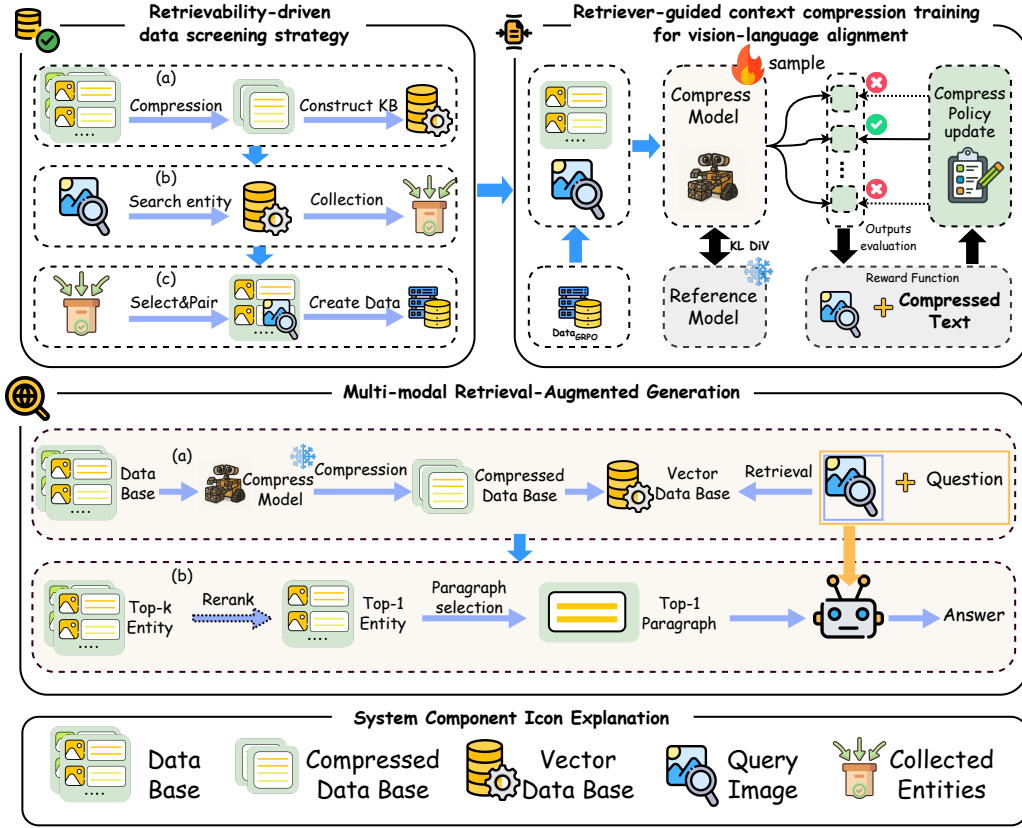


Figure 2: In the **Retrieval-driven data screening strategy** module: (a) represents the construction of vector database; (b) represents the candidate entity collection; (c) represents the data filter and construction. In the **Retriever-guided context compression training for vision-language alignment** module: in the image-entity pair, only the article part of the entity is compressed. The compressed text and the image are then used to calculate similarity using Eva-CLIP, serving as a reward signal to guide the model’s training. In the **multi-modal retrieval-augmented generation** module: (a) represents the process of vector database construction and retrieval; (b) represents reranking (an optional operation) and knowledge-based visual question answering.

Eva-CLIP, and indexed with Faiss to construct a vector database.

- **Step 2: Candidate Entity Collection.** For each query image Q^v , we retrieve the top- m candidate entities $\{\text{Entity}_i\}_{i=1}^m$ based on similarity scores. In our experiments, we set $m = 50$ for both InfoSeek and E-VQA, balancing retrieval effectiveness and computational efficiency.
- **Step 3: Data Filtering and Training Dataset Construction.** An image–golden entity pair is retained only if the golden entity appears outside the Top- k but within the Top- m results (i.e., $\{\text{Entity}_i\}_{i=k}^m$), and discarded otherwise. The value of k is determined through empirical retrieval analysis, which evaluates how different ranking ranges impact the quality and

coverage of the training data (see Section 4.4 and Appendix D.2 for details). Based on this analysis, we set $k=10$ for InfoSeek and $k=20$ for E-VQA. This strategy preserves alignment-friendly yet challenging samples, ensuring that the training dataset D_{grpo} contains the diversity and complexity of real-world retrieval scenarios while remaining aligned with the retriever’s preferences.

Building on this dataset, we further identify entities that are poorly modeled or weakly separated in the LLM representation space, thereby uncovering latent cross-modal correlations. These alignment-friendly yet challenging samples provide more informative retrieval-guided signals for subsequent compression training. This selection process lays the foundation for retriever-guided context compression, which alleviates distributional and gran-

ularity misalignment by condensing long contexts into retriever-preferred representations while preserving critical visual semantics (see Section 3.3).

3.3 Retriever-guided context compression training for vision-language alignment

To explicitly address both distributional and granularity misalignment, we propose a retriever-guided context compression framework optimized via Group Relative Policy Optimization (GRPO) (Zhang and Zuo, 2025), as illustrated in Figure 2. We adopt GRPO instead of supervised fine-tuning (SFT) because there is no unique ground-truth compressed context label, and the compression model does not directly take visual information as input during SFT training. GRPO allows us to introduce a visual-similarity reward as an external optimization signal, encouraging the model to preserve textual evidence aligned with salient visual cues while discarding abstract or encyclopedic content irrelevant to the image. A frozen reference model π_{ref} constrains policy updates, preventing the compression model from deviating excessively from its original text compression capabilities. Both the compression and reference models are initialized from the same pretrained weights, while the compression model is optimized as the target policy.

The GRPO training dataset D_{grpo} is constructed from high-quality image–entity pairs identified by a retrievability-driven data screening strategy. For each entity, only the textual article is fed into the compression model. From each input, a set of G candidate outputs $o_1, \dots, o_G$ is sampled from the current compression policy. The similarity between each candidate and its corresponding query image, computed using a retriever (e.g., Eva-CLIP), serves as the reward signal for GRPO, encouraging retention of text segments aligned with salient visual features.

Formally, the reward for a candidate compressed text o_i is defined as the cosine similarity with the query image Q_v :

$$R(o_i) = \text{Cos}(En(o_i), En(Q^v)) \quad (1)$$

where $En(*)$ denotes the embedding produced by the retriever. The advantage of each candidate is computed relative to the group mean $\bar{R}$:

$$\bar{R} = \frac{1}{G} \sum_{i=1}^G R(o_i). \quad (2)$$

Then, the advantage A_i for o_i is defined as:

$$A_i = R(o_i) - \bar{R}. \quad (3)$$

This advantage A_i is used for policy updates. The policy parameters θ are optimized by maximizing the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q, \{o_i\} \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^G (\mathcal{L}_i - \beta \mathbb{D}_{\text{KL}}(\pi_{\theta} || \pi_{\text{ref}})) \right] \quad (4)$$

$$\mathcal{L}_i = \min(r_i A_i, \text{clip}(r_i, 1 - \epsilon, 1 + \epsilon) A_i) \quad (5)$$

$$r_i = \frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, \quad (6)$$

where $\pi_{\theta_{\text{old}}}$ is the old policy, π_{θ} is the current policy being optimized, and ϵ is a clipping hyperparameter. The $\mathbb{D}_{\text{KL}}$ term represents the regularization of KL divergence, penalizing deviations of the current policy π_{θ} from a reference policy π_{ref} .

By aligning the compressed text with the visual input, the model bridges the gap between fine-grained image-level features and coarse-grained knowledge-base texts. As a result, both distributional and granularity misalignments are mitigated. In addition, the retrievability-driven data screening strategy selects alignment-friendly training samples that provide more informative supervision for subsequent compression training, enabling the model to generate compact textual representations that are visually aligned and preferred by the retriever.

4 Experiments

4.1 Datasets

In this study, two challenging KB-VQA datasets, InfoSeek and E-VQA (Yan and Xie, 2024b; Mensink et al., 2023; Chen et al., 2023), were employed to evaluate our proposed method. (1) E-VQA comprises 221K unique question–answer pairs (Q, A) spanning 16.7K fine-grained entities from iNaturalist (Van Horn et al., 2021) and Google Landmarks Dataset v2 (Weyand et al., 2020). Each question–answer pair is linked to up to five diverse instance images, resulting in a total of 1M visual–question–answer triplets (Q, V, A) . (2) InfoSeek contains 1.3M (Q, V, A) triplets associated with approximately 11K distinct entities from OVEN (Hu et al., 2023). Detailed statistics for

Datasets	Samples				#Entity Articles	
	Train	Re. Train	Com. Train	Test	Train	Valid/Test
InfoSeek	900k	-	37k	71335	100k	100k
E-VQA	900k	177k	26k	4750	2M	2M

Table 1: The statistical information of the InfoSeek and E-VQA datasets used in our experiments. Train and Test denote the numbers of samples in the training and test sets, respectively; Com. Train and Re. Train represent the numbers of samples used in the model compression stage and the re-ranking stage, respectively.

both datasets, including sample splits and knowledge base sizes, are shown in Table 1. More details on the datasets and their evaluation protocols are provided in Appendix B.

