Progressive Re-ranking for Multimodal Retrieval-Augmented Generation via Curriculum Learning

Zhu Min[1], Yanchao Hao[1,*], Jian Liu[2], Shizhu He[3], Xi Chen[1]
[1]Tencent [2]University of Science and Technology Beijing
[3]Institute of Automation, Chinese Academy of Sciences
{krystalmzhu, marshao, jasonxchen}@tencent.com
jian.liu@ustb.edu.cn, shizhu.he@nlpr.ia.ac.cn

Abstract

Retrieval-augmented generation (RAG) can enhance large language models (LLMs) by providing external knowledge and helping reduce hallucinations. In multimodal RAG, however, retrieval remains challenging because a single retriever may fail to capture fine-grained multimodal semantics, and visually or semantically similar entities may still contain misleading information for answer generation. We propose a progressive multimodal re-ranking framework with curriculum learning to improve CLIP-based visual coarse-grained retrieval. Our framework progressively refines retrieval results through two stages: fine-grained section-level re-ranking and multimodal section reassessment. To better align re-ranking with multimodal queries, we introduce a curriculum-learning strategy that trains the model with hard negatives that are visually or semantically similar but contain misleading information. Experiments on InfoSeek and Enc-VQA show that our method achieves state-of-the-art answer accuracy and competitive retrieval performance.

1 Introduction

The advancement of large language models (LLMs) and their multimodal extensions (MLLMs) has propelled progress in natural language processing and cross-modal understanding. However, their dependency on static knowledge limits their effectiveness in dynamic or domain-specific tasks such as knowledge-based visual question answering (Chen et al., 2023; Mensink et al., 2023), often leading to inaccurate or unsubstantiated responses, a phenomenon known as "hallucination". As illustrated in Figure 1, a user might ask, "Who is in charge of maintaining this park?" which requires robust multimodal knowledge retrieval and reasoning. The retrieval-augmented generation (RAG) paradigm significantly enhances the factual accuracy and response reliability of existing LLMs in knowledge-intensive tasks by dynamically retrieving relevant information from external knowledge sources. RAG performance hinges on the quality of retrieved documents and their effective utilization. Existing multimodal RAG methods typically focus on two optimization directions: improving the retriever to acquire more relevant and reliable knowledge from multimodal knowledge bases (Caffagni et al., 2024; Yan and Xie, 2024), or guiding LLMs to evaluate and selectively incorporate retrieved content (Zhang et al., 2024b; Cocchi et al., 2025). Our work focuses on improving multimodal retrievers to achieve a more detailed and accurate understanding of multimodal queries.

CLIP-based models (Radford et al., 2021) are prevalent in multimodal retrieval due to their strong cross-modal alignment capabilities. However, their global alignment training strategy can hinder their ability to understand fine-grained semantic information, such as extracting query-relevant region-specific representations from images, limiting their effectiveness in complex scenarios. Multimodal re-

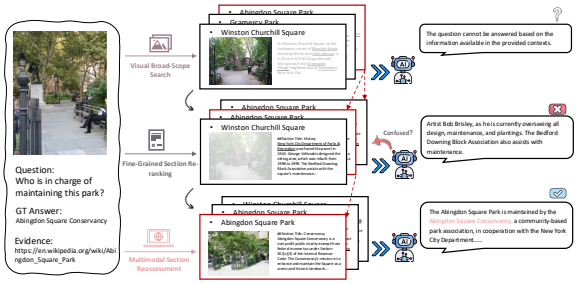

Figure 1: An overview of our method. Given a multimodal knowledge corpus and the query, CLIP-based visual coarse-grained search retrieves the top-k candidate knowledge visual entities. Subsequently, our proposed fine-grained section re-ranking focuses on calculating fine-grained textual section similarities. Finally, multimodal section reassessment evaluates the similarity of multimodal knowledge sections to obtain the top-1 candidate section for answer generation.

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 41135–41147

ranking offers a cost-effective solution by reassessing the relevance of candidate results before response generation. Related research (Yan and Xie, 2024; Yang et al., 2025) employs the multi-stage retrieval strategy to evaluate the correlation between queries and documents from multiple perspectives, incrementally optimizing precision with controlled overhead. Despite some progress in related research, multimodal retrieval-augmented generation still faces several challenges: (1) Multimodal fine-grained semantic alignment: The system needs to accurately identify multimodal fine-grained semantic associations to achieve precise filtering of the retrieved multimodal document collection. (2) Objective mismatch between the retrieval and generation stages: Retrievers are typically trained using pre-trained cross-modal representations or static annotation data, making it difficult to verify if retrieved content is correctly utilized during generation. This can lead to irrelevant or confusing information, compromising the factual consistency and coherence of the generated responses.

We propose a progressive multimodal re-ranking framework, as shown in Figure 1, that improves LLM performance by progressively locating fine-grained multimodal knowledge. Built on CLIP-based visual coarse-grained retrieval, the framework first identifies relevant textual sections through fine-grained late-interaction re-ranking, and then performs multimodal section reassessment to further refine the retrieved evidence. To train the lightweight re-ranking module, we introduce a curriculum-learning strategy with hard negatives that are semantically similar to positive samples but contain misleading information. We conducted extensive experimental evaluations on two widely used knowledge-augmented VQA benchmark datasets—Infoseek (Chen et al., 2023) and Enc-VQA (Mensink et al., 2023). The results demonstrate that our method achieves state-of-the-art question-answering accuracy and competitive retrieval performance.

