OMGM: Orchestrate Multiple Granularities and Modalities for Efficient Multimodal Retrieval

Wei Yang*, **Jingjing Fu**†, **Rui Wang, Jinyu Wang, Lei Song, Jiang Bian**
Microsoft Research Asia
wyang6621@gmail.com, {jifu,ruiwa,jinywan,lesong,jiabia}@microsoft.com

Abstract

Vision-language retrieval-augmented generation (RAG) has become an effective approach for tackling Knowledge-Based Visual Question Answering (KB-VQA), which requires external knowledge beyond the visual content presented in images. The effectiveness of Vision-language RAG systems hinges on multimodal retrieval, which is inherently challenging due to the diverse modalities and knowledge granularities in both queries and knowledge bases. Existing methods have not fully tapped into the potential interplay between these elements. We propose a multimodal RAG system featuring a coarse-to-fine, multi-step retrieval that harmonizes multiple granularities and modalities to enhance efficacy. Our system begins with a broad initial search aligning knowledge granularity for cross-modal retrieval, followed by a multimodal fusion reranking to capture the nuanced multimodal information for top entity selection. A text reranker then filters out the most relevant fine-grained section for augmented generation. Extensive experiments on the InfoSeek and Encyclopedic-VQA benchmarks show our method achieves state-of-the-art retrieval performance and highly competitive answering results, underscoring its effectiveness in advancing KB-VQA systems. Our code can be found at https://github.com/ChaoLinAViy/OMGM.

1 Introduction

Visual Question Answering (VQA) involves answering questions about a given query image by comprehensively understanding its semantic content, demanding proficiency in both visual and textual understanding. Large Language Models (LLMs) have demonstrated remarkable generalization and reasoning capabilities in text-based

*Work done during an internship at Microsoft Research Asia.
†Corresponding author

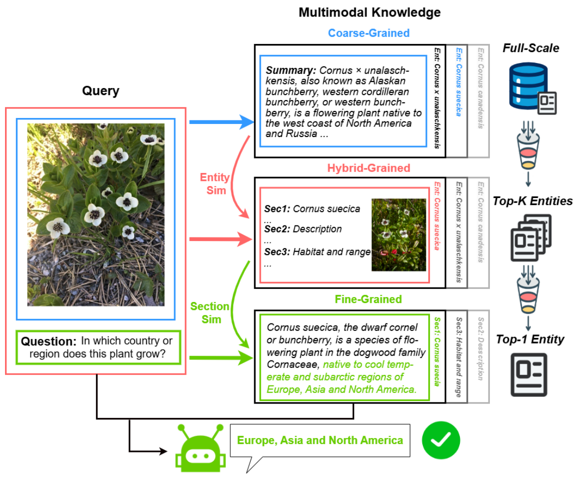

Figure 1: An illustration of our method. First, a coarse-grained cross-modal entity search is performed between entity summaries and the query image to retrieve the top-k entity candidates. Next, a hybrid-grained multimodal fusion reranker uses the multimodal query to retrieve image-section pairs, refining the selection of the most relevant entity. Finally, within knowledge associated with the top-1 entity, fine-grained textual filtering is applied to extract most relevant section, which is used to enhance generation in downstream generator.

tasks (Raffel et al., 2020; Brown et al., 2020). By integrating visual encoders with LLMs, Multimodal Large Language Models (MLLMs) have emerged as an effective approach for handling VQA tasks, as they can jointly model both image and text representations for enhanced comprehension and reasoning (Alayrac et al., 2022; Liu et al., 2024; Li et al., 2023b). Knowledge-Based Visual Question Answering (KB-VQA) extends this challenge by requiring the incorporation of external world knowledge that transcends the visible elements of the image. In KB-VQA, questions are designed to probe for information pertaining to the image's subject matter, but necessitate insights not directly present within the image itself.

Retrieval-Augmented Generation (RAG) offers a cost-effective and efficient solution to the challenges of KB-VQA by retrieving query-relevant

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 24545–24563

knowledge from the knowledge base and integrating it as contextual information for response generation (Karpukhin et al., 2020; Si et al., 2023). The effectiveness of this approach depends on efficient retrieval mechanisms capable of identifying the most relevant information from large-scale, heterogeneous knowledge bases. However, multimodal retrieval in KB-VQA introduces complexities beyond the standard text-based retrieval used in most RAG systems. These complexities stem from two key factors:

- **Multiple Modalities.** Both queries and knowledge bases comprise multiple modalities, such as images and text, necessitating diverse relevance assessment strategies that may involve single-modal, cross-modal, or multi-modal approaches. The choice of retrieval schema is influenced by the retrieval model's capabilities and the requirements of specific tasks.
- **Hybrid Granularities.** Queries and knowledge bases often involve information at different levels of granularities. For instance, a query might include a query image with coarse-grained knowledge identifying a subject, paired with a question seeking fine-grained details about the subject. Similarly, a knowledge base might consist of articles with coarse-grained overview reference images and titles as well as fine-grained detailed sections containing in-depth information.

Recent research efforts have explored both single-step and multi-step multimodal retrieval strategies to enhance retrieval effectiveness and ultimately improve answer generation. Single-step approaches retrieve passages directly using a multimodal query (Lin and Byrne, 2022; Deng et al., 2025; Jian et al., 2024; Lin et al., 2024). While effective, these methods often require expensive task-specific pretraining and incur high computational costs during inference due to exhaustive full-range searches. Conversely, multi-step retrieval methods adopt hierarchical retrieval strategies that enhance searching efficiency by progressively narrowing the search space. These methods often employ different retrieval modalities at each step, enabling the evaluation of knowledge relevance from multiple perspectives (Caffagni et al., 2024; Yan and Xie, 2024; Qi et al., 2024). Nevertheless, the potential of multi-step retrieval remains underexplored. Current approaches frequently overlook the intricate interplay between retrieval modalities, granularities, and the sequencing of retrieval steps, limiting their overall effectiveness and adaptability.

We propose a multimodal RAG system, **OMGM**, which employs a coarse-to-fine, multi-step retrieval strategy to effectively **O**rchestrate **M**ultiple **G**ranularities and **M**odalities across queries and knowledge bases, enhancing multimodal retrieval. As illustrated in Figure 1, our system operates in three stages: it begins with a coarse-grained cross-modal retrieval to identify an initial pool of entity article candidates, followed by a hybrid-grained multimodal reranker that leverages both coarse-grained and fine-grained knowledge to rerank candidates and select the most relevant entity. Finally, a fine-grained text reranker filters the selected entity's sections to extract the most pertinent sections for augmented response generation. Throughout the process, query and candidates are aligned based on their granularities, and embedding models are carefully selected to ensure effective multimodal representation. More importantly, these retrieval steps interact sequentially, with similarity scores from earlier stages propagated forward and fused in subsequent steps, enabling a cohesive and context-aware retrieval process. We conduct extensive retrieval and VQA experiments on two KB-VQA datasets, where our method achieved state-of-the-art retrieval performance and competitive question-answering results compared to other existing multi-step retrieval methods.

In summary, the main contributions of our work can be summarized as follows:

- We propose that multimodal retrieval should be tailored to the characteristics of KB-VQA, specifically in data modality and knowledge granularity. To achieve this, we introduce a coarse-to-fine multi-step retrieval strategy that progressively enhances retrieval quality by leveraging retrieval steps with varying modalities and granularities.
- We introduce a trainable multimodal reranker to maximize the utilization of full-modal information while minimizing inference costs by restricting the reranking scope, thereby ensuring both effectiveness and efficiency.
- We conducted extensive experiments on the InfoSeek and Encyclopedic-VQA (E-VQA) benchmarks, showcasing the effectiveness of the proposed method. Furthermore, a comprehensive ablation study validates the contribution of each retrieval step, offering valuable in-

sights for designing multimodal retrieval systems in KB-VQA tasks.