4.2 Implementation Details

In this section, we detail the main experimental procedures. The prompts used for LLM-based text compression and MLLM-based visual question answering are provided in Appendix A.

Compression Model Training. We initialize the compression model from Qwen2.5-1.5B-Instruct due to its strong semantic modeling and instruction-following capabilities. Detailed training hyperparameters are provided in Appendix C.2.

Retrieval setting. We employ the trained compression model to condense each entity’s article in the knowledge base, using the prompts summarized in Table 26. For efficient entity retrieval, a FAISS index is constructed over the compressed article representations, and image-text similarity is computed from pooled embeddings of the last encoder layer. Following prior work, we retrieve the top-20 entities ($k = 20$) for a fair comparison.

Importantly, compression is applied only at the retrieval stage as a coarse-grained mechanism for entity-level candidate selection. The retriever outputs candidate entity IDs based on compressed representations, which are then used in the reranking stage to access the full, uncompressed articles for fine-grained, section-level selection. Consequently, no original content is lost during reranking or answer generation, ensuring all answer-critical evidence remains fully accessible. Moreover, because compression is performed entirely as offline knowledge-base preprocessing, it introduces no additional online inference cost or latency to the VQA pipeline.

Reranker setting. We use BLIP-2 (Li et al., 2023) to rerank the first-stage retrieved entities (top-20) and select the most similar one as the

top-1 entity. Details of reranker training and inference scoring are provided in Appendix C.3 and Appendix C.4, respectively.

Generator Training and Inference. Given the top-1 entity, we use BGE-Reranker-v2-m3 (Multi-Granularity, 2024) to select the question-relevant top-1 section, which is then fed into the MLLM together with the question and image for answer generation. We evaluate our method under the OMGM QA setting. For E-VQA, we directly use OMGM’s fine-tuned LLaVA-1.5-7B. Since the InfoSeek model weights are not released, we fine-tune LLaVA-1.5-7B with LoRA following OMGM’s training strategy. More details are provided in Appendix C.5.

4.3 Main results

A comparative evaluation of our method against existing works is presented in Tables 2 and 3, focusing on the tasks of entity retrieval and VQA performance.

Retrieval results. As shown in Table 2, our method achieves the best overall performance on both E-VQA and InfoSeek. In the retrieval stage, it consistently outperforms all baselines across different Recall@K metrics. Specifically, on E-VQA, our method achieves R@1, R@5, R@10, and R@20 scores of 23.5, 43.2, 52.7, and 60.2, respectively, surpassing the strongest baseline, OMGM. On InfoSeek, it also delivers strong performance, achieving 56.1 R@1 and further improving to 82.7 R@10 and 86.9 R@20.

In the reranking stage, our method maintains clear advantages over EchoSight and OMGM, improving R@1 on E-VQA by 8.7 and 2.6 percentage points, respectively, with similar gains on InfoSeek. These results indicate that enhanced retrieval not only increases candidate recall but also provides higher-quality candidates for reranking, thereby boosting the overall VQA pipeline performance.

VQA results. As shown in Table 3, on InfoSeek, RGCC achieves an overall accuracy of 32.1. While a direct comparison with the reported figures of OMGM is hindered by unreleased model weights, RGCC significantly outperforms the reproduced OMGM baseline (29.4) under identical generative conditions. Importantly, this performance advantage is further validated on the E-VQA dataset, where RGCC explicitly surpasses both the reported and reproduced results of OMGM (51.8 vs. 50.2 and 49.5). Taken together, these results provide robust empirical evidence that optimizing

Method	E-VQA				InfoSeek			
	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
CLIP I-T	3.3	7.7	12.1	16.5	32.0	54.0	61.6	68.2
Wiki-LLaVA	3.3	-	9.9	13.2	36.9	-	66.1	71.9
LLM-RA	-	-	-	-	47.3	53.8	-	-
mR ² AG	-	-	-	-	38.0	-	65.0	71.0
ReflectiVA*	15.6	36.1	-	49.8	56.1	77.6	-	86.4
VLM-PRF	-	-	-	-	45.6	54.9	-	-
EchoSight								
<i>w/o. reranking</i>	13.3	31.3	41.0	48.8	45.6	67.1	73.0	77.9
<i>w. reranking</i>	36.5	47.9	48.8	48.8	53.2	74.0	77.4	77.9
OMGM								
<i>w/o. reranking</i>	19.1	41.2	49.8	58.7	52.6	73.9	80.0	84.8
<i>w. reranking</i>	<u>42.6</u>	<u>55.5</u>	<u>58.0</u>	<u>58.7</u>	<u>63.4</u>	<u>80.1</u>	<u>83.5</u>	<u>84.8</u>
RGCC (Ours)								
<i>w/o. reranking</i>	23.5	43.2	52.7	60.2	56.1	76.9	82.7	86.9
<i>w. reranking</i>	45.2	57.8	59.6	60.2	65.9	82.6	85.8	86.9

Table 2: Retrieval and re-ranking results on the E-VQA test set and the InfoSeek validation set. “w/o reranking” and “w. reranking” denote the settings without and with reranking, respectively. ReflectiVA* is evaluated on a different number of test samples than our method and other baselines. Best in bold, second-best underlined.

Method	Generation Model	Gen. FT	Ret. FT	E-VQA	InfoSeek		
					Unseen-Q	Unseen-E	Overall
RoRA-VLM	LLaVA-1.5-7B	✓	✗	20.3	27.3	25.1	-
Wiki-LLaVA	LLaVA-1.5-7B	✓	✗	21.8	30.1	27.8	28.9
LLM-RA	BLIP2-Flan-T5XL	✓	✓	-	26.1	20.9	23.1
mR ² AG	LLaVA-1.5-7B	✓	✓	-	40.6	39.8	40.2
ReflectiVA	LLaVA-MORE-8B	✓	✓	35.5	40.4	39.8	40.1
VLM-PRF	InternVL3-8B	✓	✓	40.1	43.5	42.1	42.5
EchoSight	Mistral-7B LLaMA3-8B	✗	✓	41.8	-	-	31.3
OMGM†	LLaVA-1.5-7B	✓	✓	50.2	43.5	43.5	43.5
OMGM*	LLaVA-1.5-7B	✓	✓	49.5	32.1	27.1	29.4
RGCC (ours)	LLaVA-1.5-7B	✓	✓	51.8	33.5	30.9	32.1

Table 3: VQA accuracy comparison with the baselines. Gen. FT indicates whether the generator was fine-tuned. † The originally reported VQA accuracy of OMGM on InfoSeek (43.5) could not be directly replicated due to unreleased generator weights (achieving 29.4 under our controlled reproduction), while * indicates our reproduced results. Crucially, on the E-VQA dataset, RGCC consistently outperforms both the originally reported and reproduced results of OMGM (51.8 vs. 50.2 and 49.5). The best result in bold font. Detailed definitions of Unseen-Q and Unseen-E can be found in Appendix B.

compressed knowledge for retriever alignment consistently improves downstream question-answering quality across different knowledge-intensive benchmarks.

4.4 Ablation study

We conduct extensive experiments to assess the effectiveness of our framework.

The Impact of Data Selection Strategies. For a fair comparison, we employ a random sampling strategy that selects a subset of training samples through random sampling, while keeping the number of selected samples identical to that produced by the Retrieval-driven Data Screening Strategy. In addition, we conduct experimental analyses

on the hyperparameters k and m . The specific retrieval results of different strategies are shown in Table 4. Compared with random sampling, our data selection strategy consistently improves performance across all recall metrics on E-VQA. In particular, the significant gain on R@1 demonstrates that our method effectively enhances high-confidence retrieval, indicating superior data quality and stronger semantic alignment. The experiment on the InfoSeek dataset can be found in Appendix D.2.

Impact of Modalities and Granularity on Retrieval Performance. We use the query image as the retrieval query, where the image uniquely contains the target entity. Candidate represen-

Strategy	E-VQA			
	R@1	R@5	R@10	R@20
Random	19.7	40.9	49.2	57.8
1-10	20.1	41.7	50.2	58.9
10-50	21.0	41.6	51.2	59.1
20-50	23.5	43.2	52.7	60.2

Table 4: Different data strategies on E-VQA. “1–10” denotes training with samples whose retrievability ranks fall between top-1 and top-10, and other settings follow the same definition.