In summary, our contributions are as follows:

1. We propose a progressive re-ranking framework for multimodal retrieval-augmented generation. This framework introduces a multi-stage, multi-perspective knowledge comparison process to locate key section information of candidate knowledge entries and enhances the accuracy of LLMs in answering questions.

2. We introduce a curriculum learning-based training strategy for multimodal re-ranking, which progressively guides the retriever to distinguish semantically similar knowledge entries through similar visual entity paragraph sampling and hard-negative sampling based on LLM decoding-rank perturbation.

3. Our model achieves state-of-the-art VQA performance on two benchmark datasets - Infoseek and Enc-VQA, demonstrating the effectiveness of the proposed method. Furthermore, a series of ablation experiments validate the effectiveness of the proposed components.

2 Related Works

Visual Question Answering Benchmarks: Early VQA tasks (Antol et al., 2015; Goyal et al., 2017) focused on identifying attributes and detecting objects within images, often relying solely on visual information. Recently, knowledge-based VQA has gained prominence, challenging models to integrate external knowledge for reasoning and answering. Representative datasets include OK-VQA (Marino et al., 2019), InfoSeek (Chen et al., 2023), and Enc-VQA (Mensink et al., 2023).

OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022) assess a model's ability to integrate common-sense knowledge, which existing MLLMs (Chowdhery et al., 2023) may already possess through their pre-training data. Conversely, InfoSeek (Chen et al., 2023) and Enc-VQA (Mensink et al., 2023) emphasize information retrieval capabilities, requiring fine-grained open-world knowledge search to answer questions about detailed attributes of visual entities. These datasets utilize large-scale Wikipedia datasets as a multimodal knowledge corpus. Because the knowledge corpus rarely appears in MLLM training corpora, even state-of-the-art MLLMs (Chowdhery et al., 2023) struggle with implicit knowledge insufficiency when answering factual queries about detailed attributes. This poses a significant challenge for building robust and generalizable multimodal question-answering systems.

Retrieval-augmented generation for VQA: To address hallucination and knowledge obsolescence in LLM systems, retrieval-augmented generation (RAG) has become a prevalent solution for knowledge-intensive tasks. Existing methods for multimodal RAG typically focus on: (1) Retriever performance optimization: Early approaches (Gao et al., 2022; Gui et al., 2022) used DPR or CLIP

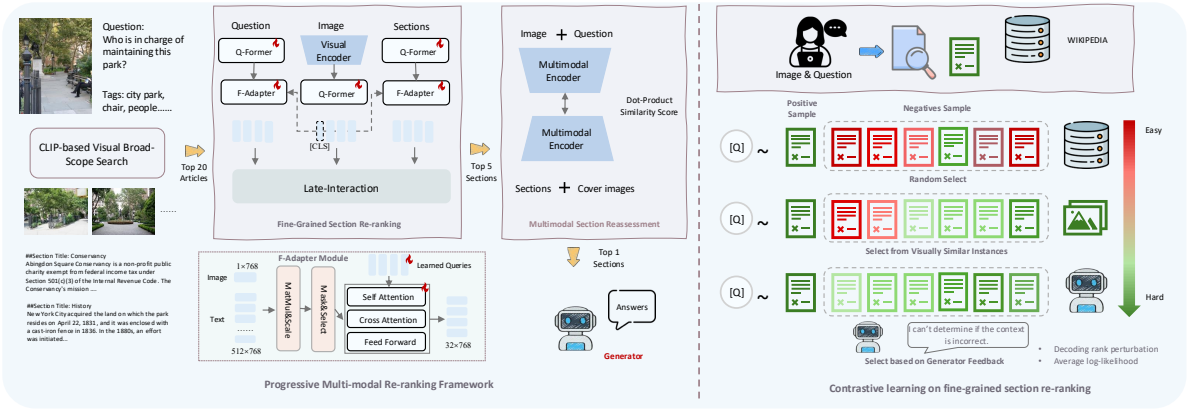

Figure 2: The overview of our progressive multimodal re-ranking framework for LLM generations. The contributions are: a progressive multimodal re-ranking framework based on multi-perspective, multi-granularity alignment; and a curriculum learning optimization strategy for the re-ranking contrastive training.

architectures to achieve fine-grained feature alignment between multimodal queries and textual knowledge. Wiki-LLaVA (Caffagni et al., 2024) pioneered hierarchical retrieval based on CLIP and Contriever (Izacard et al., 2022), refining the retrieved Wikipedia corpus through segmentation. Others (Yan and Xie, 2024; Yang et al., 2025) employ trainable Q-Former (Li et al., 2023) modules and multi-stage re-ranking strategies to improve recall performance with Wikipedia corpora. (2) Knowledge fusion optimization: mR2AG (Zhang et al., 2024b) introduces a multimodal retrieval-reflection-augmented generation process, using special tokens to identify whether knowledge paragraphs should be introduced, enabling adaptive retrieval. MMKB-RAG (Ling et al., 2025) leverages the inherent knowledge of MLLMs and employs a three-stage process to determine retrieval necessity and relevance, ensuring accurate and robust multimodal reasoning. Our method focuses on retriever performance optimization, improving accuracy by introducing a multi-level, fine-grained, progressive multimodal retrieval framework.