2 Related Work

2.1 KB-VQA

Traditional VQA tasks focus primarily on the visual content within images to answer the related textual question (Antol et al., 2015). However, KB-VQA expands this by incorporating external knowledge bases to address questions requiring information beyond the image. Datasets such as OK-VQA and A-OKVQA (Marino et al., 2019; Schwenk et al., 2022) involve visual question answering that require outside knowledge, which consists of general commonsense that lacks distinctiveness.

The emergence of the E-VQA and InfoSeek (Mensink et al., 2023; Chen et al., 2023) datasets presents a greater challenge for end-to-end LLMs/MLLMs by incorporating varying granularities of encyclopedic knowledge alongside extensive multimodal information. In response to these characteristics of KB-VQA task, our work adopts a multi-step, granularity-aligned retrieval framework, resulting in improved retrieval and VQA performance compared to the current multimodal RAG system (Caffagni et al., 2024; Lerner et al., 2024; Yan and Xie, 2024).

2.2 Vision-language RAG

Retrieval-Augmented Generation (RAG) (Guu et al., 2020) enhances the generation performance of LLMs by retrieving external documents relevant to the input query and using them as guiding context in prompts. In addition to text augmentation, recent work (Lin and Byrne, 2022; Xia et al., 2024) has targeted generation enhancement specifically for vision-language tasks, which are more closely aligned with complex real-world scenarios.

Vision-language retrieval-augmented generation typically entails handling multimodal documents and queries, requiring the retrieval process across these modalities. For example, the approaches introduced in PreFLMR and MuKA (Lin et al., 2024; Deng et al., 2025) encodes features across various modalities and dimensions, and separately concatenates the retrieval matrices for the query and the candidate, thereby facilitating a fine-grained knowledge search. Wiki-LLaVA (Caffagni et al., 2024) and EchoSight (Yan and Xie, 2024) employ a hierarchical retrieval strategy, achieving efficient cross-step multimodal retrieval. LLM-RA (Jian et al., 2024) and RoRA-VLM (Qi et al., 2024) adopt LLM-based and similarity-based approaches, respectively, for fine-grained denoising on queries and knowledge, thereby improving the accuracy of both retrieval and question answering. Additionally, mR^{2}AG (Zhang et al., 2024) and ReflectiVA (Cocchi et al., 2024) emphasize reflective processing. By fine-tuning large models, they leverage token outputs to drive the retrieval process and subsequently perform relevance-based re-screening and modifications on both the retrieved content and the generated answers. Unlike previous methods, our system constructs a multi-step retrieval process that integrates granularity alignment with cross-step joint retrieval. At each step, we align queries with the corpus using techniques such as summary extraction, multimodal fusion, and question-based section denoising, then fuse results via similarity fusion to boost recall and enhance downstream KB-VQA performance of MLLMs/LLMs.

3 Methodology

To address the challenges in KB-VQA, we introduce a multimodal RAG system that features efficient coarse-to-fine, multi-step multimodal retrieval. This system is capable of extracting pertinent information from a vast multimodal knowledge base with millions of entries. The retrieved data is subsequently used to enhance the generator's responses. An overview of the proposed framework is shown in Figure 2, which comprises three key components: coarse-grained cross-modal entity searching, hybrid-grained multimodal-fused reranking, and fine-grained section-augmented generation.

3.1 Coarse-Grained Cross-Modal Entity Searching

To preliminarily filter out wiki entity information related to the query from a large corpus, we design a coarse-grained entity retrieval method. In this step, the query image indicating a subject serves as a coarse-grained query, while entity summaries act as coarse-grained candidates, ensuring appropriate information granularity alignment.

Summary Generation. Since the complete wiki entity articles are too redundant as entity information for effective retrieval indexing, we align them with the macro-level entity information of the query image. Specifically, we use article summaries as the retrieval index for wiki entities, these concise

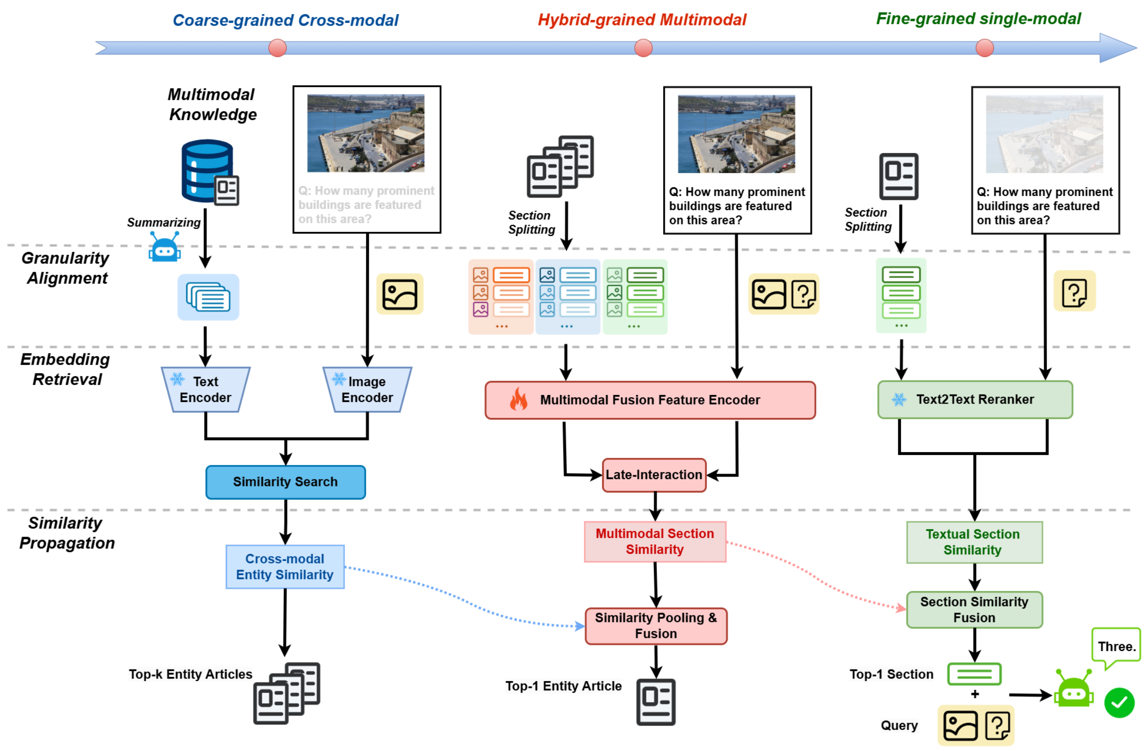

Figure 2: An overview of our framework. Our framework first performs a coarse-grained, cross-modal entity search by using offline-generated knowledge summaries as indices and a query image as the query; it then conducts a hybrid-grained multimodal-fused reranking that aligns image and text details via feature fusion and integrates cross-step similarities. Finally, a fine-grained section-augmented generation selects the most relevant knowledge section through combined text and multimodal similarities to support answer generation.

abstracts densely encapsulate key aspects of the entity's information. To maintain real-time retrieval efficiency, we generate summaries for all entity articles in the knowledge base offline. Given an entity article a_i, we employ a pre-trained language model M_s with an instruction prompt template P to generate a summary s_i, ensuring that the summary is both informative and well-aligned with the retrieval task.

$$s_i = M_s(P, a_i) \tag{1}$$

Image-to-Summary Entity Searching. After obtaining the wiki entity summaries, we use the query image from the VQA triplet, uniquely representing the target entity, as the query object for entity searching. Specifically, both the query image and candidate entity summaries are transformed into feature vectors, which are used for similarity-based matching for entity searching.

$$Ent_k = F(E_v(I_q), E_t(S), k) \tag{2}$$

where I_q denotes the query image, and $S = s_1, s_2, ..., s_n$ represents all entity summaries in the knowledge base. These are encoded through the visual encoder E_v and textual encoder E_t of CLIP (Radford et al., 2021), respectively. We use the Faiss library F (Johnson et al., 2019) to index the feature vectors of the entity summaries and perform embedding matching based on the inner product. Finally, we retain the top-k most relevant entity summaries along with their corresponding entity information Ent_k.