Ret. Method	E-VQA			
	R@1	R@5	R@10	R@20
Image→Article	13.2	27.7	35.5	41.7
Image→Image	13.4	31.8	41.9	48.8
Image→Title	17.5	31.9	38.6	44.8
Image→Vanilla.Comp	18.3	37.8	48.3	56.1
Image→Aligned.Comp	23.5	43.2	52.7	60.2

Table 5: Ablation Study on the Effects of Modality and Granularity on Retrieval Performance. Vanilla.Comp denotes Vanilla Compressed Text, while Aligned.Comp denotes Aligned Compressed Text.

tations are constructed from different modalities and knowledge granularities. Table 5 reports the retrieval results when using images, article bodies, titles, vanilla compressed text, and aligned compressed text as candidates. The results show that Image-to-Aligned.comp achieves the best performance across all reported evaluation metrics. While vanilla.comp and the query image also deliver competitive results, Image-to-Aligned.comp consistently outperforms other candidates. These findings demonstrate the advantage of aligned compressed text, highlighting the effectiveness of alignment-aware compression for image-text retrieval. The experiment on the InfoSeek dataset can be found in Appendix D.1.

Impact of Compression Model for Retrieval.

Table 6 analyzes the impact of different compression models on E-VQA retrieval. Llama3-8B-Instruct achieves competitive performance, sug-

Model	E-VQA			
	R@1	R@5	R@10	R@20
Llama3-8B-Instruct	19.1	41.2	49.8	58.7
Qwen2.5-1.5B-Instruct				
<i>w/o. training</i>	18.3	37.8	48.3	56.1
<i>w. training</i>	23.5	43.2	52.7	60.2

Table 6: Impact of Compression Model for retrieval on E-VQA.

gesting that strong instruction-tuned LLMs can generate informative compressed representations. However, the untrained Qwen2.5-1.5B-Instruct performs worse, particularly at Recall@1 and Recall@5, indicating that general-purpose compression may fail to preserve retrieval-relevant evidence.

After training, Qwen2.5-1.5B-Instruct shows substantial improvements, increasing Recall@1 from 18.3 to 23.45 and Recall@5 from 37.8 to 43.2. Notably, it also outperforms the much larger Llama3-8B-Instruct across all Recall@K metrics. These results demonstrate that our training strategy effectively adapts the compression model to the retrieval objective, making task-specific retrieval-guided compression more beneficial than relying on larger off-the-shelf LLMs. Results on InfoSeek are reported in Appendix D.4.

InfoSeek	R@1	R@5	R@10	R@20
Vanilla.Comp	52.6	73.9	80	84.8
Aligned.Comp	55	76.2	82.4	86.8

Table 7: Cross-dataset generalization from E-VQA to InfoSeek. The compression model trained on E-VQA is directly evaluated on InfoSeek. Vanilla.Comp and Aligned.Comp denote vanilla and aligned compressed text, respectively.

Cross-dataset generalization. To further assess generalization, we conduct cross-dataset evaluations by training the compression model on one dataset and directly testing it on another without further fine-tuning. As shown in Table 7, the compressed texts learned from E-VQA generalize well to InfoSeek, indicating that our retriever-guided compression captures transferable visual-textual alignment patterns rather than overfitting to the training data. Additional generalization experiments are presented in Table 19.

Effect of different reward models. We investigate the effects of the reward model and test-time retriever on E-VQA. As shown in Table 8, reward training consistently improves retrieval performance. Replacing CLIP with EVA-CLIP as the test-time retriever further yields substantial gains. The best performance is achieved by combining CLIP-based reward training with the EVA-CLIP retriever, obtaining R@1/R@5/R@10/R@20 scores of 19.7/40.5/50.2/60.0. These results demonstrate the complementary benefits of reward supervision and a stronger test-time retriever.

Effect of the KL coefficient. We further study

RM	Retriever	RT	R@1	R@5	R@10	R@20
–	CLIP	×	3.3	8.1	12.3	18.4
EVA-CLIP	CLIP	✓	4.0	11.0	16.3	23.2
–	EVA-CLIP	×	13.2	27.7	35.5	41.7
CLIP	EVA-CLIP	✓	19.7	40.5	50.2	60.0

Table 8: Ablation study on different reward models and test-time retrievers on E-VQA. RM denotes the reward model, while RT indicates whether reward training is applied.

KC	R@1	R@5	R@10	R@20
0.1	20.5	40.2	49.7	58.9
0.01	20.9	42.2	49.8	58.0
0.001	23.5	43.2	52.7	60.2

Table 9: Ablation study of different KL-divergence Coefficients on E-VQA. KC denotes the KL-divergence coefficient.

the effect of the KL-divergence coefficient λ_{KL} by varying it among 0.1, 0.01, and 0.001. As reported in Table 9, decreasing the coefficient generally improves retrieval performance. In particular, setting $KC = 0.001$ achieves the best results across all evaluation metrics, with R@1, R@5, R@10, and R@20 scores of 23.5, 43.2, 52.7, and 60.2, respectively. Compared with $KC = 0.1$, it improves R@1 by 3.0 points and R@20 by 1.3 points. This suggests that an excessively strong KL constraint may limit the model’s ability to preserve retrieval-relevant information. Therefore, we set $KC = 0.001$ in all main experiments.

Impact of Training Strategies. We compare no training, SFT, and RGCC on E-VQA and InfoSeek. SFT uses the candidate text with the highest image–text similarity score as a silver-label target.

As shown in Table 10, RGCC achieves the best performance on both datasets. On E-VQA, it improves R@1 from 18.3 to 23.5 and consistently outperforms both baselines across all metrics. On InfoSeek, RGCC achieves R@1, R@5, R@10, and R@20 scores of 56.1, 76.9, 82.7, and 86.9, respectively. SFT performs below the no-training baseline on E-VQA, possibly because its objective does not explicitly incorporate visual information, leading to a mismatch with image-to-text retrieval. In contrast, RGCC directly optimizes retrieval performance and better preserves visually relevant information, demonstrating the effectiveness of reward-based optimization for multimodal retrieval.

Effect of Reward Design. We investigate different reward designs for E-VQA image-to-text

Dataset	Strategy	R@1	R@5	R@10	R@20
E-VQA	No training	18.3	37.8	48.3	56.1
	SFT	15.9	33.1	42.7	51.4
	RGCC	23.5	43.2	52.7	60.2
InfoSeek	No training	52.9	73.6	79.7	84.5
	SFT	51.7	74.1	80.5	85.4
	RGCC	56.1	76.9	82.7	86.9

Table 10: Impact of different training strategies on retrieval performance.

Reward Design	R@1	R@5	R@10	R@20
Image-only	23.5	43.2	52.7	60.2
Image+question	2.3	9.5	14.2	19.6
Performance Drop	21.2	33.7	38.5	40.6

Table 11: Ablation study of different reward designs on E-VQA. Performance is evaluated using Recall@K (R@K).

retrieval. As shown in Table 11, the image-only reward substantially outperforms the joint image-question reward across all metrics (R@1/5/10/20: 23.5/43.2/52.7/60.2 vs. 2.3/9.5/14.2/19.6). Incorporating the question may overemphasize textual correspondence and weaken visual grounding, leading to noisy supervision. Thus, directly optimizing image-text relevance provides a more task-aligned learning signal.

5 Conclusion

We introduce RGCC, a Retriever-Guided Context Compression framework that enhances Multimodal Retrieval-Augmented Generation (MMRAG) for knowledge-based VQA (KB-VQA). RGCC effectively tackles two key challenges: distributional misalignment and granularity misalignment. To address these, we employ a retrievability-driven data screening strategy to select high-quality image–entity pairs and leverage Group Relative Policy Optimization (GRPO) with a visual-similarity reward to train a compression model that condenses long contexts into retriever-preferred representations without losing critical visual information. Extensive experiments on InfoSeek and E-VQA show that RGCC achieves state-of-the-art retrieval and reranking performance, substantially improving downstream VQA accuracy while introducing no additional inference-time overhead. Overall, RGCC offers an effective and scalable solution for bridging these gaps in multimodal retrieval.

Limitations

Our approach primarily focuses on retrieval, where we have observed significant performance improvements. However, several limitations remain: 1) The multimodal database contains both text and images at varying granularities, and our current method does not fully exploit the image information to optimize the retrieval process. This presents a promising direction for future research. 2) Although we have successfully enhanced reranking efficiency and improved visual question answering performance through better retrieval quality, the reranking process is still constrained by slow inference speed, limiting its practical applicability. Further research is needed to optimize the inference speed of reranking and explore how generative models can more effectively utilize the retrieved knowledge to further enhance answer quality. 3) Our current approach mainly targets knowledge-intensive VQA tasks centered on concrete entities, such as landmarks, species, and other visually identifiable objects. As a result, its generalization to broader VQA scenarios involving abstract concepts, events, attributes, or commonsense reasoning remains underexplored.