3 Proposed Method

We propose a progressive re-ranking framework for multimodal retrieval-augmented generation, as illustrated in Figure 2. Considering the multimodal and multi-granularity nature of queries and Wikipedia corpora, the system introduces a multi-perspective progressive multimodal retrieval method to extract the most relevant multimodal information and feed it into the generator as a contextual prompt. In this section, we first briefly describe the process of retrieval-augmented VQA, then detail the progressive multimodal retrieval strategy, and demonstrate how to train the re-ranking module to robustly evaluate relevant knowledge paragraphs.

3.1 Problem Formulation

Retriever $\mathcal{M}_R$: Given an image I and a question q, the retriever identifies relevant knowledge from a knowledge corpus K through a multi-stage retrieval and re-ranking strategy. The knowledge corpus $\mathcal{K}$, derived from Wikipedia web pages, contains N entities, each with a textual description and associated images: $\mathcal{K} = \{d_1, d_2, \ldots, d_N\} = \{(p_1, t_1), (p_2, t_2), \ldots, (p_N, t_N)\}$, where d_i represents the i-th knowledge entity, t_i represents the textual explanation, and p_i represents the associated images. The retriever encodes both the multimodal query (I, q) and the knowledge entity d_i, evaluating their relevance using a similarity assessment function: $score_i = \mathcal{M}_R(I, q, d_i)$. The document d with the highest score is typically selected as input for the generator.

Generator $\mathcal{M}_G$: The retrieved document d is concatenated with the query, enabling the generator to produce an answer: $y = \mathcal{M}_G(I, q \circ d)$. To further enhance response accuracy, LoRA (Hu et al., 2022) is introduced as learnable parameters within the LLMs. The model is trained using the cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{V} y_{ij} \log p_{ij}, \tag{1}$$

where N is the number of training examples, V is the vocabulary size, y_{ij} represents the target distri-

bution of the j-th token in the i-th training example, and p_{ij} represents the predicted probability.

3.2 Progressive Re-ranking Framework

CLIP (Sun et al., 2023) is commonly used for visual coarse-grained search. Specifically, the CLIP models encode all reference images $\{p_1, p_2, \ldots, p_N\}$ in the Wikipedia corpus and the multimodal query (I, q) into global embeddings: $v_I, v_p^i = \text{CLIP}_v(I, p_i), v_q = \text{CLIP}_t(q)$. These multimodal query vectors contribute equally to the retrieval of knowledge base images through simple averaging:

$$\text{sim}_p(I, q, p_i) = \left(\frac{1}{2}\left[\frac{v_I}{\|v_I\|_2} + \frac{v_q}{\|v_q\|_2}\right]\right)^\top \frac{v_{p_i}}{\|v_{p_i}\|_2}, \quad (2)$$

where v_I is the CLIP visual embedding of the input image I, v_q is the CLIP textual embedding of the question q, and v_{p_i} is the visual embedding of the i-th candidate image. All embeddings are L_2-normalized before similarity computation.

To better capture the multimodal nature of the query, we leverage MLLMs (Liu et al., 2023) to generate image tags related to the question, supplementing the textual query. Finally, we use a FAISS index (Douze et al., 2024) for efficient visual entity retrieval, retaining the top 20 most relevant images and their corresponding Wikipedia articles. We then employ a fine-grained progressive re-ranking framework with fine-grained section re-ranking and multimodal section re-ranking to efficiently improve retrieval accuracy.

Fine-Grained Section Re-ranking: The top 20 Wikipedia articles from the initial retrieval are divided into a candidate set of textual sections $S = \{s_1, s_2, \ldots, s_M\}$, where M denotes the number of sections. We then employ a fine-grained late-interaction approach to identify the most relevant section using the multimodal query. This approach utilizes learnable embedding vectors to represent fine-grained image and text features. A late-interaction mechanism, similar to ColBERT (Khattab and Zaharia, 2020), is used to calculate similarity scores and re-rank the candidate sections.

Specifically, the Q-Former (Li et al., 2023) framework first extracts image features from the input image I using a frozen ViT architecture (Dosovitskiy et al., 2021). Then, a set of learnable query embedding vectors interacts with these image features through cross-attention, generating 32 aggregated visual embeddings:

$$\{z_I^i\}_{i=1}^{32} = Q\text{-Former}(I, q), \quad (3)$$

where z_I^i represents the i-th visual token embedding of image I conditioned on the question q. These vectors, pre-trained via image-text contrastive learning, capture rich image information and are aligned with textual embedding vectors generated by the same architecture.

Furthermore, the Q-Former generates 512 textual token-level embeddings for both the question q and each candidate section s:

$$\{z_q^i\}_{i=1}^{512},\ \{z_s^i\}_{i=1}^{512} = Q\text{-Former}(q, s), \quad (4)$$

where z_q^i and z_s^i denote the i-th token-level embeddings of the question q and candidate section s, respectively. To reduce noise and computation, we further retain only the top 32 question and section tokens most relevant to the visual query.

Considering that the [CLS] embedding may not fully capture the semantic information of the text, and that using all token-level embeddings for fine-grained similarity calculation can be noisy and computationally expensive, we retain the top 32 question and section token embeddings, denoted as $\{\hat{z}_q^i\}_{i=1}^{32}$ and $\{\hat{z}_s^i\}_{i=1}^{32}$, based on their cosine similarity to the image [CLS] embedding z_I^0.