3.2 Hybrid-Grained Multimodal-Fused Reranking

After the initial search, we obtain the top-k entity candidates most similar to the query subject. Within this candidate set, we extract hybrid-grained multimodal fusion feature matrices from the coarse-grained images and fine-grained texts provided by the query and candidate knowledge. Then, by employing a late-interaction mechanism, we obtain a fine-grained section similarity. By integrating the entity similarity from the previous step, we derive a coarse-grained reranking similarity that yields a more accurate ordering of relevant entities.

Multimodal Fusion Feature Matching. Specifically, we leverage the Q-Former (Li et al., 2023b) architecture to extract multimodal fusion features from both query and candidates. The pipeline for fusion feature extraction is as follows:

$$Q = E_m(q), C_{sec_e^h} = E_m(d_e^h) \tag{3}$$

As the input to the q-former multimodal encoder E_m, we use the query image I_q and textual question T_q on the query side, forming the input pair $q = (I_q, T_q)$. On the candidate side, we construct the input as $d_e^h = (I_e, sec_e^h)$, where I_e is the main image of entity and sec_e^h is the h-th section of the entity article $a_e = \{sec_e^1, sec_e^2, \ldots, sec_e^p\}$. The fusion feature matrix Q and $C_{sec_e^h}$ are the output vectors corresponding to the query tokens of encoder E_m. We calculate the fine-grained multimodal similarity $sim_m^{sec_e^h}$ between the query q and candidate d_e^h using the late-interaction (Khattab and Zaharia, 2020), which fully leverages the correlations between each token vector in the feature matrix and integrates them into a single similarity score through a Max-Sum operation. l_Q and l_C denote the total number of tokens in Q and $C_{sec_e^h}$, respectively.

$$sim_m^{sec_e^h} = sim_m(q, d_e^h) = \sum_{i=1}^{l_Q} \max_{j=1}^{l_C} Q_i C_{sec_e^h}^{j\ \top} \tag{4}$$

Finally, we obtain the coarse-grained entity similarity by maximizing sim_m of all the sections in the entity article a_e, with section size l_{a_e}. We then compute the final reranking similarity scores for the top-k entities Ent_k by performing a weighted summation with the initial entity similarity sim_c^e from the previous cross-modal retrieval step. The most relevant entity e_{top1} is selected based on the highest reranking score. The weighting factor α balances the contributions of the two similarity measures, optimizing reranking performance.

$$e_{top1} = \arg\max_{e \in Ent_k} (\alpha \cdot sim_c^e + (1-\alpha) \cdot \max_{h=1}^{l_{a_e}} sim_m^{sec_e^h}) \tag{5}$$

Multimodal Reranker Training. To train the multimodal fusion encoder, we employ contrastive learning with the hard negative samples from previous retrieval step. Specifically, the coarse-grained retrieval in earlier step generates a top-k candidate entity set for each training sample. Most of these entities share similar coarse-grained characteristics, but differ in fine-grained details. To construct training pairs, we randomly select negative pairs by pairing the main images of candidate entities with non-evidentiary sections from their articles. In contrast, positive pairs consist of the main image and the evidence section of the correct entity.

Datasets	#Samples				#Entity Articles	
	Gen. Train	Ret. Train	Valid	Test	Train	Valid/Test
Infoseek	100K	-	-	71,335	100K	100K
E-VQA	100K	191k	11,696	4,750	2M	2M

Table 1: Statistics of Infoseek and E-VQA datasets used in our experiments. Gen. Train and Ret. Train represent the number of training samples for the generator and the multimodal fusion module in reranking, respectively.

$$\mathcal{L} = -\log \frac{\exp(sim_m(q, d_+)/\mathcal{T})}{\sum_{j=1}^{N} \exp(sim_m(q, d_j)/\mathcal{T})} \tag{6}$$

By constructing each training sample containing N contrastive pairs d, we train the reranker to match queries with the positive candidate pairs d_+ based on hybrid-grained multimodal information with the adaptive temperature $\mathcal{T}$ for the smoothness of the softmax distribution. This multimodal-fused retrieval approach not only complements the coarse-grained entity retrieval in step one but also paves the way for subsequent fine-grained refinement.

3.3 Fine-Grained Section-Augmented Generation

After completing the first two steps of entity ranking, we identify e_{top1} as the entity most relevant to the query. In this step, we perform fine-grained knowledge filtering on the entity information and use the filtered knowledge as auxiliary context to enhance the generation of the downstream generator. Specifically, we employ a pre-trained textual reranker R_t to calculate the textual similarities $sim_t^{sec} = R_t(T_q, sec)$ between sections of the entity article and question. These textual similarities are then combined with the fine-grained multimodal-fused similarity sim_m^{sec} obtained from the previous step through weighted summation, enabling us to extract the most relevant entity section $sec_{e_{top1}}^{best}$ for the query.

$$sec_{e_{top1}}^{best} = \arg\max_{sec \in e_{top1}} (\beta \cdot sim_m^{sec} + (1-\beta) \cdot sim_t^{sec}) \tag{7}$$

Here, β serves as a balancing factor to weigh the similarities derived from retrieval at different steps. Finally, the most relevant section is provided as input context along with the query to the generator for question-answering.

4 Experiments

4.1 Datasets

We utilize two challenging KB-VQA datasets, InfoSeek and E-VQA (Chen et al., 2023; Mensink et al., 2023), for training and testing. To ensure fairness in both retrieval and VQA evaluations, we adopt the same setup used by many previous studies, with the specific configuration detailed in Table 1. For the images associated with the entity articles in the knowledge base, we crawled multiple images, including the main image, from each entity's corresponding wikipedia page to constitute the knowledge base's image set. Our experiments primarily focus on assessing both retrieval and question-answer performance: retrieval is evaluated using Recall@K, while VQA performance is measured using the official metrics for each dataset (e.g., BEM score (Zhang et al., 2019) for E-VQA and both VQA accuracy (Antol et al., 2015) and relaxed accuracy (Methani et al., 2020) for InfoSeek). More details on the datasets and their evaluation methods can be found in the Appendix B.

4.2 Implementation Details

In this section, we briefly introduce some details of the various steps in our framework. Regarding the models and prompts used, please refer to Appendix C.1 and Appendix A, respectively.

Initial Entity Searching. To produce high-quality summaries of entity articles, we leveraged LangGPT (Wang et al., 2024) to develop the "Wiki Summary Generator Assistant" prompt. For efficient entity search using the FAISS library, we employed the pooled embeddings from the last layer of encoder for calculating image-text similarity. We set k to 20 and retrieved only the top-20 most relevant entities to strike a balance between the speed and effectiveness of subsequent retrieval and question-answering tasks, as validated by the experimental results presented in the Table 7.

Multimodal-Fused Encoder Training and Inference. The encoder model is initialized with pre-trained q-former weights using the LAVIS Library (Li et al., 2023a). We select the top 32 embeddings from the model output as our multi-modal fusion feature matrix, corresponding to the position and number of defined query tokens. Since InfoSeek training samples do not include labels for the evidence section, our encoder is trained on the E-VQA training set and then tested on both datasets. This setting aligns with the training requirements and characteristics of the datasets, while also allows us to evaluate the model's generalization ability. Given that each question-answer pair in the training samples corresponds to multiple query images, we use only the first query image to form the triplet with the question-answer pair as the actual training sample. This strategy ensures training quality while enhancing efficiency. Additional details regarding this step can be found in Appendix C.2. Results of ablation experiments on the hyperparameter for cross-stage similarity propagation are presented in Appendix D.2.