Acknowledgements

The research work has been supported by the Natural Science Foundation of China under Grant No. 62476271

References

- Huda Diab Abdulgalil and Otman A Basir. 2025. Next-generation image captioning: A survey of methodologies and emerging challenges from transformers to multimodal large language models. *Natural Language Processing Journal*, page 100159.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. Deep reinforcement learning: A brief survey. *IEEE signal processing magazine*, 34(6):26–38.
- Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. 2023. A suite of generative tasks for multi-level multimodal webpage understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 1917–1947.
- 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826.
- Wang Chen, Wenhan Yu, Guanqiang Qi, Weikang Li, Yang Li, Lei Sha, Deguo Xia, and Jizhou Huang. 2025. Cmrage: Co-modality-based visual document retrieval and question answering. *arXiv preprint arXiv:2509.02123*.
- 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? *arXiv preprint arXiv:2302.11713*.
- Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2025. Augmenting multimodal llms with self-reflective tokens for knowledge-based visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9199–9209.
- Mohsen Fayyaz, Ali Modarressi, Hinrich Schuetze, and Nanyun Peng. 2025. Collapse of dense retrievers: Short, early, and literal biases outranking factual evidence. *arXiv preprint arXiv:2503.05037*.
- Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G Bellemare, Joelle Pineau, and 1 others. 2018. An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, 11(3-4):219–354.
- Demian Ghalandari, Chris Hokamp, and Georgiana Ifrim. 2022. Efficient unsupervised sentence compression by fine-tuning transformers with reinforcement learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1267–1280.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.
- Yuyang Hong, Jiaqi Gu, Qi Yang, Lubin Fan, Yue Wu, Ying Wang, Kun Ding, Shiming Xiang, and Jieping Ye. 2025. Knowledge-based visual question answer with multimodal processing, retrieval and filtering. *arXiv preprint arXiv:2510.14605*.

- Chan-Wei Hu, Yueqi Wang, Shuo Xing, Chia-Ju Chen, Suofei Feng, Ryan Rossi, and Zhengzhong Tu. 2025. mrag: Elucidating the design space of multi-modal retrieval-augmented generation. *arXiv preprint arXiv:2505.24073*.
- 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12075.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. 2024. Preflmr: Scaling up fine-grained late-interaction multi-modal retrievers. *arXiv preprint arXiv:2402.08327*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.
- Thomas Mensink, Jasper Uijlings, Lluís 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3113–3124.
- Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/cvf winter conference on applications of computer vision*, pages 1527–1536.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.
- Multi-Linguality Multi-Functionality Multi-Granularity. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.
- Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. 2022. Benchmarking robustness of multimodal image-text models under distribution shift. *arXiv preprint arXiv:2212.08044*.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.
- Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.
- Wei Tang, Yanpeng Sun, Shan Zhang, Xiaofan Li, Piotr Koniusz, Wei Li, Na Zhao, and Zechao Li. 2025. Artemis: Structured visual reasoning for perception policy learning. *arXiv preprint arXiv:2512.01988*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Yujan Ting, Xu Tang, Terrence Chen, and Weijing Huang. 2026. Bridging the memorization-utilization gap: Near-lossless context compression via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14949–14972.
- Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.
- Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisín Mac Aodha. 2021. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893.
- Shuhe Wang, Shengyu Zhang, Jie Zhang, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu, Guoyin Wang, and Eduard Hovy. 2024. Reinforcement learning enhanced llms: A survey. *arXiv preprint arXiv:2412.10400*.
- Weixi Weng, Jieming Zhu, Xiaojun Meng, Hao Zhang, Rui Zhang, and Chun Yuan. 2024. Learning to compress contexts for efficient knowledge-based visual question answering. *arXiv preprint arXiv:2409.07331*.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584.

Yibin Yan and Weidi Xie. 2024a. Echosight: Advancing visual-language models with wiki knowledge. *arXiv preprint arXiv:2407.12735*.

Yibin Yan and Weidi Xie. 2024b. [EchoSight: Advancing visual-language models with Wiki knowledge](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1538–1551, Miami, Florida, USA. Association for Computational Linguistics.

Wei Yang, Jingjing Fu, Rui Wang, Jinyu Wang, Lei Song, and Jiang Bian. 2025. Omgm: Orchestrate multiple granularities and modalities for efficient multimodal retrieval. *arXiv preprint arXiv:2505.07879*.

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

Jixiao Zhang and Chunsheng Zuo. 2025. [GRPO-LEAD: A difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5642–5665, Suzhou, China. Association for Computational Linguistics.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, and 1 others. 2024. mr2ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. *arXiv preprint arXiv:2411.15041*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix

A Prompt Used in Our Methodology

In this section, we present the prompt templates used in our methodology for invoking LLM/MLLM, covering both LLM-based summary generation and visual question answering (VQA) across multiple datasets. We use the summary generation and VQA templates provided by OMGM to ensure fairness in the method comparison. Table 26 displays the prompt template for guiding the LLMs in generating offline entity knowledge summaries, while Tables 27 and 28 show the prompt

templates for the E-VQA and InfoSeek datasets, respectively. Since InfoSeek adopts a strict exact match evaluation standard, we use a one-shot prompting approach to generate answers.

B Dataset Details and metrics

Encyclopedic VQA. Encyclopedic VQA (Yan and Xie, 2024b; Mensink et al., 2023; Chen et al., 2023) is a knowledge-intensive visual question answering benchmark designed to evaluate models’ ability to answer entity-centric questions with external knowledge. It contains approximately 221K question-answer pairs grounded in 16.7K fine-grained entities, where each entity is associated with up to five images. The entities and images are sourced from iNaturalist 2021 and Google Landmarks Dataset V2 (Van Horn et al., 2021; Weyand et al., 2020). E-VQA further provides a controlled knowledge corpus built from Wiki-Web2M (Burns et al., 2023), consisting of around 2M image-rich Wikipedia articles that contain evidence for answering the questions. Based on the required reasoning process, the questions are categorized into single-hop and two-hop types. The original dataset is split into training, validation, and test sets with about 1M, 13K, and 5.8K instances, respectively. Following prior work, we restrict both training and evaluation to single-hop questions.

Infoseek. The dataset is a large-scale visual information-seeking benchmark constructed from OVEN entities (Hu et al., 2023). It contains approximately 1.3 million image-question-answer triplets associated with around 11,000 visual entities. The dataset includes 8.9K human-authored visual information-seeking questions, together with 1.3M automatically generated questions. These samples are split into training, validation, and test subsets, comprising about 934K, 73K, and 348K instances, respectively. Since the test split does not provide ground-truth answers, we report performance on the validation set. Notably, both the validation and test splits are designed to evaluate generalization, as they contain questions involving either unseen entities or query types that do not appear in the training data. InfoSeek also provides a large-scale knowledge base with 6 million Wikipedia entities. Following prior work such as EchoSight (Yan and Xie, 2024a), we use a 100K-entity subset of this knowledge base, which covers the 6,741 entities appearing in the training and validation questions. During the construction of the

retrieval corpus, we collect images from the corresponding Wikipedia pages of each entity. We observe that a small number of validation samples are linked to correct entities for which no associated images are available. Therefore, these samples are removed from evaluation. The final validation set contains 71,335 samples, accounting for 96.9% of the original validation split, which suggests that this filtering has only a marginal impact on the overall evaluation.

Metrics. For evaluation, we use Recall@K and BEM(Zhang et al., 2019) to assess retrieval quality and answer generation performance, respectively. Recall@K measures whether the ground-truth entity is included among the top-K retrieved candidates, thereby reflecting the effectiveness of both the retrieval and reranking modules. For answer evaluation on E-VQA, we follow the official evaluation protocol and adopt BEM as the primary metric. BEM is based on a BERT-based answer equivalence model that compares generated responses with reference annotations. Compared with exact-match evaluation, which only rewards identical surface forms, BEM can better capture semantic equivalence between predictions and ground-truth answers despite differences in wording.

For InfoSeek, we follow its official evaluation setting and adopt answer-type-specific metrics. For string-based questions, including those requiring entity names, we report VQA accuracy (Antol et al., 2015), which supports multiple valid annotations and considers a prediction correct if it matches any accepted answer. For numerical questions, we use relaxed accuracy (Methani et al., 2020), which allows small numerical deviations by treating predictions as correct when they fall within a predefined tolerance range of the ground-truth answer.