Inspired by the Q-Former architecture, we introduce a smaller aggregation module, F-Adapter, to generate more compact textual embeddings. This module consists of a set of learnable query embedding vectors and a stack of N transformer layers. It interacts with the selected token-level textual embeddings through cross-attention to produce aggregated textual embeddings:

$$\begin{aligned}\{\tilde{z}_q^i\}_{i=1}^{32} &= \text{F-Adapter}\left(\{\hat{z}_q^j\}_{j=1}^{32}\right), \\ \{\tilde{z}_s^i\}_{i=1}^{32} &= \text{F-Adapter}\left(\{\hat{z}_s^j\}_{j=1}^{32}\right),\end{aligned} \quad (5)$$

where $\tilde{z}_q^i$ and $\tilde{z}_s^i$ are the aggregated textual embeddings produced by the F-Adapter.

The fine-grained multimodal similarity score sim_s between the multimodal query $e_q = \{z_I^i\}_{i=1}^{32} \cup \{\tilde{z}_q^i\}_{i=1}^{32}$ and a candidate section $e_s = \{\tilde{z}_s^i\}_{i=1}^{32}$ is calculated using late-interaction. This approach uses a Max-Sum operation to integrate inter-token relationships into a unified similarity measure:

$$sim_s(e_q, e_s) = \sum_{i=1}^{l_q} \max_{1 \le j \le l_s} e_q^i \cdot (e_s^j)^T, \quad (6)$$

where e_q is the multimodal query representation composed of visual token embeddings $\{z_I^i\}_{i=1}^{32}$ and question-side aggregated textual embeddings

$\{\tilde{z}_q^i\}_{i=1}^{32}$, e_s is the section representation composed of section-side aggregated textual embeddings $\{\tilde{z}_s^i\}_{i=1}^{32}$, and l_q and l_s denote the numbers of embeddings in e_q and e_s, respectively. In our implementation, $l_q = 64$ and $l_s = 32$.

Finally, $sim_s(e_q, e_s)$ is combined with a visual entity similarity score $sim_p(I, q, p)$ in a weighted manner to produce the final candidate section similarity score:

$$\text{sim}_{\text{total}} = \lambda \ \text{sim}_p(I, q, p) + (1 - \lambda) \ \text{sim}_s(e_q, e_s), \tag{7}$$

where $\lambda \in [0, 1]$ balances the contributions of visual similarity and fine-grained section similarity. In our experiments, λ is treated as a fixed, dataset-dependent hyperparameter and is set differently across datasets, as described in Appendix A.3.

Multimodal Section Reassessment: To incorporate visual knowledge information, we extract the cover images corresponding to the top 5 Wikipedia sections identified by the previous re-ranking. These image-section pairs are then used for a multimodal reassessment process. We use the pre-trained multimodal re-ranking model (Bai et al., 2023) to generate global representations for both the multimodal query and the candidate sections and then calculate the cosine similarity. Finally, a voting strategy combines the results of the initial and multimodal re-ranking stages to select the most relevant section. This selected section is then concatenated with the question to serve as the input context for the generator.

3.3 Contrastive Training Strategy for the Multimodal Re-ranking

We use contrastive learning to train the fine-grained section re-ranking module. Following the standard contrastive learning setup, we select annotated queries $\mathcal{D}$ and their corresponding Wikipedia sections from the training corpus as positive samples. We employ visual coarse-grained search to obtain k_1 sections as negative samples, which are similar with the positive samples but lack direct semantic associations, fail to provide valuable knowledge for LLM generation.

Furthermore, to guide the retriever in distinguishing more challenging sections, we adopt a curriculum learning (Bengio et al., 2009) setting to further increase the difficulty of negative samples. We define difficult negative samples as sections that are semantically similar to positive samples but contain misleading information, which LLMs are challenging to identify and may generate high-confidence but incorrect responses. We leverage a decoding rank elevation strategy and generation likelihood to assess knowledge effectiveness. Unlike previous methods (Zhang et al., 2024a; Wang et al., 2025), which evaluate document quality based on the generation results without the knowledge supplement, our task focuses on identifying hard negative samples that exhibit subtle differences from the positive sample. To eliminate the spurious positives caused by the generator's internal implicit knowledge, we use LLMs (Grattafiori et al., 2024) that only introduce questions and paragraphs as input context without visual information to evaluate the validity of the sections.

Specifically, we sample N candidate outputs $\{\hat{y}^{(m)}\}_{m=1}^{N}$ conditioned on the question q together with either the positive section s^+ or a visually similar negative section s^-. For each sampled output $\hat{y}^{(m)} = (\hat{y}_1^{(m)}, \ldots, \hat{y}_{T_m}^{(m)})$ with length T_m, we compute its average token log-likelihood (used as a confidence score) under evidence section s:

$$\ell(\hat{y}^{(m)}; q, s) = \frac{1}{T_m} \sum_{t=1}^{T_m} \log P\Big(\hat{y}_t^{(m)} \mid \hat{y}_{<t}^{(m)}, \ q \circ s\Big), \tag{8}$$

where $\hat{y}^{(m)}$ is the m-th sampled output, T_m is its length, and $q \circ s$ denotes the concatenation of the question and the section.

Based on this confidence score, we rank the N sampled outputs in descending order of $\ell(\hat{y}^{(m)}; q, s)$ and compute the rank of the expected (gold) answer y^*:

$$R(s) = 1 + \sum_{m=1}^{N} \mathbb{I}\Big[\ell(\hat{y}^{(m)}; q, s) > \ell(y^*; q, s)\Big], \tag{9}$$

where N is the number of sampled outputs and $\mathbb{I}[\cdot]$ is the indicator function.