Generator Training and Inference. we primarily employ pre-trained LLM/MLLM models for ablation studies and most of the main experiments. To assess the performance of the generator fine-tuned under our retrieval system, we experiment with lightweight and efficient fine-tuning on LLaVA-1.5-7B (Liu et al., 2024) and evaluate its VQA performance. More details are presented in Appendix C.3.

4.3 Main Results

The results of our method compared with other works are presented in Tables 2 and Table 3, primarily on entity retrieval and VQA performance.

Retrieval Result. Table 2 showcases the retrieval performance of various multimodal RAG approaches across two datasets. The "CLIP I-T" refers to the naive approach, where CLIP is used for the cross-modal similarity search, linking the query image to the wiki article with embeddings. By examining the Recall@1 results, our method (w. reranking) significantly outperforms other methods, which demonstrates the robust retrieval capability of our retrieval method. Additionally, the reranking in step two improves Recall@1 by 23.7% and 11.4% for the two datasets compared to step one alone, underscoring the effectiveness of multimodal retrieval and cross-step similarity propagation in boosting reranking performance. Furthermore, even without reranking, our method outperforms the full-scale retrieval strategies commonly used in other works across all retrieval ranges, demonstrating the significant advantage provided by the granularity alignment between the summary and the query image in step 1.

VQA Result. Table 3 presents a comparison of our VQA results with the state-of-the-art methods. RORA-VLM (Qi et al., 2024) is a retrieval-

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	-	<u>49.8</u>	<u>56.1</u>	<u>77.6</u>	-	**86.4**
EchoSight								
w/o. reranking	13.3	31.3	41.0	48.8	45.6	67.1	73.0	77.9
w. reranking	<u>36.5</u>	<u>47.9</u>	48.8	48.8	53.2	74.0	77.4	77.9
OMGM (ours)								
w/o. reranking	19.1	41.2	<u>49.8</u>	58.7	52.6	73.9	<u>80.0</u>	84.8
w. reranking	**42.8**	**55.7**	**58.1**	**58.7**	**64.0**	**80.8**	**83.6**	<u>84.8</u>

Table 2: Retrieval results on the E-VQA test set and InfoSeek validation set. "w/o. reranking" and "w. reranking" represent the entity retrieval results after step one and step two, respectively. Best in bold, second-best underlined.

Method	Generator Model	Gen. FT	Ret. FT	E-VQA	InfoSeek		
					Unseen-Q	Unseen-E	Overall
RoRA-VLM	LLaVA-1.5-7B	✓	✗	20.29	27.34	25.10	-
Wiki-LLaVA	LLaVA-1.5-7B	✓	✗	21.8	30.1	27.8	28.9
LLM-RA	BLIP2-Flan-T5XL	✓	✓	-	26.12	20.90	23.14
EchoSight	Mistral-7B | LLaMA3-8B	✗	✓	41.8	-	-	31.3
mR^{2}AG	LLaVA-1.5-7B	✓	✓	-	<u>40.6</u>	<u>39.8</u>	<u>40.2</u>
ReflectiVA	LLaVA-MORE-8B	✓	✓	35.5	40.4	39.8	40.1
OMGM (ours)	InternVL-2.5-8B	✗	✓	48.72	37.16	35.1	36.1
	LLaMA3-8B	✗	✓	<u>49.94</u>	35.26	33.61	34.42
	LLaVA-1.5-7B	✓	✓	**50.17**	**43.46**	**43.53**	**43.49**

Table 3: VQA accuracy comparison with the baselines. Gen. FT and Ret. FT indicate whether the generator and retriever of the method were fine-tuned, respectively. Best in bold, second-best underlined.

augmented VLM system that removes irrelevant information based on token-level embedding similarity and incorporates noise-resilient retrieval-augmented training. mR^{2}AG (Zhang et al., 2024) and ReflectiVA (Cocchi et al., 2024) center on reflection. By fine-tuning large models, they leverage token outputs to drive the retrieval process and perform re-screening and modifications on retrieved content and generated answers. Our method achieves superior VQA results on both datasets compared to existing approaches, demonstrating the improved generation performance of downstream models after enhancing retrieval capabilities. Notably, our method, which only fine-tunes the retriever, outperforms most approaches that fine-tune downstream generators, highlighting its efficiency. We test the VQA performance of mainstream LLMs and MLLMs under zero-shot settings and our framework on two datasets and provide results in Appendix D.1. Our method demonstrates excellent VQA results across both MLLM and LLM models, underscoring its ability to generalize retrieval optimization to enhance generation across different types of downstream generators.

4.4 Ablation Study

We conduct extensive experiments to assess the effectiveness of our framework and the design of each step, focusing on retrieval and question-answering performance. Due to space limitations, we have placed additional ablation experiments in Appendix D.2 and Appendix D.3, and the results of some ablation experiments on InfoSeek can be found in Appendix D.4.

Impact of Retrieval Steps on VQA Results. Table 4 records the VQA results on two datasets, as each step of our framework is executed in sequence. When only step one is completed, the generator receives the top-1 entity article as context. After step two is executed, the generator's context is provided with the top-1 section. Each step in our multi-step framework progressively improves the VQA performance of the downstream generator, confirming the effectiveness of its design. Notably, step two offers the most significant enhancement, highlighting the powerful retrieval capability of our multimodal-fused reranker with similarity propagation.

step-1	step-2	step-3	E-VQA	InfoSeek
✗	✗	✗	17.2	9.3
✓	✗	✗	25.45	18.87
✓	✓	✗	39.81	31.38
✓	✓	✓	41.81	33.29

Table 4: The ablation study on the impact of different steps in our framework on the VQA results, using LLaVA-1.5-7B as the answer generator.

Impact of Retrieval Design on Initial Entity Search. As the first step of multi-step retrieval, we perform a coarse-grained search for entity-related knowledge within a large knowledge base. We use the query image as the query, which uniquely contains entity information. Retrieval candidates can be selected based on different modalities and knowledge granularities. Table 5 presents our retrieval experiments using images, articles, titles, and summaries as candidates. The results indicate that the Image-to-Summary method achieves the best retrieval performance, suggesting that summaries offer better alignment with query images in terms of information granularity, leading to improved retrieval outcomes.

Ret. Method	R@1	R@5	R@10	R@20
$Image \rightarrow Article$	13.2	27.7	35.5	41.7
$Image \rightarrow Image$	13.4	31.8	41.9	48.8
$Image \rightarrow Title$	17.5	31.9	38.6	44.8
$Image \rightarrow Summary$	**19.1**	**41.2**	**49.8**	**58.7**

Table 5: Ablation study on modalities and granularities design for entity retrieval in step 1.

Effect of Multimodal Fusion Reranking. In step two, we rerank a small subset of entities obtained from the initial search to identify the most relevant one to the query. Consequently, we must leverage question-oriented, fine-grained textual information, such as the query question and the sections associated with each entity. In VQA tasks, both the query and the knowledge base are inherently multimodal. Therefore, it is reasonable to consider incorporating multi-modal fusion features into the retrieval process. Based on this rationale, we experimented with four retrieval approaches: purely text-based retrieval, multimodal-to-text retrieval, text-to-multimodal retrieval, and fully multimodal retrieval. Using Q-Former as the base encoder model, we fine-tune all four retrieval modality approaches under the same training configuration and evaluated their performance. The results presented in Table 6 indicate that the method employing multimodal fusion features on both the query and candidate sides achieve the best retrieval performance. This finding directly demonstrates the effectiveness and comprehensiveness of the multimodal-fused reranking design in our framework.