Unseen-Q. Unseen-Q corresponds to the Unseen Question setting, where specific question types (or templates) are held out during training and only included during evaluation, assessing generalization to new questions.

Unseen-E. Unseen-E refers to the Unseen Entity setting, where certain entities are held out during training and only appear during evaluation, testing the model’s generalization to new entities.

C More Experimental Details

C.1 Overall Framework and Model Checkpoints.

Given that current language models (such as the Qwen series) exhibit excellent understanding and compression capabilities, and can efficiently follow instructions, while considering computational resource consumption and training efficiency, we have chosen to use Qwen2.5-1.5B-Instruct as the compression model. This model can effectively compress entity knowledge, and the compressed text length meets our needs. We use Eva-CLIP (Sun et al., 2023) as the retriever to extract embeddings of the query image and the compressed text, calculating the similarity between them. After obtaining the candidate article IDs produced by the retriever based on the compressed representations, we first use BLIP-2 to measure the relevance between the sections and images of each candidate article and the question-query image pair, thereby identifying the final top-1 article. Then, within this top-1 article, we use BGE-Reranker-v2-m3 (Multi-Granularity, 2024) to compute the semantic similarity between each section and the question, and select the section with the highest similarity as the final retrieval result. These, along with the image and the question, are then input into a multimodal language model (such as LLaVA-1.5-7B) to generate the final answer.

C.2 Training Details of the Compression Model.

To achieve visual-language alignment during text compression, we propose an image-text similarity reward. Specifically, we use the retriever to compute the similarity between the compressed text and the query image, and use this score as the reward signal to guide the iterative optimization of the compression model. This encourages the compressed text to retain information that is more relevant to the visual query, thereby alleviating domain bias and improving visual-language alignment.

We do not explicitly incorporate the query question into the reward design, because our retrieval setting focuses on matching the query image with textual documents in the knowledge base during the retrieval stage. Importantly, compression is used only as an offline, coarse-grained mechanism for article-level candidate selection. The retriever first produces candidate article IDs based on compressed representations, while the subse-

quent reranking stage accesses the full, uncompressed articles for fine-grained, section-level selection. Therefore, no original content is discarded during reranking or answer generation, ensuring that all answer-critical evidence remains fully available. Since compression is performed entirely as offline knowledge-base preprocessing, it also introduces no additional online inference cost or latency to the VQA pipeline.

In contrast, the question typically specifies the target attribute to be answered, which may not have direct semantic correspondence with the encyclopedic documents. Incorporating the question into the reward could therefore introduce noise into the retriever-guided compression objective. In addition, we do not introduce a separate length-based reward, as the compression length can be effectively controlled through the instruction prompt. This design keeps the reward focused on retrieval-guided vision-language alignment.

We use Qwen2.5-1.5B-Instruct as the base compression model and optimize it with GRPO. During training, we generate 8 rollouts for each input, set the sampling temperature to 1.0, train the model for 5 epochs, and use a learning rate of 3e-6. The KL coefficient β is set to 0.001. The EVA-CLIP-8B model is kept frozen throughout training and is only used to compute the semantic similarity score between the query image and the compressed text. We conducted the experiment on a server with 4x NVIDIA A800.

C.3 Reranker Training.

The retrieval strategy we employ is $IT_Q \rightarrow IT_D$. In IT_Q , "I" and "T" represent the image and its corresponding question, respectively. In IT_D , "I" and "T" refer to the first image and the section within the entity, respectively. We crawl the images based on the URLs provided by OMGM. For any invalid URL links, we replace the images with gray image. We use contrastive learning and employ hard negative samples from the previous retrieval step. We randomly select negative pairs by pairing the main images of candidate entities with non-evidentiary sections from their articles.

$$\mathcal{L} = -\log \frac{\exp(\text{sim}_m(q, d_+)/\mathcal{T})}{\sum_{j=1}^N \exp(\text{sim}_m(q, d_j)/\mathcal{T})} \quad (7)$$

We construct each training sample using N contrastive pairs and train the reranker to match queries

with the positive candidate pairs d_+ by leveraging hybrid-grained multimodal information. The reranker is optimized with an adaptive temperature $\mathcal{T}$, which dynamically adjusts the smoothness of the softmax distribution. We use the LAVIS library to train only the weights of the BLIP-2's Q-Former (Li et al., 2023). We select the top 32 embeddings from the model output as our multimodal fusion feature matrix. The learning rate is set to 1e-5, hard negative is set to 15, and the model is trained for one epoch. This configuration allows the training to complete in 12 hours on 1 Nvidia A800 (40G)

C.4 Reranker Inference and Scoring.

The reranker adopts a dual-encoder architecture to estimate the semantic relevance between the user query and each candidate knowledge pairs. In this architecture, the query pair consists of the user-provided image and the question, while the candidate pair is constructed from a candidate image and its associated textual section in the knowledge base. The two pairs are encoded separately but with a shared multimodal encoder, ensuring that their representations are projected into the same embedding space. Each pair is formulated as a unified multimodal sequence: the visual input is first encoded by a visual backbone, such as ViT (Vision Transformer), and the textual input is encoded by a transformer-based text encoder, such as BERT. The resulting visual and textual features are then fused via cross-modal attention to produce a joint multimodal embedding, which is subsequently used for relevance scoring and reranking.

Formally, let E_{IQ} denote the embedding of the user image-question pair, and let E_{IS} denote the embedding of the knowledge-base image-section pair. The reranking score is computed using cosine similarity:

$$s = \cos(E_{IQ}, E_{IS}) = \frac{E_{IQ}^\top E_{IS}}{\|E_{IQ}\|_2 \|E_{IS}\|_2} \quad (8)$$

During inference, the base retriever first retrieves the top- K candidate entities from the knowledge base, each containing candidate images and textual sections. The reranker then computes a similarity score between the query image-question representation and each knowledge-base image-section representation. Finally, all candidates are ranked in descending order according to their reranking scores.

Ret. Method	InfoSeek			
	R@1	R@5	R@10	R@20
Image→Article	44.5	64.5	70.7	76
Image→Image	45.6	68.6	74.6	77.9
Image→Title	51.5	69.2	74.8	79.1
Image→Vanilla.Comp	52.6	73.9	80	84.8
Image→Aligned.Comp	56.1	76.9	82.7	86.9

Table 12: Ablation Study on the Effects of Modality and Granularity on Retrieval Performance. Vanilla Compressed Text denotes Vanilla.Comp, while Aligned Compressed Text denotes Aligned.Comp.

C.5 Generator Training.

We randomly selected 100k data entries from the InfoSeek training set. After the retrieval and reranking steps, we selected the top-1 entity. Then, we used BGE-Reranker-v2-m3 (Multi-Granularity, 2024) to filter out the most relevant section for the question from that entity. Finally, we combined the question, query image, relevant section and corresponding prompt to create a training dataset.

We use the LLaVA (Liu et al., 2023) framework to perform LoRA fine-tuning on LLaVA-1.5-7B, setting the epoch to 1 and the learning rate to 2e-5.

We directly use the fine-tuned LLaVA-1.5-7B model provided by OMGM for E-VQA evaluation in our experiments. For InfoSeek, we follow OMGM’s training strategy and fine-tune LLaVA-1.5-7B with LoRA. The resulting fine-tuned model is then used for VQA evaluation on InfoSeek.

D More Ablation Studies

D.1 The Impact of Different Modalities and Granularities on InfoSeek.

Table 12 reports the retrieval performance on the InfoSeek dataset. As shown, all methods exhibit a consistent improvement in Recall@K as K increases, indicating stable ranking behavior. Among the compared approaches, Image→Article and Image→Image achieve relatively lower performance, suggesting that directly aligning images with full articles or relying solely on visual similarity is challenging due to semantic noise and granularity mismatch. In contrast, Image→Title yields a notable improvement at R@1, demonstrating that more concise and focused textual representations are beneficial for visual-language alignment. Furthermore, Image→Vanilla.Comp consistently outperforms Image→Title across all K values, highlighting the advantage of summary-level text in

Strategy	InfoSeek			
	R@1	R@5	R@10	R@20
Random	55.4	75.2	80.1	85.3
1-10	56.4	76.2	81.9	86.3
5-50	55.9	76.2	82.1	86.5
10-50	56.1	76.9	82.7	86.9
20-50	54.9	75.8	81.9	86.3

Table 13: Different data selection strategies on InfoSeek. “1–10” denotes training with samples whose retrievability ranks fall between top 1 and top 10; other intervals are defined analogously. “Random” denotes training data selected via random sampling.