We then define the decoding rank perturbation induced by a negative section s^- as the rank difference between the negative and positive sections:

$$\Delta R(s^-) = R(s^-) - R(s^+), \tag{10}$$

where s^+ denotes the positive section and s^- denotes a negative section. A larger $\Delta R(s^-)$ indicates that the negative section is more difficult for the LLM to distinguish from the positive one.

We rank negative samples in descending order of $\Delta R(s^-)$; ties are further broken by the confidence score (e.g., $\ell(y^*; q, s^-)$). We then sample a

subset of negatives from the training set to obtain the k_2 hardest negatives, and further optimize the re-ranking module with a contrastive loss to adapt to the generator's prediction preferences.

4 Experiments

Datasets: The InfoSeek dataset (Chen et al., 2023) comprises 1.3 million image-question pairs, covering over 11,000 visual entities from OVEN (Hu et al., 2023). Following prior work, we utilize a controlled Wikipedia knowledge base of 100,000 articles as the knowledge source. The Enc-VQA dataset (Mensink et al., 2023) contains 221,000 question-answer pairs and 2 million Wikipedia articles, with images sourced from iNaturalist 2021 (Van Horn et al., 2021) and Google Landmarks Dataset V2 (Weyand et al., 2020). Each question-answer pair in Enc-VQA is annotated with the precise knowledge source. For a fair comparison, we adopt the training/testing split used in previous studies (Yan and Xie, 2024); sample statistics and evaluation metrics are presented in the appendix.

Implementation Details: We employ LLaVA-1.5-7B (Liu et al., 2023) to generate image-related tags and Eva-CLIP-8B (Sun et al., 2023) to generate image and text embeddings, extracting the top 20 articles with the highest similarity as a candidate set. The fine-grained section re-ranking module is initialized with pre-trained BLIP-2 (Li et al., 2023) parameters and trained on the Enc-VQA dataset. Notably, since InfoSeek lacks paragraph-level annotations, we adopt a zero-shot transfer approach. The multimodal reassessment stage utilizes the bge-qwen-vl-2b (Bai et al., 2023) for feature extraction. During the answer generation stage, we use a pre-trained LLaMA3-8B (Grattafiori et al., 2024) and a LoRA-fine-tuned LLaVA-1.5-7B (Liu et al., 2023) for inference. During training, we randomly sampled $k_1 = 24$ sections as negative samples and trained for 6 epochs. To improve efficiency, we sampled a subset of 100,000 samples and extracted $k_2 = 4$ sections as hard negative samples, training for 1 epoch. The experimental training is conducted on two NVIDIA H20 80GB GPUs. Prompt settings and further training details are provided in the Appendix.

4.1 Comparison with State-of-the-Art Methods

Table 1 compares the VQA performance of our proposed method against several state-of-the-art approaches. Our method primarily focuses on effectively localizing relevant knowledge and enhancing VQA accuracy through a progressive multimodal re-ranking strategy and curriculum learning incorporating decoding rank perturbation. The results demonstrate that our framework consistently achieves superior performance across two datasets, demonstrating the significant potential of enhancing multimodal retrieval for improved answer accuracy. Specifically, our method achieves optimal VQA performance with LLMs in a zero-shot setting on the Enc-VQA dataset, without explicitly incorporating visual information during generation.

4.2 Retrieval Performance Analysis

Table 2 presents the performance outcomes of various retrieval methods. The methods OMGM (Yang et al., 2025) and ReflectiVA (Cocchi et al., 2025), which employ ChatGPT to generate textual summaries for all Wikipedia articles, exhibit superior performance compared to other retrievers without preprocessing. Our methodology primarily compares performance with baselines that utilize Wikipedia images for initial retrieval. By incorporating image tags as supplementary data, our visual coarse-grained search method achieves an enhancement of 2.5% in Recall@20 for Enc-VQA and 4% for Infoseek. Additionally, our progressive re-ranking strategy significantly improves Recall@1 performance and consistently outperforms the baseline. Specifically, our approach overcomes the performance bottleneck of initial visual retrieval through a two-stage re-ranking strategy, achieving superior Recall@1 performance compared to OMEM (Yang et al., 2025). This demonstrates that our progressive re-ranking strategy effectively filters relevant documents from the initial retrieval results, ensuring the quality of knowledge ultimately fed into the generator, even with imperfect initial retrieval.

4.3 Ablation Study

Contribution of different retrieval stages to VQA results: Table 3 evaluates the contribution of different retrieval strategies to VQA results. The results indicate that incorporating Wikipedia knowledge enhances answer generation, as it provides valuable external information to the model. The progressive, multimodal re-ranking strategy focuses on refining the precision of relevant sections to provide reliable and low-noise information for subsequent generation. The fine-grained

Model	LLM	Enc-VQA	InfoSeek		
			Unseen-q	Unseen-e	Overall
Retrieval process optimization.					
RoRA-VLM (Qi et al., 2024)	Llava-1.5-7b†	20.3	27.3	25.1	-
Wiki-LLaVA (Caffagni et al., 2024)	Vicuna-7B†	21.8	30.1	27.8	28.9
EchoSight (Yan and Xie, 2024)	LLaMA3-8B	41.8	-	-	31.3
OMGM (Yang et al., 2025)	LLaMA3-8B	49.94	35.26	33.61	34.42
OMGM (Yang et al., 2025)	LLaVA-1.5-7B-FT†	50.17	43.46	43.53	43.49
Generation process optimization.					
mR2AG (Zhang et al., 2024b)	Llava-1.5-7b†	-	40.6	39.8	40.2
ReflectiVA (Cocchi et al., 2025)	LLaMA-3.1-8B†	35.5	40.4	39.8	40.1
MMKB-RAG (Ling et al., 2025)	Qwen2-7B†	39.7	36.4	36.3	36.4
Ours	LLaMA3-8B	**50.32**	36.42	33.81	35.07
Ours	LLaVA-1.5-7B-FT†	50.21	**46.99**	**44.22**	**45.56**

Table 1: VQA performance comparison with SOTA methods. † indicates that the LLMs are fine-tuned on domain-specific datasets. Our proposed method achieved the highest prediction performance on the Enc-VQA and InfoSeek benchmarks.