Ret. Modality	Sec. R@1	R@1	R@5	R@10
$T \rightarrow T$	24.6	30.7	51.8	57.4
$(I, T) \rightarrow T$	22.5	28.7	51.3	57.0
$T \rightarrow (I, T)$	24.3	30.3	51.1	56.9
$(I, T) \rightarrow (I, T)$	**32.8**	**40.2**	**54.8**	**57.8**

Table 6: Ablation study on retrieval modality of step two on E-VQA. "I" indicates image modality and "T" indicates text modality. "Sec. R@1" refers to the recall of the top-1 section. α is set to 0 for direct comparison.

Effect of Reranking Scope on Retrieval Performance. The reranking scope k influences both the number of entities filtered during the initial search and the range of multi-modal fused features extraction during reranking. Table 7 presents the retrieval results and the average retrieval time for the first two steps with varying k values. Our framework can consistently improve the retrieval capacity while increasing k from 10 to 100, accompanied with more retrieval time as well. To balance retrieval quality and efficiency, we set the reranking scope k to 20 for experiments across both datasets.

k	Sec. R@1	R@1	R@5	R@10	R@20	Time
10	31.6	39.0	48.8	49.8	-	0.630
20	34.7	42.8	55.7	58.1	58.7	1.110
50	37.4	45.9	61.3	64.6	66.7	2.420
100	38.8	47.4	65.0	69.4	72.2	4.642

Table 7: The effect of the retrieval scope K on the retrieval results and time after step two on E-VQA.

Generalization on Additional Benchmark with Differen Document Structures. For E-VQA and InfoSeek documents, which consist of entity-based Wikipedia articles, we segmented them into sections based on their inherent structure. This method allowed us to leverage the natural organization of Wiki-style content, which often includes headings and subheadings, to create meaningful sections for retrieval. For KB-VQA datasets with document structures that differ significantly from E-VQA or InfoSeek, alternative segmentation strategies can be employed to ensure effective organization of knowledge. One potential approach is to use rule-based method, which relies on predefined heuristics such as paragraph breaks, headings, or specific keywords to define section boundaries. Another approach is semantic clustering, which groups text segments based on semantic similarity, enabling the creation of fine-grained, section-like knowledge units even in the absence of explicit document structure.

To evaluate the generalizability of OMGM on different segmenation setting, we tested our method on the OK-VQA(Marino et al., 2019) dataset, which lacks a highly structured format and segmented using a rule-based segmentation method. As shown in Table 8, OMGM demonstrated strong performance, achieving higher retrieval accuracy (Pseudo Recall@5) and VQA scores compared to the strong baseline PreFLMR(Lin et al., 2024). Specifically, OMGM achieved a Pseudo Recall@5 of 73.4 and a VQA score of 66.57, outperforming PreFLMR by a notable margin. These results suggest that OMGM's framework is robust and adaptable to varying document structures across different KB-VQA datasets.

Method	Pseudo Recall@5	VQA score
PreFLMR	70.9	61.88
OMGM	**73.4**	**66.57**

Table 8: Comparison of the retrieval and VQA results of OMGM and PreFLMR on OK-VQA.

Quantitative evaluation on the efficiency of OMGM's step-by-step approach. To further perform the efficiency of the proposed OMGM framework, we conducted a comparative evaluation of OMGM, the one-step multimodal RAG method PreFLMR(Lin et al., 2024), and the direct use of LLaVA-1.5-7B for VQA on the E-VQA dataset. We evaluated three key metrics: average retrieval time, average inference time, and VQA performance. It is worth noting that PreFLMR preprocesses and encapsulates all passage embeddings prior to inference, which reduces retrieval time during runtime to primarily consist of query embedding and similarity matching. To ensure a fair comparison, OMGM's retrieval time calculation also focuses on these two components.

Method	Avg. Ret. Time	Avg. Inf. Time	VQA Result
LLaVA-1.5-7B	-	1.432	17.00
PreFLMR	0.984	2.196	54.45
OMGM	**0.402**	**2.023**	**63.39**

Table 9: Comparison of average retrieval and inference time as well as VQA performance on E-VQA.

As shown in Table 9, OMGM achieves substantial improvements in VQA performance, surpassing both LLaVA-1.5-7B and PreFLMR. Despite employing a step-by-step retrieval strategy, OMGM maintains competitive inference efficiency. Compared to direct use of LLaVA-1.5-7B, OMGM delivers significantly better VQA results while introducing only a minimal increase in inference time.

Additionally, when compared to the one-step PreFLMR, OMGM demonstrates a notable reduction in retrieval time, decreasing from 0.98s to 0.4s. This improvement is due to its orchestrated retrieval process, which is specifically designed to optimize the integration of different modalities and knowledge granularities at each step, achieving an effective balance between retrieval efficiency and performance.

5 Conclusion

In this paper, we propose a RAG system with multi-step multimodal retrieval. By employing queries and candidates of appropriate modalities at each step, the system aligns the information granularities for better retrieval. Our system capitalizes on cross-step similarity propagation to enhance retrieval interactions and employs a multimodal-fused design to fully exploit the rich multimodal information present in queries and candidates. Experimental results on mainstream KB-VQA datasets show that our approach surpasses existing approaches in retrieval performance. The comprehensive retrieval pipeline enables pre-trained models and lightly fine-tuned models to outperform the systems heavily reliant on extensive fine-tuning, and we reveal the rationality and effectiveness of it by ablation studies. These findings provide valuable insights for designing effective multimodal retrieval systems tailored to KB-VQA tasks.

Limitations

Our RAG system focuses primarily on the design of the multimodal retrieval module. While it has demonstrated significant performance improvements, several limitations remain: 1) Multimodal knowledge bases often include not only coarse-grained main images of entities but also numerous fine-grained secondary images linked to specific sections. Our current method does not exploit these secondary images to optimize multimodal retrieval, presenting a promising area for future exploration.

2) Although our approach enhances VQA performance through improved retrieval, further exploration is needed to determine how generator models can more effectively utilize multimodal fusion features to enhance answer quality during the generation phase.

References

AI@Meta. 2024. Llama 3 model card.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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, Singapore. Association for Computational Linguistics.

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.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

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? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, Singapore. Association for Computational Linguistics.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Augmenting multimodal llms with self-reflective tokens for knowledge-based visual question answering. *arXiv preprint arXiv:2411.16863*.

Lianghao Deng, Yuchong Sun, Shizhe Chen, Ning Yang, Yunfeng Wang, and Ruihua Song. 2025. MuKA: Multimodal knowledge augmented visual information-seeking. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9675–9686, Abu Dhabi, UAE. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

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.

Pu Jian, Donglei Yu, and Jiajun Zhang. 2024. Large language models know what is key visual entity: An LLM-assisted multimodal retrieval for VQA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10939–10956, Miami, Florida, USA. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Paul Lerner, Olivier Ferret, and Camille Guinaudeau. 2024. Cross-modal retrieval for knowledge-based visual question answering. In *European Conference on Information Retrieval*, pages 421–438. Springer.

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. 2023a. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Toronto, Canada. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. 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 and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. 2024. PreFLMR: Scaling up fine-grained late-interaction multi-modal retrievers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5294–5316, Bangkok, Thailand. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-VQA: A Visual Question Answering Benchmark Requiring External Knowledge.

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.

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.

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. *arXiv 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, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge.

Qingyi Si, Yuchen Mo, Zheng Lin, Huishan Ji, and Weiping Wang. 2023. Combo of thinking and observing for outside-knowledge VQA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10959–10975, Toronto, Canada. Association for Computational Linguistics.

Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. 2024. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin 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.