Method	E-VQA			
	R@1	R@5	R@10	R@20
Echosight	36.5	47.9	48.8	48.8
OMGM	42.6	55.6	58.1	58.7
Ours	43.9	57.1	59.6	60.14

Table 14: Re-rank performance on E-VQA without additional training.

providing richer yet compact semantic information. Notably, Image→Aligned.Comp achieves the best results on all metrics, with clear gains especially at R@1, indicating that the Aligned Compressed Text enables more accurate semantic matching between images and text, thereby significantly enhancing image-text retrieval performance.

D.2 The Impact of Data Selection Strategies.

The experimental results presented in Table 13 on the InfoSeek dataset demonstrate that the Retrievability-driven data screening strategy consistently outperforms the Random strategy across all Recall@K metrics. Specifically, the Retrieval-Guided strategy achieves recall rates of 56.1%, 76.9%, 82.7%, and 86.9% at R@1, R@5, R@10, and R@20, respectively, showing steady improvements over the Random baseline. Notably, the most significant gain is observed at R@10, highlighting that incorporating retrieval guidance is particularly effective in enhancing retrieval quality for medium-sized Top-*k* candidate sets. These results validate our approach improves the relevance of retrieved results, making it a practical and effective strategy for information retrieval tasks.

Method	InfoSeek			
	R@1	R@5	R@10	R@20
Echosight	53.2	74	77.4	77.9
OMGM	63.8	80.07	83.5	84.8
Ours	65.4	82.4	85.7	86.9

Table 15: Re-rank performance on InfoSeek without additional training.

D.3 The reranking performance on E-VQA and InfoSeek.

To validate the effectiveness of our approach, we used the rerank model weights provided by OMGM, which were not trained on our retrieval results. The corresponding results are shown in Table 14 and Table 15. On the E-VQA, our method outperforms both the Echosight and OMGM methods across all metrics, achieving the best performance. For the R@1 metric, Ours achieved 43.9%, significantly higher than Echosight at 36.5% and OMGM at 42.6%. In the R@5, R@10, and R@20 metrics, Ours scored 57.1%, 59.6%, and 60.14%, respectively, also outperforming the other two methods. These results indicate that the Ours method provides more accurate predictions in re-ranking tasks, demonstrating a significant improvement in identifying relevant answers in the top ranks. On the InfoSeek, EchoSight performs the worst across all evaluation metrics, with scores of 53.2 for R@1, 74 for R@5, 77.4 for R@10, and 77.9 for R@20. OMGM outperforms EchoSight in all metrics, achieving R@1 of 63.8, R@5 of 80.07, R@10 of 83.5, and R@20 of 84.8. Ours method achieves the best performance across all evaluation metrics, with R@1 of 65.4, R@5 of 82.4, R@10 of 85.7, and R@20 of 86.9, significantly outperforming both EchoSight and OMGM. Overall, our method demonstrates the strongest ranking ability on the InfoSeek dataset, particularly improving R@1 by 1.6 compared to OMGM, indicating superior ranking performance. Even without retraining on our retrieval results, our method still outperforms the results of OMGM, which fully demonstrates that enhancing retrieval performance can effectively promote the efficiency of reranking.

D.4 Impact of Compression Model for retrieval.

The results are shown in Table 16. We compared different compression models and their impact

Model	InfoSeek			
	R@1	R@5	R@10	R@20
Llama3-8B-Instruct	52.6	73.9	80	84.8
Qwen2.5-1.5B-Instruct				
w/o. Training	52.9	73.6	79.7	84.5
w. Training	56.1	76.9	82.7	86.9

Table 16: Impact of Compression Model for retrieval on InfoSeek.

on retrieval performance before and after training. Due to resource constraints, we only explored the Qwen2.5-1.5B model. On InfoSeek, the Qwen model had the same impact on retrieval performance as Llama. After fine-tuning Qwen2.5-1.5B using the GPRO, our retrieval performance saw a significant improvement. On E-VQA, at R@1, it increased by points 4.3, it achieve a significant improvement. On infoSeek, at R@1, it increased by 3.2 points, and even at R@20, we were also able to achieve a good improvement. This reflects the effectiveness of our data filtering and compression model training.

Retriever	R@1	R@5	R@10	R@20
Siglip2	0.5	3.1	4.5	6.9
CLIP	3.3	8.1	12.3	18.4
EVA-CLIP	13.2	27.7	35.5	41.7

Table 17: Image-to-Text Retrieval Performance of Different Retrievers on E-VQA.

Retriever	R@1	R@5	R@10	R@20
Siglip2	4.5	9.7	12.9	16.1
CLIP	24	45.6	54.4	62.5
EVA-CLIP	44.5	64.5	70.7	76.0

Table 18: Image-to-Text Retrieval Performance of Different Retrievers on InfoSeek.

D.5 Performance of Different Retrievers.

For a fair comparison with prior methods that use EVA-CLIP as the retriever, we adopt EVA-CLIP in our main experiments. We further compare it with several widely used retrievers, including CLIP and SigLIP-2. The image-to-text retrieval results on E-VQA and InfoSeek are reported in Tables 17 and 18, respectively.

CLIP is a vision-language model trained with contrastive learning on large-scale image-text pairs, enabling strong cross-modal representation alignment and zero-shot transfer capabilities across a

wide range of visual understanding tasks. SigLIP-2(Tschannen et al., 2025) extends the SigLIP(Zhai et al., 2023) framework by replacing the standard contrastive objective with a sigmoid-based loss and incorporating improved training strategies and data scaling, leading to enhanced retrieval performance and stronger multimodal representation quality.

On E-VQA, EVA-CLIP substantially outperforms CLIP and SigLIP-2. Specifically, EVA-CLIP achieves Recall@1, Recall@5, Recall@10, and Recall@20 scores of 13.2, 27.7, 35.5, and 41.7, respectively, whereas CLIP obtains only 3.3, 8.1, 12.3, and 18.4. SigLIP-2 performs worst on this dataset, with 0.5 Recall@1 and 6.9 Recall@20. These results suggest that E-VQA poses a particularly challenging retrieval setting, where fine-grained visual entities require stronger cross-modal matching ability.

On InfoSeek, all retrievers achieve higher recall scores than on E-VQA, suggesting that InfoSeek is relatively less challenging for image-to-text retrieval. Nevertheless, EVA-CLIP still maintains a clear advantage, achieving Recall@1, Recall@5, Recall@10, and Recall@20 scores of 44.5, 64.5, 70.7, and 76.0, respectively.

Overall, these results demonstrate that the choice of retriever has a substantial impact on retrieval performance. EVA-CLIP consistently delivers the strongest results on both datasets, supporting its use as the default retriever in our framework.

Furthermore, the mRAG study (Hu et al., 2025) shows that EVA-CLIP consistently outperforms other mainstream retrievers on both InfoSeek and E-VQA across different retrieval granularities. This further supports our choice of EVA-CLIP as the default retriever for fair and competitive evaluation.

E-VQA	R@1	R@5	R@10	R@20
Vanilla.Comp	18.3	37.8	48.3	56.1
Aligned.Comp	18.7	39.2	48.7	57.1

Table 19: Cross-dataset generalization from InfoSeek to E-VQA, where the compression model trained on InfoSeek is directly evaluated on E-VQA. Vanilla.Comp and Aligned.Comp denote vanilla compressed text and alignment-aware compressed text, respectively.

D.6 Analysis of Retriever Bias and Cross-Dataset Generalization.

To further examine the generalization ability of the proposed compression strategy, we conduct cross-dataset evaluations. Specifically, we train the com-

pression model on one dataset and directly evaluate its retrieval performance on the other dataset without further fine-tuning. The results are reported in Tables 7 and 19.

As shown in Table 7, when trained on E-VQA and evaluated on InfoSeek, Aligned.Comp consistently outperforms Vanilla.Comp across all Recall@K metrics. It improves R@1 from 52.6 to 55.0, R@5 from 73.9 to 76.2, R@10 from 80.0 to 82.4, and R@20 from 84.8 to 86.8. These results indicate that the learned compression strategy can transfer to InfoSeek and produce retriever-aligned representations beyond the source dataset.

Table 19 presents the reverse setting, where the model is trained on InfoSeek and evaluated on E-VQA. Aligned.Comp again improves over Vanilla.Comp across all metrics, increasing R@1 from 18.3 to 18.7, R@5 from 37.8 to 39.2, R@10 from 48.3 to 48.7, and R@20 from 56.1 to 57.1. Although the gains are relatively smaller, the consistent improvements suggest that the aligned compression model maintains cross-dataset robustness.

Overall, these results demonstrate that retriever-guided alignment improves the transferability of compressed representations across KB-VQA datasets. The aligned compression model generalizes beyond the dataset on which it is trained, suggesting that it learns transferable retrieval-relevant compression patterns rather than merely fitting dataset-specific characteristics.