Model	Recall@K on Enc-VQA			Recall@K on InfoSeek		
	K=1	K=5	K=20	K=1	K=5	K=20
Retrieved from query to Wikipedia summary generated from ChatGPT.						
ReflectiVA (Cocchi et al., 2025)	15.6	36.1	49.8	56.1	77.6	86.4
OMGM (Yang et al., 2025) w/o. re-ranking	19.1	41.2	58.7	52.6	73.9	84.8
OMGM (Yang et al., 2025) w. re-ranking	42.8	55.7	58.7	64.0	80.8	84.8
Retrieved from query to Wikipedia images.						
EVA-CLIP-8B I-T (Sun et al., 2023)	3.3	7.7	16.5	32.0	54.0	68.2
EchoSight (Yan and Xie, 2024) w/o. re-ranking	13.3	31.3	48.8	45.6	67.1	77.9
EchoSight (Yan and Xie, 2024) w. re-ranking	36.5	47.9	48.8	53.2	74.0	77.9
Ours w/o. re-ranking	13.4	33.7	51.3	46.7	70.5	81.9
Ours w. re-ranking	**43.6**	**50.1**	**51.3**	**66.4**	**78.5**	**81.9**

Table 2: Retrieval results on the Enc-VQA and InfoSeek datasets. The underline indicates the best performance among all multimodal retrieval methods. ReflectiVA and OMGM achieve excellent performance in the initial retrieval process by introducing article summaries from the entire knowledge corpus. Our approach achieved the highest recall performance compared to other methods that retrieve images within the knowledge base as preliminary results. Simultaneously, our two-stage re-ranking strategy delivered the best performance on the Recall@1 metric.

Method	Enc-VQA	InfoSeek
LLaMA3-8B ZS	18.78	1.56
+ Top1 article of step1	20.17	29.29
+ Top1 section of step2	46.59	32.35
+ Top1 section of step3	**50.32**	**35.07**

Table 3: Contribution of different retrieval stages to VQA results under the LLaMA3-8B zero-shot setting. The introduction and iterative refinement of knowledge retrieval significantly enhance answer generation accuracy.

Methods	R@1	R@5	R@20
CLS Emb & ICL	57.2	76.8	81.9
Late-Interaction	**59.0**	**78.2**	**81.9**

Table 4: ICL or Late-Interaction settings for multimodal section re-ranking on the Infoseek.

section re-ranking module significantly improved performance, demonstrating the effectiveness of its design and training.

Contribution of fine-grained late-interaction approach for re-ranking: We evaluated the impact of two settings on search results: (i) direct similarity computation using text CLS embeddings and (ii) fine-grained similarity score calculation via late interaction. Results presented in Table 4 demonstrate that the late-interaction approach, through a more comprehensive and fine-grained token-level similarity comparison, more accurately identifies the semantic information expressed within the section. Consequently, the weighted fusion of this refined textual similarity with the visual similarity score yields superior retrieval performance.

Impact of negative sampling strategies on fine-

Sample Strategies	R@1	R@5	R@20
Random Select	53.5	74.3	81.9
Visually Similar Entries	57.6	77.5	81.9
Generator Feedback	**59.0**	**78.2**	**81.9**

Table 5: Impact of negative sampling strategies for InfoSeek. Sampling difficult negative examples can guide the retriever to identify sections that are visually or semantically similar but contain misleading content.

Number of samples	K	R@1	R@5
50,000	4	58.32	77.81
50,000	9	58.57	77.94
100,000	4	59.07	**78.28**
100,000	9	59.14	78.17
200,000	4	**59.11**	78.24

Table 6: The effect of the number of hard negative samples on recall metrics.

grained section re-ranking: Table 5 assesses the impact of negative sample selection strategies for the fine-grained section re-ranking on InfoSeek's overall retrieval performance after the second stage. Instead of randomly sampling Wikipedia articles, we implemented a curriculum learning strategy with negative sampling at increasing difficulty levels. This approach initially trains the retriever to differentiate between similar visual knowledge entities and subsequently enhances its ability to discern texts with subtle inconsistencies. This curriculum learning approach improves the generalization performance of the re-ranking module. Furthermore, Table 6 demonstrates that training with a small sample of hard negative examples can achieve significant performance gains at a manageable cost.

Comparison of retrieval efficiency and performance across different multimodal RAG approaches: Table 7 illustrates the retrieval efficiency and performance comparison between two multimodal re-ranking methods. Despite adopting a stepwise retrieval strategy, our method incurs only minimal additional retrieval overhead with small-scale re-ranking while delivering significantly superior VQA results. This demonstrates that our approach achieves an effective balance between retrieval efficiency and performance.