Ming Wang, Yuanzhong Liu, Xiaoming Zhang, Songlian Li, Yijie Huang, Chi Zhang, Daling Wang, Shi Feng, and Jigang Li. 2024. Langgpt: Rethinking structured reusable prompt design framework for llms from the programming language. *Preprint*, arXiv:2402.16929.

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.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024. RULE: Reliable multimodal RAG for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093, Miami, Florida, USA. Association for Computational Linguistics.

Yibin Yan and Weidi Xie. 2024. 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.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, and 1 others. 2024. mr^{2}ag: 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 utilized for invoking LLM/MLLM in our methodology, encompassing processes such as LLM-based summary generation and VQA across various datasets. The prompt template used to instruct the LLM to generate summaries offline for entity knowledge is presented in Table 15. It is evident that the entire template is structured in an agent-like format. Under this format, we expect the LLM to strictly adhere to the instructions to efficiently produce high-quality entity summaries.

During the question-answering phase in step 3, we have designed corresponding prompt templates tailored to the characteristics of different datasets and the capabilities of LLM/MLLM. This is to ensure that the true efficacy of our method in question-answering is demonstrated as accurately as possible. As demonstrated by the prompt for E-VQA testing shown in the Table 16, when deploying the MLLM, simple visual constraints and a query image for the query are incorporated into the prompt. For the InfoSeek dataset, which employs the stringent exact match evaluation criteria, it is crucial to thoroughly guide the downstream generator to adhere to the prescribed output format in order to showcase genuine question-answering performance. Consequently, we have incorporated additional format instructions and a one-shot example from the training set into the prompt presented in Table 17.

B Dataset Details

Encyclopedic VQA (Mensink et al., 2023) The dataset encompasses approximately 221k question-answer pairs linked to 16.7k distinct fine-grained entities, with each entity represented by up to five images. The fine-grained entities and associated images are derived from the iNaturalist 2021 dataset and the Google Landmarks Dataset V2 (Van Horn et al., 2021; Weyand et al., 2020). Moreover, the dataset provides a controlled knowledge base derived from WikiWeb2M (Burns et al., 2023) with 2M Wikipedia articles with images, which contain evidences to support each answer. According to the number of reasoning steps required, the questions in the dataset can be divided into single-hop and two-hop questions. The dataset triplets are split into training, validation, and test subsets, containing 1 million, 13k, and 5,800 samples, respectively. For performance comparison with other related works, we also only use single-hop questions for training and testing. Therefore, we adopt the sample allocation method as shown in the Table 1.

To evaluate the performance of our proposed retrieval-augmented QA LLM framwork in E-VQA, we utilize the standard metric Recall@K and BEM score (Zhang et al., 2019) as metrics to evaluate its retrieval capability and question-answering capability, respectively. Recall@K evaluates the proportion of test samples whose top-k retrieved entities contain the correct entity, thereby reflecting the retrieval performance within the top-k scope. As the specific evaluation metric for the E-VQA dataset, the BEM score is obtained by comparing the predicted answer with the correct answers using a BERT model specifically fine-tuned for answer similarity assessment. This method correctly evaluates candidate answers that are valid but do not exactly match the reference answers in annotations, as opposed to common VQA metrics.

InfoSeek (Chen et al., 2023) The dataset comprises 1.3 million image-question-answer triplets, corresponding to approximately 11,000 visual entities from OVEN (Hu et al., 2023). There are 8.9K human-written visual info-seeking questions and 1.3M automated generated questions in InfoSeek. The triplets are partitioned into training, validation, and test sets, containing approximately 934k, 73k, and 348k samples, respectively. Due to the lack of ground truth for test split, our evaluation is conducted on the validation set. In particular, both the validation and test sets feature questions pertaining to unseen entities or queries that are not encountered during training. Additionally, the dataset includes a knowledge base consisting of 6 million Wikipedia entities. To be consistent with related works, like EchoSight (Yan and Xie, 2024), we utilize the subset of 100,000 entities, ensuring the inclusion of the 6,741 entities corresponding to the questions from the training and validation splits. When collecting images from Wikipedia pages for Wikipedia entities, we found that a very small portion of validation samples in the infoseek dataset corresponded to correct entities that lacked associated images. As a result, we filter out these samples and conduct evaluations on the remaining 71,335 validation samples, which still account for 96.9% of the original dataset and ensures that the final results are not significantly affected.

For the retrieval evalution, we retained Recall@K as the evaluation metric, consistent with E-VQA. Following the specific question-answering evaluation criteria of InfoSeek, we employed two different metrics based on the question types. For questions requiring string-based answers, such as entity names, we report accuracy using the VQA accuracy metric (Antol et al., 2015). This metric allows for multiple valid answers by considering slight variations in phrasing as correct. The model is evaluated based on whether its answer exactly matches any of these valid responses. For questions requiring numeric answers, we use relaxed accuracy (Methani et al., 2020), which considers an answer correct if it falls within an acceptable tolerance range around the ground truth.

C More Experimental Details

C.1 Model Checkpoints.

We adopted LLaMA-8B-Instruct (AI@Meta, 2024) as the summarization model, whose robust ability to follow instructions guarantees the reliability of the output, while its open-source nature and compact size ensure high efficiency for a massive knowledge base. With the Eva-CLIP encoder (Eva-CLIP-8B) (Sun et al., 2024), we extract the embeddings of the query images and entity summaries. When selecting the best section from the entity article in step 3, we used BGE-Reranker-v2-m3 (Chen et al., 2024a) for obtaining text similarity, which is a lightweight, efficient text reranking model that has been pre-trained for optimal performance. Regarding the answer generator, we tested not only LLMs such as LLaMA-8B-Instruct and Mistral-7B-Instruct-v0.2, but also MLLMs like LLaVA-1.5-7B and InternVL-2.5-8B (AI@Meta, 2024; Jiang et al., 2023; Liu et al., 2024; Chen et al., 2024b). Additionally, we further tested the performance of GPT-4 and GPT-4-o, which refer to GPT-4-1106 and GPT-4-o-2024-08-06, respectively.

C.2 More Details about Step 2 Experiments.

When employing hard negative sampling for multimodal-fused retriever training, we set $N = 16$ image-section pairs for one training sample, which contain only one positive pair and 15 negtive pairs. For the image in the positive pair, we offline select the most similar image to the query image from multiple wiki images associated with the evident entity, using Eva-CLP for similarity ranking. And we directly choose the first image of the entity article as the image for the negative pair. Among the 15 negative pairs, we select up to three harder negative pairs from other sections of the article that contains the evidence section. Through the above contrastive pairs construction method, we efficiently obtain positive pairs most relevant to the query image and question, as well as diverse and challenging negative pairs. During retrieval training, we use learning rate 1e-5 and batch size 8, training for 1 epoch on a total of 191k training samples processed from E-VQA. This configuration allow the training to complete in 11 hours on 1 Nvidia A100 (80G). In the reranking inference, we set the similarity fusion hyper-parameter α to 0.9, indicating that fine-grained multimodal similarity can be used to fine-tune the coarse-grained entity similarity, which can be valiated by the results as shown in Appendix D.2

C.3 More Details about Step 3 Experiments.

For the best section similarity fusion, we set β to 0.2 to incorporate multimodal information with a small weight into unimodal information, which can be proved by the results as presented in Appendix D.2. For the LLaVA-1.5-7B fine-tuning samples, we randomly selected 100k samples from the training set of each dataset and used our system's retrieval step to match the corresponding section paragraphs, thereby constructing training samples in the RAG format. Regarding the fine-tuning settings, we adopted the official LLaVA-1.5 LoRA training parameters, using a learning rate of 2e-5 and a batch size of 8x16. Lightweight training was performed for 1 epoch on 1 Nvidia A100 (80G).