Method	E-VQA	InfoSeek		
		Unseen-Q	Unseen-E	Overall
GPT-4V	28.1	15.0	14.3	14.6
RGCC (ours)	51.8	33.5	30.9	32.1

Table 20: Comparison with GPT-4V

D.7 Comparison with Closed-Source model.

Table 20 compares RGCC with GPT-4V on E-VQA and InfoSeek. RGCC consistently outperforms GPT-4V across all evaluation settings. On E-VQA, RGCC achieves an accuracy of 51.8, surpassing GPT-4V by 23.7 points. On InfoSeek, RGCC improves the overall accuracy from 14.6 to 32.1, with substantial gains on both the Unseen-Q and Unseen-E splits. These results indicate that, although GPT-4V exhibits strong multimodal understanding capabilities, its parametric knowledge alone may be insufficient to cover the fine-grained external knowledge required for

knowledge-intensive VQA. In contrast, by explicitly retrieving external knowledge and improving retrieval quality through retriever-guided context compression, RGCC can acquire more relevant evidence and achieve stronger answer accuracy.

The consistent improvements over GPT-4V further demonstrate the importance of retrieval-augmented reasoning in knowledge-intensive visual question answering. In addition, our method focuses on open-source models, which ensures reproducibility and facilitates fair comparison with existing methods.

D.8 Impact of Token Compression Ratio on Retrieval Performance

We investigate how the compressed context length and the original context length affect retrieval performance on InfoSeek and E-VQA. Retrieval quality is evaluated using Recall@K ($R@K$), where $K \in \{1, 5, 10, 20\}$.

We investigate how the compressed context length and the original context length affect retrieval performance on InfoSeek and E-VQA. Retrieval quality is evaluated using Recall@K ($R@K$), where $K \in \{1, 5, 10, 20\}$.

Effect of compressed context length. Table 21 compares different compressed lengths while fixing the original context length to 1,000 tokens. On InfoSeek, compressing the context to 100 tokens achieves the best overall performance, obtaining 52.9 $R@1$, 74.2 $R@5$, 80.1 $R@10$, and 84.9 $R@20$. Increasing the compressed length from 100 to 128 tokens does not further improve retrieval performance, suggesting that a compact representation already preserves sufficient retrieval-relevant information for InfoSeek.

E-VQA exhibits a different pattern. Increasing the compressed length generally improves retrieval performance, particularly at larger retrieval depths. Compressing the context to 128 tokens produces the highest $R@1$, $R@10$, and $R@20$ scores of 18.3, 48.3, and 56.1, respectively. Compared with a compressed length of 77 tokens, the 128-token setting improves $R@1$ by 1.2 points and $R@10$ by 1.4 points. The 77-token setting achieves a slightly higher $R@5$ score (38.1 versus 37.8), but its overall performance is lower. These results indicate that E-VQA requires a relatively larger compressed representation to retain the fine-grained visual and entity-related information needed for retrieval.

Effect of original context length. Table 22 examines the influence of the original context length while fixing the compressed length to 128 tokens. Retrieval performance on InfoSeek remains relatively stable across original lengths ranging from 500 to 2,500 tokens. The 700-token setting achieves the highest $R@1$ and $R@5$ scores, whereas the 2,500-token setting obtains the best $R@10$ and $R@20$ scores. Nevertheless, the differences among these settings are small: the maximum variations are only 0.6 points for $R@1$ and 0.5 points for $R@20$. This demonstrates that the compression module is robust to changes in the original context length on InfoSeek.

For E-VQA, the effect of original context length is more pronounced and non-monotonic. Using 1,000 original tokens gives the best deep-retrieval performance, reaching 48.3 $R@10$ and 56.1 $R@20$. Although the 2,500-token setting achieves the highest $R@1$ and $R@5$ scores of 18.7 and 38.2, its $R@20$ score decreases to 54.6. Moreover, increasing the original length to 2,000 tokens results in a noticeable performance drop. This suggests that adding more context does not necessarily improve retrieval, as additional tokens may introduce irrelevant or redundant information.

Overall findings. The results show that retrieval performance depends more strongly on the compressed length than on the amount of original context. A compressed length of 100 tokens is sufficient for InfoSeek, while E-VQA benefits from retaining 128 tokens. Considering performance across both datasets, we use 128 tokens as the default compressed length in subsequent experiments. We set the original context length to 1,000 tokens because it provides a favorable balance between retrieval effectiveness and computational efficiency, especially for E-VQA.

D.9 Comparison of Original, Vanilla-Compressed, and Aligned-Compressed Texts

As shown in Figure 3, we present the differences among the original Wikipedia text, the vanilla compressed text, and the aligned compressed text in the visual retrieval task. The original encyclopedic text on the left contains extensive background knowledge, definitions, and distributional information. Although it is comprehensive, it is lengthy and introduces substantial noise that is irrelevant to visual matching. The vanilla compressed text in the mid-

Dataset	Length (Original → Compressed)	R@1	R@5	R@10	R@20
InfoSeek	1000 → 77	52.6	73.8	79.9	84.6
	1000 → 100	52.9	74.2	80.1	84.9
	1000 → 128	52.9	73.6	79.7	84.5
E-VQA	1000 → 77	17.1	38.1	46.9	54.6
	1000 → 100	17.4	35.9	45.2	54.8
	1000 → 128	18.3	37.8	48.3	56.1

Table 21: Effect of Compression Length on Retrieval Performance

Dataset	Length (Original → Compressed)	R@1	R@5	R@10	R@20
InfoSeek	500 → 128	52.5	73.9	80.1	84.8
	700 → 128	53.1	74.4	80.2	84.8
	1000 → 128	52.9	73.6	79.7	84.5
	1500 → 128	52.8	73.9	80.0	84.7
	2000 → 128	52.9	74.3	80.0	84.8
	2500 → 128	52.9	74.1	80.3	85.0
E-VQA	500 → 128	16.7	37.1	44.6	51.9
	700 → 128	17.5	37.3	46.2	53.3
	1000 → 128	18.3	37.8	48.3	56.1
	1500 → 128	18.6	37.2	46.1	54.8
	2000 → 128	15.8	36.5	46.4	54.4
	2500 → 128	18.7	38.2	46.9	54.6

Table 22: Effect of Original Context Length on Retrieval Performance

dle is shorter, but it still mainly preserves general encyclopedic facts and pays insufficient attention to the visual characteristics of the target entity. In contrast, the aligned compressed text on the right preserves the core semantics while placing greater emphasis on image-relevant visual attributes, such as a small body size, sturdy build, short limbs, and a long tail. These results show that aligned compression can generate text representations that better match the retriever’s preferences, thereby improving image-text matching performance.

D.10 Analysis of the Reproduction Gap

We compare RGCC with OMGM on the E-VQA and InfoSeek benchmarks. As shown in Table 23, our reproduced OMGM results are close to the originally reported result on E-VQA, achieving 49.5 compared with the reported score of 50.2. However, a substantial reproduction gap is observed on InfoSeek, where the reproduced overall score is 29.4, considerably lower than the reported score of 43.5.

We attribute this discrepancy primarily to the difference in the granularity of training supervision provided by the two datasets. E-VQA provides the ground-truth evidence section for each question, allowing the generator to be directly trained with accurately localized supporting evidence. In contrast, InfoSeek provides only the target entity

identifier without annotating the corresponding ground-truth section. Since each entity usually contains one answer-relevant section together with many irrelevant sections, constructing generator training data for InfoSeek requires an additional section-selection step. Specifically, we employ BGE-Reranker-v2-M3 to rank all sections of the target entity according to their semantic relevance to the question and select the highest-scoring section as the generator input.

Nevertheless, the section with the highest semantic relevance score does not necessarily contain the evidence required to answer the question. Incorrect, incomplete, or only superficially relevant sections may therefore be selected as training evidence, introducing noisy supervision into generator fine-tuning. Such noise weakens the model’s ability to learn reliable correspondences among the question, supporting evidence, and answer. This problem is further amplified by the strict answer-matching evaluation adopted by InfoSeek, which makes the final score particularly sensitive to errors in evidence localization and answer generation. In addition, the original generator weights, LoRA configuration, and complete training details of OMGM are not publicly available, further limiting exact reproduction.