4.4 Quantitative Result Analysis

We present a comparison of our method's answer generation on real-world examples from the Enc-VQA dataset against state-of-the-art VLMs, including GPT-4o and Claude-3.5, as illustrated in Figure 3. These cases highlight the effectiveness of our approach in addressing complex questions. For instance, precise paragraph localization mitigates knowledge deficiencies during the generation process of LLMs and VLMs, leading to improved prediction accuracy.

Methods	Avg Time	R@1	VQA Score
EchoSight	**0.429**	36.5	41.8
Ours	0.496	**43.6**	**50.32**

Table 7: Comparison of retrieval efficiency and performance across various multimodal re-ranking methods. Our approach provides significant improvements in VQA performance while incurring only a minimal increase in retrieval time.

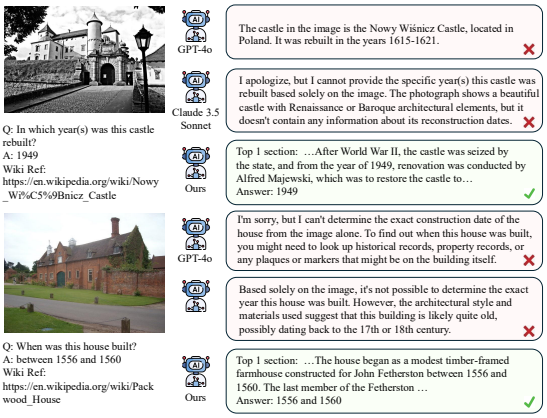

Figure 3: Quantitative result analysis on Enc-VQA datasets.

5 Conclusion

This paper introduces a progressive multimodal re-ranking framework for knowledge-based VQA. The system focuses on optimizing retrieval performance over the multi-modal knowledge corpus, introducing a progressive, multi-granularity multimodal retrieval framework. To enhance the performance of the retriever, we fine-tune a small-scale re-ranking module by introducing a curriculum learning training strategy, guiding the retrieval module to better understand multi-modal query intent at a low cost. Compared to other existing research on the InfoSeek and Enc-VQA datasets, our method achieves significant improvements in VQA performance, demonstrating the superiority of the proposed method. This research provides new insights into the rational design of multi-modal re-ranking modules and the layout of retrieval processes.

Limitations

Despite its strong performance, our method still has two main limitations. First, the overall upper bound is constrained by the quality of visual coarse-grained retrieval. To balance efficiency and performance, we do not substantially expand the initial retrieval scope, which leaves room for improvement in later stages. Second, hard-negative sampling based on decoding-rank perturbation is computationally expensive because it relies on LLM inference, making it difficult to scale to the full training corpus. Moreover, the current sampling strategy is static. Future work could explore dynamic hard-negative selection based on generator feedback, for example through reinforcement learning.

Acknowledgments

This work was supported in part by the Beijing Major Science and Technology Project under Contract no. Z251100008125025.

References

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.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

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 Workshops*, pages 1818–1826.

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9209.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. *arXiv preprint arXiv:2401.08281*.

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5067–5077.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander G Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. Kat: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. LoRA: Low-rank adaptation of large language models. *The Tenth International Conference on Learning Representations*, 1(2):3.

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.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

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.

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.

Zihan Ling, Zhiyao Guo, Yixuan Huang, Yi An, Shuai Xiao, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. 2025. Mmkb-rag: A multi-modal knowledge-based retrieval-augmented generation framework. *arXiv preprint arXiv:2504.10074*.

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, 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3113–3124.

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.

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*.

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.

Shaohan Wang, Licheng Zhang, Zheren Fu, and Zhendong Mao. 2025. Dacl-rag: Data augmentation strategy with curriculum learning for retrieval-augmented generation. *arXiv preprint arXiv:2505.10493*.

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. 2024. EchoSight: Advancing visual-language models with wiki knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1538–1551.

Wei Yang, Jingjing Fu, Rui Wang, Jinyu Wang, Lei Song, and Jiang Bian. 2025. OMGM: Orchestrate multiple granularities and modalities for efficient multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24545–24563.

Peitian Zhang, Zheng Liu, Shitao Xiao, Zhicheng Dou, and Jian-Yun Nie. 2024a. A multi-task embedder for retrieval augmented LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3537–3553.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, and 1 others. 2024b. mR^{2}AG: Multimodal retrieval-reflection-augmented generation for knowledge-based VQA. *arXiv preprint arXiv:2411.15041*.

A Appendix

A.1 Dataset setup

InfoSeek Dataset: The InfoSeek dataset comprises 1.3 million samples derived from a Wikipedia corpus. Given the non-public availability of the original external knowledge base, we utilize 100,000 Wikipedia articles provided by EchoSight as our knowledge source. This dataset features 8.9K human-authored visual information-seeking questions and 1.3M automatically generated questions, partitioned into training, validation, and test sets.

	# QA Pairs				Knowledge Corpus	
	# Ret Train	# Gen Train	# Valid	# Test	# Train	# Valid
InfoSeek	-	10k	-	71335	100k	100k
Enc-VQA	901674	10k	11696	4750	2M	2M

Table 8: Dataset statistics for InfoSeek and Enc-VQA datasets.

Following established protocols, we employ the validation set for experimental evaluation. This validation set is further divided into two distinct subsets: Unseen Entity and Unseen Question. We report evaluation results separately for each of these subsets. During the fine-tuning phase for LLaVa, we randomly select 100,000 questions from the training set for training. Acknowledging the absence of section-level annotations in the labeled data, we employ a trained retriever to identify the most relevant sections, which are then incorporated as supplementary contextual information.