Regarding the baseline details in our main results, we consider mR²AG and ReflectiVA to have fine-tuned their MLLM to perform a reranking-like relevance reflection on the retrieved content, and thus we treat them as having fine-tuned the retriever. Additionally, the E-VQA results for mR²AG were not adopted because they utilize the online Google Lens retrieval results provided by E-VQA, which differ from the settings in other works where the retrieval system is custom-designed, so no comparison is made.

D More Experiments and Ablation studys

D.1 Consistency of our work across LLMs and MLLMs.

As shown in Table 10, we tested the VQA performance of mainstream LLMs and MLLMs un-

Generator Model	W/OMGM	E-VQA	InfoSeek
LLaVA-1.5-7B	✗	17.2	9.3
	✓	41.81	33.29
InternVL-2.5-8B	✗	25.56	10.95
	✓	48.72	36.1
Mistral-7B	✗	17.71	0
	✓	48.08	17.14
LLaMA3-8B	✗	18.78	1.56
	✓	49.94	34.42
GPT-4	✗	18.23	-
	✓	50.88	-
GPT4-o	✗	35.92	38.35
	✓	51.18	42.09

Table 10: The ablation study on the impact of our method on the VQA results of different generator models. We use the Overall Score as the primary metric for the VQA results on the InfoSeek dataset. Regarding the performance variations of closed-source models on InfoSeek, due to the vast number of samples in the InfoSeek test set and the cost constraints of API calls, we only evaluated the most critical model, GPT4-o.

der zero-shot settings and our framework on two datasets. Firstly, from the perspective of closed-source and open-source models, we can observe that our framework enhances the performance of small open-source models to approach or even surpass that of powerful closed-source models in a zero-shot setting (some cases are shown in Appendix E), while also significantly improving the VQA performance of closed-source models, which demonstrates the generalizability of our RAG system for generation models. Additionally, we observed that for LLMs, if the retrieval performance is sufficiently good, their question answering results can approach or even surpass those of MLLMs (49.94 for LLaMA3-8B and 48.72 for InternVL-2.5-8B in E-VQA), when query images are not accessible. This highlights the correct retrieved content about visual entity can help LLMs in analyzing and answering visual questions.

D.2 Effect of different hyper-parameters' value in step 2 and 3

In this section, we investigate the impact of the hyperparameters α and β, which control the similarity mixing in step 2 and 3 of our method, on the retrieval and question-answering results.

In step 2, we set α as the weight to integrate the coarse-grained similarities obtained from the initial entity search into the multimodal-fused similarities, thereby achieving a comprehensive reranking similarity. In the experiments with α, we tested its variation from 0 to 1 and observed the changes in the final retrieval entity recall of step 2, as illustrated in the Figure 3.

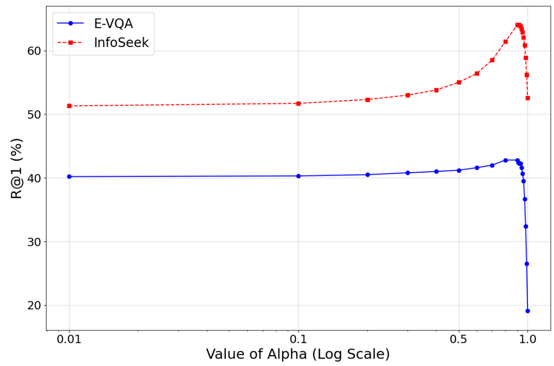

Figure 3: The variation of Entity Recall@1 with the change of step 2 similarity fusion hyper-parameter α on the E-VQA and InfoSeek

From the Figure 3, we can see that the fusion of similarities at different steps has a very positive impact on the reranking effect. When α is set to 0, we solely rely on the similarity from step 2 for reranking, achieving relatively satisfactory results on both datasets (51.3% and 40.2%), highlighting the excellent entity reranking capability of the multimodal-fused reranker. Notably, for E-VQA, the retrieval outcomes in step 2 significantly surpass those of step 1, which is attributed to the fact that this dataset was used as the training set for the reranking encoder, demonstrating the remarkable effectiveness of multimodal fusion retrieval. Moreover, on InfoSeek, the approach exhibits strong generalizability, as combining similarities results in a substantial improvement in reranking performance (52.6% to 64.0%). The best retrieval performance is achieved when the similarity from step 1 is mixed at a higher proportion (0.9), indicating that the similarity of multimodal-fused retrieval optimizes the coarse-grained entity similarity obtained from the initial search in a "fine-tuning" manner.

In the step 3, we set β as the weight to integrate the multimodal-fused section similarity into the direct text similarity, aiming for better knowledge denoising. In the experiments with β, we tested its variation from 0 to 1 and observed the changes in the retrieval and question-answering effectiveness on two datasets in step 2, as presented in Figure 4 and Figure 5.

Compared to the significant impact of α on re-

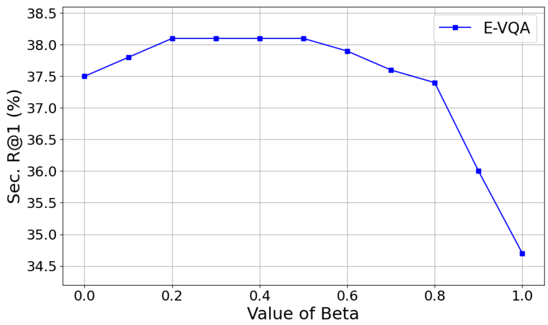

Figure 4: The changes in Section Recall@1 on the E-VQA dataset under varying values of the step 3 similarity fusion hyper-parameter β

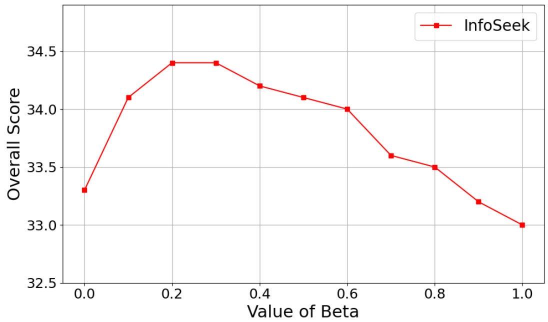

Figure 5: The changes in Overall Score on the InfoSeek dataset under varying values of the step 3 similarity fusion hyper-parameter β

trieval, the positive impact of β is relatively limited. This is because the knowledge denoising in step 3 is conducted in a top-1 document context, without involving entity-level filtering, so simply using a pretrained text reranker can achieve good denoising results. The performance improvement obtained by integrating the multimodal fusion similarity at a small proportion also demonstrates the comprehensiveness and generalization ability of cross-step similarity propagation.

D.3 Methods for computing multimodal fusion feature similarities

In order to achieve the best retrieval performance, we need to adopt an appropriate similarity computation method to match the multimodal fusion feature matrix. Based on previous studies (Yan and Xie, 2024; Radford et al., 2021; Khattab and Zaharia, 2020), we evaluated three commonly used methods for computing similarity in Table 11 and Table 12. Q-Former's Image-to-Text Correspondence (ITC) computes the highest pairwise similarity between each multimodal query token embedding and the pooling token embedding of the multimodal candidate. CLIP's ITC computes the similarity between the first token embeddings of the multimodal query and candidate. Late-Interaction gets the retrieval score by aggregating the maximum dot products over all query tokens with respect to all candidate tokens. From the results in the Table 11 and Table 12, it is clear that, whether at the entity level or the section level, retrieval performance obtained through Late-Interaction is superior. This finding indicates that a fine-grained, comprehensive token-level similarity computation method is more suitable for multimodal-fused retrieval.