Despite this reproduction gap, RGCC consistently outperforms the reproduced OMGM base-

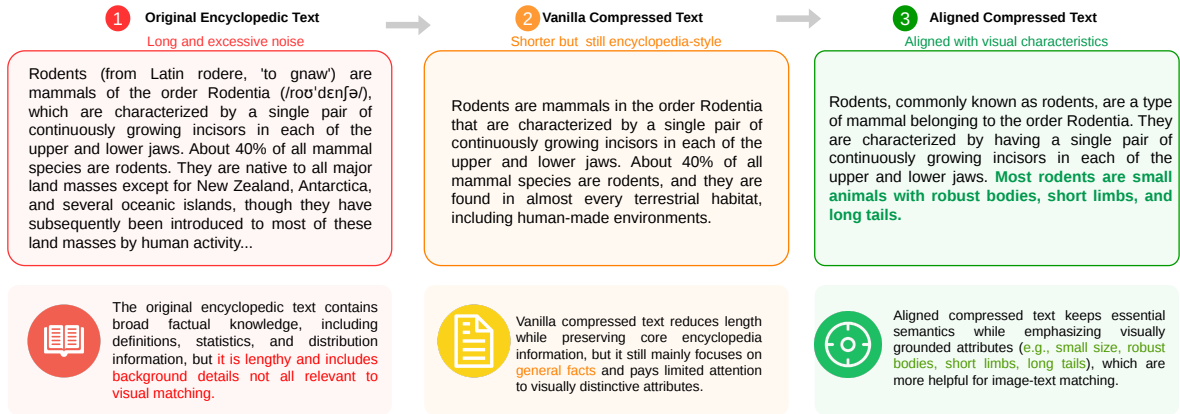


Figure 3: Comparison of original encyclopedic text, vanilla compressed text, and aligned compressed text. The original text contains broad factual knowledge but includes noise irrelevant to visual matching. Vanilla compression shortens the text while mainly preserving general encyclopedic information. In contrast, aligned compression emphasizes visually grounded attributes, producing compact text that better aligns with the retriever’s image-text matching preference.

line under the same fine-tuning setting. As reported in Table 23, RGCC improves the E-VQA score from 49.5 to 51.8. On InfoSeek, it improves the Unseen-Q, Unseen-E, and overall scores from 32.1, 27.1, and 29.4 to 33.5, 30.9, and 32.1, respectively. These results correspond to absolute improvements of 1.4, 3.8, and 2.7 points on the three InfoSeek evaluation settings.

To reduce the influence of generator-specific factors and enable a more controlled comparison, we further evaluate both methods using the same untuned InternVL2.5-8B generator. Neither SFT nor LoRA fine-tuning is performed, and both methods use identical prompts, decoding parameters, candidate-set sizes, and reranking configurations. The results are presented in Table 24.

Without reranking, RGCC improves OMGM by 3.10 points on E-VQA and by 1.00, 1.20, and 1.10 points on InfoSeek Unseen-Q, Unseen-E, and Overall, respectively. When reranking is enabled, RGCC achieves improvements of 1.98 points on E-VQA, 0.63 points on InfoSeek Unseen-E, and 0.22 points on InfoSeek Overall. Although RGCC is 0.22 points lower on InfoSeek Unseen-Q, it still obtains a higher overall InfoSeek score. These controlled results demonstrate that the performance gains of RGCC do not depend on a particular generator or fine-tuning configuration. Instead, the improvements mainly originate from its compressed knowledge index and its ability to preserve answer-relevant information during retrieval.

Method	E-VQA	InfoSeek Unseen-Q	InfoSeek Unseen-E	InfoSeek Overall
OMGM (reported)	50.2	43.5	43.5	43.5
OMGM (reproduced)	49.5	32.1	27.1	29.4
RGCC (Ours)	51.8	33.5	30.9	32.1

Table 23: Comparison with OMGM on E-VQA and InfoSeek using the fine-tuned LLaVA-1.5-7B generator.

Reranking	Method	E-VQA	InfoSeek Unseen-Q	InfoSeek Unseen-E	InfoSeek Overall
No	OMGM	32.60	32.50	31.10	31.80
No	RGCC (Ours)	35.70	33.50	32.30	32.90
No	Improvement	+3.10	+1.00	+1.20	+1.10
Yes	OMGM	48.90	37.16	34.95	36.03
Yes	RGCC (Ours)	50.88	36.94	35.58	36.25
Yes	Improvement	+1.98	-0.22	+0.63	+0.22

Table 24: Comparison with OMGM on E-VQA and InfoSeek using the same untuned InternVL2.5-8B generator.

D.11 Effect of Retrieval-Driven Data Screening.

We investigate the effect of different screening thresholds to analyze the trade-off between training sample quality and quantity. As shown in Table 25, the screening interval of 20–50 achieves the best performance on E-VQA across all retrieval metrics, obtaining R@1, R@5, R@10, and R@20 scores of 23.0, 43.2, 52.7, and 60.2, respectively. In particular, its R@1 score exceeds those of the 1–10 and 10–50 intervals by 2.9 and 2.0 points, respectively, while retaining 26K training samples. These results indicate that excluding overly easy samples allows the model to learn from more informative and challenging training instances.

Dataset	Strategy	R@1	R@5	R@10	R@20	# Samples
E-VQA	1–10	20.1	41.7	50.2	58.9	3.8K
	10–50	21.0	41.6	51.2	59.1	35K
	20–50	23.0	43.2	52.7	60.2	26K
InfoSeek	1–10	56.4	76.2	81.9	86.3	66K
	5–50	55.9	76.2	82.1	86.5	48K
	10–50	56.1	76.9	82.7	86.9	37K
	20–50	54.9	75.8	81.9	86.3	24K

Table 25: Effect of the retrievability-driven data screening strategy. “Strategy” denotes the rank interval used to select training samples.

On InfoSeek, the 10–50 interval provides the best overall performance, achieving the highest R@5, R@10, and R@20 scores of 76.9, 82.7, and 86.9, respectively. Although its R@1 score of 56.1 is slightly lower than the 56.4 obtained by the 1–10 interval, the difference is only 0.3 points. Meanwhile, the 10–50 strategy retains 36K samples and consistently outperforms the more restrictive 20–50 strategy. Overall, these results show that an appropriate screening interval is necessary to balance sample difficulty, quality, and quantity. We therefore adopt 20–50 for E-VQA and 10–50 for InfoSeek in our subsequent experiments.

Overall, these results demonstrate that moderately difficult samples provide a better balance between training data quality and quantity than either overly easy samples or excessively restrictive screening.

System:

You are a Wiki Summary Generator Assistant. Following is some information about you:

Profile

- name: Wiki Summary Generator Assistant
- language: English
- description: The Wiki Summary Generator Assistant is designed to create concise and informative summaries based on provided Wikipedia content. It extracts key aspects of the entity mentioned in the Wiki article, covering various dimensions such as history, characteristics, significance, appearance and impact.

Workflows

1. Input the provided Wikipedia content into the system.
2. Identify the main sections and key information related to the entity.
3. Synthesize this information into a well-structured summary.
4. Review and refine the summary for clarity, coherence, and completeness before finalizing.

Rules

1. Focus on summarizing key details across multiple aspects (e.g., appearance, features, impact) of the entity.
2. Ensure the summary is concise, clear, and free of irrelevant details.
3. Retain the original meaning and context of the Wiki content while rephrasing it into a summary.

User:

Following is the input Wikipedia content:

{ *Wikipedia Content* }

Based on the above Wikipedia content, I would like you to generate a summary of the Wikipedia content. Here is the summary of the Wikipedia content:

Table 26: Prompt template used to instruct the LLM to generate summaries offline for entity knowledge

System:

Answer the encyclopedic question about the given image. Don't mention the visual content of image in your output. Directly output the answer of the question according to the context.

If the context does not contain the information required to answer the question, you should answer the question using internal model knowledge.

User:

{ Query Image }

- Context: { Entity section }

- Question: { Textual question }

The answer is:

Table 27: Prompt used for the VQA process of MLLM in E-VQA.

System: Answer the encyclopedic question about the given image. Don't mention the visual content of image in If you need to answer questions about numbers or time, please output the corresponding numerical format directly. If the context does not contain the information required to answer the question, you should answer the question using internal model knowledge. There is an example: - Context: # Wiki Article: Dolomites ## Section Title: Dolomites The Dolomites, also known as the Dolomite Mountains, Dolomite Alps or Dolomitic Alps, are a mountain range located in northeastern Italy. The Dolomites are located in the regions of Veneto, Trentino-Alto Adige/Südtirol and Friuli Venezia Giulia, covering an area shared between the provinces of Belluno, Vicenza, Verona, Trentino, South Tyrol, Udine and Pordenone. - Question: Which city or region does this mountain locate in? Just answer the questions , no explanations needed. Short answer is: Province of Belluno User: { Query Image } - Context: { Entity section } - Question: { Textual question } Just answer the questions , no explanations needed. Short answer is:
--

Table 28: Prompt used for the VQA process of MLLM in InfoSeek.