Enc-VQA Dataset: Enc-VQA comprises 221,000 question-answer pairs and 2 million Wikipedia articles and encompasses questions pertaining to 16.7K distinct entities. The dataset features a diverse range of question-answering tasks, including single-hop, two-hop, and multi-answer questions. Consistent with prior work, our evaluation focuses solely on single-hop and multi-answer question types, excluding two-hop questions that necessitate more extensive retrieval and reasoning processes. Enc-VQA provides associated Wikipedia articles, fine-grained section annotations, and supporting fact information for each question-answer pair, which we leverage for model training and evaluation. Acknowledging the challenges associated with downloading corresponding images for all candidate paragraphs, we employ pre-extracted CLIP features of all image entities from EchoSight for visual coarse-grained search. Subsequently, only the cover images of the Wikipedia articles associated with the top 5 candidate paragraphs are considered as visual references corresponding to the section during the multi-modal section re-ranking stage.

Comprehensive statistics for both datasets are presented in Table 8.

A.2 Evaluation Metrics

Recall: We evaluate the retrieval performance using the standard Recall@K metric. Specifically, Recall@K measures the proportion of queries for which the correct, ground-truth Wikipedia article link is found within the top K retrieved results. A retrieval is considered successful only if there is an exact string match between the retrieved link and the ground-truth link.

Answer Accuracy on InfoSeek: To evaluate answer accuracy on the InfoSeek dataset, we employ different metrics depending on the question type. For questions requiring string or temporal answers, we use the VQA accuracy metric. For questions requiring numerical answers, we use the relaxed accuracy metric.

Answer Accuracy on Enc-VQA: To evaluate answer accuracy on the Enc-VQA dataset, we employ different metrics depending on the question type. For simple, single-hop questions, we use BERT Matching (BEM) to assess the semantic similarity between the predicted answer and the ground-truth answer. This allows us to consider answers that are semantically equivalent but phrased differently as correct. For multi-answer questions, the model's output is first transformed into a list of answers through string splitting. We then calculate the Intersection-over-Union (IoU) between the predicted answer set and the ground-truth answer set. If the IoU is greater than or equal to 0.5, the answer is considered correct. Otherwise, the BEM score is used to evaluate the correctness of the answer.

A.3 Experimental Hyperparameters and Training Setup

Retrieval Stage: During the re-ranking module training, we randomly sampled 24 visually similar passages as negative samples from a training dataset of 900,000 instances. Additionally, we sampled 4 semantically similar passages as hard negative samples from a 100,000-instance subset of the training data. The training process employed a learning rate of 1e-5 and a batch size of 8. Initially, the model was trained for 6 epochs on the entire training set using only the negative samples to facilitate comprehensive learning of retrieval requirements across diverse image and question types. To adhere to a curriculum learning paradigm and construct negative samples with varying difficulty gradients, we progressively used the hard

negative sample subset during the subsequent 1 training epochs. This progressive training strategy aims to guide the retriever towards identifying a wider range of more challenging Wikipedia articles, thereby enhancing the system's recall and generalization performance. Each training epoch required an average of 30 hours to complete on a single Nvidia H20 (80G) GPU. During inference, the weighting factor λ was set to 0.74 for InfoSeek and 0.56 for Enc-VQA, reflecting the relative contributions of visual and textual similarity.

Generation Stage: For the LLaVA training phase, we randomly selected 100,000 samples from each of the two datasets to serve as training data. We employed LoRA for fine-tuning the LLM, with the following parameter settings: r = 128 and alpha = 256. The learning rate was set to 2e-5, the batch size to 16, and the total number of epochs to 1.

A.4 Prompt Engineering

To guide the LLMs and VLLMs in generating relevant responses during image tagging and answer generation, we employed specific prompt templates, leveraging a prompt engineering approach. The prompt templates utilized for each task are detailed below:

Image Tag Generation using LLaVA-1.5-7B:

You are an image tag generator, and your task is to provide tags related to the image according to the question, thereby helping to retrieve relevant knowledge.

Image: {}

Question: {}

Directly provide 5 related image tags, use ',' as the separator:

LLaMA Generation on InfoSeek Dataset:

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 is this mountain located in?

Just answer the question. No explanation is needed. Short answer is: Province of Belluno

- Context: {}

- Question: {}

Just answer the question. No explanation is needed. Short answer is:

LLaMA Generation on Enc-VQA Dataset:

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.

- Context: {}

- Question: {}

The answer is:

LLaVA-1.5-7B Generation on InfoSeek Dataset:

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 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 is this mountain located in?

Just answer the question. No explanation is needed. Short answer is: Province of Belluno

{image}

- Context: {}

- Question: {}

Just answer the question. No explanation is needed. Short answer is:

LLaVA-1.5-7B Generation on Enc-VQA Dataset:

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.

{image}
- Context: {}
- Question: {}
The answer is:

A.5 Further Experimental Analysis

Fine-grained analysis for InfoSeek: Table 9 presents the accuracy results on the validation set of the InfoSeek dataset across different categories. From the results, we can observe that the model shows relatively better performance in handling time-related and numerical aspects. This may be due to the fact that answers for time or numerical queries are often more precise. In contrast, string-based questions tend to be more open-ended, where even high recall rates may not guarantee accurate responses.

Class	Total	Time	Number	String
Unseen Question Score	36.42	40.17	42.01	34.26
Unseen Entity Score	33.81	51.06	34.15	32.63

Table 9: Fine-grained analysis for InfoSeek based on zero-shot LLaMA-3-8B.