Sim. Calculation	Sec. R@1	R@1	R@5	R@10
CLIP's ITC	28.3	37.1	53.2	57.5
Q-Former's ITC	25.2	34.3	52.6	57.2
Late-Interaction	**32.8**	**40.2**	**54.8**	**57.8**

Table 11: The ablation study on the impact of different Step 2 similarity computation methods on entity-level and section-level retrieval results of E-VQA. The α of step 2 is set to 0 to facilitate a direct comparison.

Sim. Calculation	R@1	R@5	R@10
CLIP's ITC	47.6	76.2	82.2
Q-Former's ITC	47.5	76.6	82.4
Late-Interaction	**51.3**	**77.9**	**82.8**

Table 12: The ablation study on the impact of different step 2 similarity computation methods on entity-level retrieval results of InfoSeek. The α of step 2 is set to 0 to facilitate a direct comparison.

D.4 Some Ablation study results in InfoSeek

Table 13 and Table 14, respectively, present the results of two distinct ablation experiments on InfoSeek, whose results on E-VQA are shown respectively in Tables 5 and Tables 6. It is evident that, similar to the results on E-VQA, these findings substantiate the key conclusions of the corresponding ablation studies. This also indirectly demonstrates the soundness and generalizability of our framework design.

E Case Study

To visually assess the performance of our proposed method on the KB-VQA task, we present in the Figure 6 the qualitative results on the test dataset for fine-tuned LlaVA-1.5 using OMGM (the best one shown in Table 3) and for GPT4-o in zero-shot mode. Evidently, the strong retrieval capability enables the generator to handle a wide range of questions, including those that require precise nu-

Ret. Method	R@1	R@5	R@10	R@20
$Image \rightarrow Article$	44.5	64.5	70.7	76.0
$Image \rightarrow Image$	45.6	68.6	74.6	77.9
$Image \rightarrow Title$	51.5	69.2	74.8	79.1
$Image \rightarrow Summary$	**52.6**	**73.9**	**80.0**	**84.8**

Table 13: The ablation study on the impact of different step 1 retrieval methods on entity retrieval results of InfoSeek.

Ret. Modality	R@1	R@5	R@10
$T \rightarrow T$	25.8	65.6	79.2
$(I, T) \rightarrow T$	28.2	67.5	80.0
$T \rightarrow (I, T)$	26.9	67.6	79.8
$(I, T) \rightarrow (I, T)$	**51.3**	**77.9**	**82.8**

Table 14: The ablation study on the impact of different step 2 retrieval modalities on entity-level retrieval results of InfoSeek. Since InfoSeek does not provide annotations for evidence sections, Sec. R@1 is only reported for the results on the E-VQA dataset.

merical answers (as shown in the top-right and mid-right examples) and those that involve specialized entity knowledge (as illustrated in the top-left and mid-left examples). In contrast, GPT4-o often fails to identify specific entity information in the images, which leads to either incorrect responses or statements declaring its inability to answer. These examples qualitatively demonstrate the enhanced performance of our method for compact open-source models on the KB-VQA task.

Moreover, the bottom row also displays three failure cases. In these examples, we are generally able to retrieve the corresponding knowledge for the relevant entities. However, several factors may lead to deviations between the generated answers and the true answers: in cases where the lengthy retrieved information contains multiple potential answers, the generator might be unable to accurately extract the specific answer sought by the query (as shown in the bottom-left example); when the output format required by the query differs from that of the retrieved knowledge, the generator may provide an answer in an incorrect format (as seen in the bottom-mid example); and if the relevant content in the retrieved knowledge is nested across multiple layers, the generator might either omit part of the answer or offer a rough response (as illustrated in the bottom-right example). This indicates that although our method achieves excellent retrieval performance, the limitations in the downstream generator's instruction-following and analytical capabilities still restrict its overall VQA performance.

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 15: 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 visuall content of image in your output. Directly output the answer of the question according to the context.
You are a helpful assistant for answering encyclopedic questions.
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 16: Prompt used for the VQA process of LLM/MLLM in E-VQA. The yellow part is the content only used for MLLM. The red part is the content only used for LLM. The green part is the content used for both LLM and MLLM.

System:

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

You are a helpful assistant for answering encyclopedic questions. Do not answer anything else.

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 17: Prompt used for the VQA process of LLM/MLLM in InfoSeek. The yellow part is the content only used for MLLM. The red part is the content only used for LLM. The green part is the content used for both LLM and MLLM.

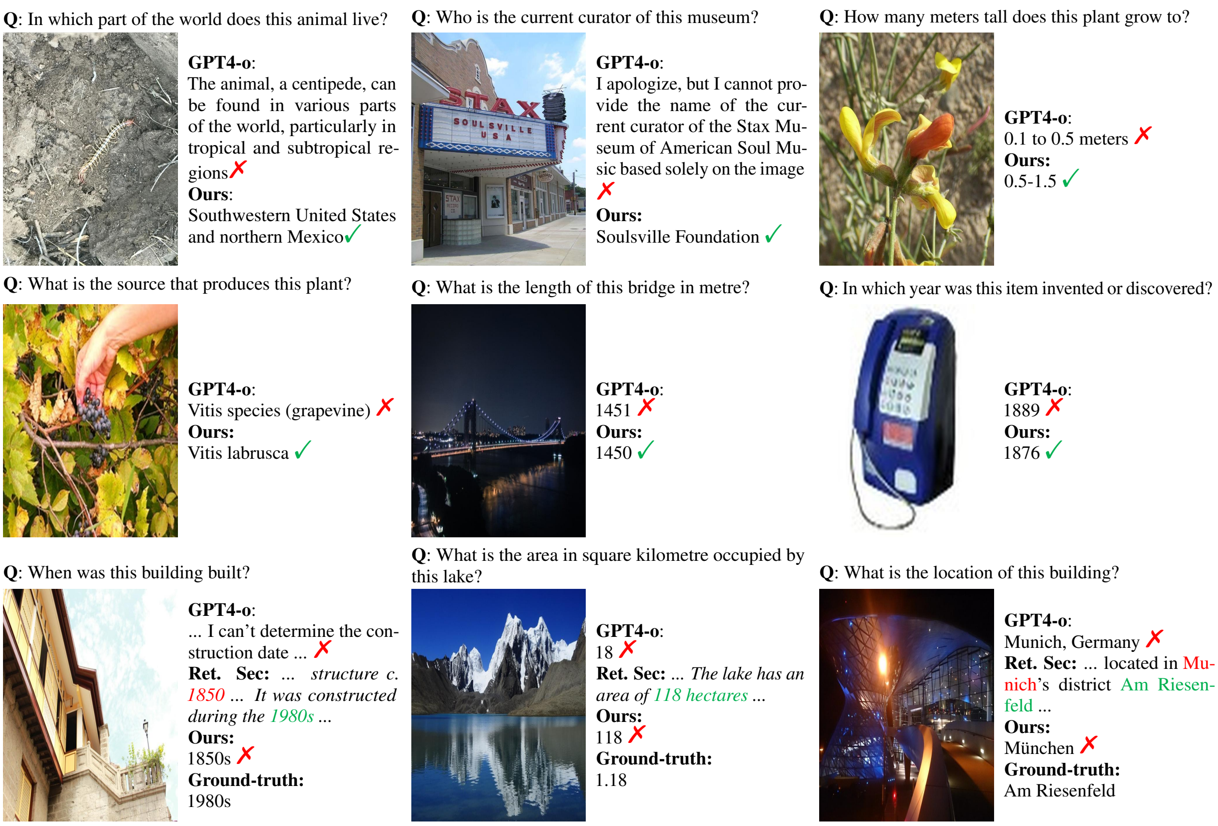

Figure 6: Qualitative VQA results comparing to GPT4-o. The first row shows results in E-VQA and the second row shows results in InfoSeek. Some failure cases are shown in the third row altogether with ground-truth.