Large Language Models Know What is Key Visual Entity: An LLM-assisted Multimodal Retrieval for VQA

Pu Jian[1,2], **Donglei Yu**[1,2], **Jiajun Zhang**[1,2,3,4]*

[1] Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Wuhan AI Research [4] Shanghai Artificial Intelligence Laboratory
{jianpu2023, yudonglei2021}@ia.ac.cn
jjzhang@nlpr.ia.ac.cn

Abstract

Visual question answering (VQA) tasks, often performed by visual language model (VLM), face challenges with long-tail knowledge. Recent retrieval-augmented VQA (RA-VQA) systems address this by retrieving and integrating external knowledge sources. However, these systems still suffer from redundant visual information irrelevant to the question during retrieval. To address these issues, in this paper, we propose **LLM-RA** , a novel method leveraging the reasoning capability of a large language model (LLM) to identify key visual entities, thus minimizing the impact of irrelevant information in the query of retriever. Furthermore, key visual entities are independently encoded for multimodal joint retrieval, preventing cross-entity interference. Experimental results demonstrate that our method outperforms other strong RA-VQA systems. In two knowledge-intensive VQA benchmarks, our method achieves the new state-of-the-art performance among those with similar scale of parameters and even performs comparably to models with 1-2 orders larger parameters.

1 Introduction

Visual question answering (VQA) task involves providing a natural language answer to a given question based on the image (Gao et al., 2022). Mainstream approaches utilize visual language models (VLMs) to accomplish VQA tasks (Li et al., 2023b; Liu et al., 2024). Previous studies suggest that VLMs struggle to capture long-tail knowledge in the real world, such as fine-grained entity categories and their detailed attributes (Chen et al., 2023d; Mensink et al., 2023). To alleviate this problem, plenty of works retrieve relevant information from external knowledge sources (Gui et al., 2022; Si et al., 2023). Such retrieval-augmented VQA (RA-VQA) systems can compensate for the limitations of VLMs in modeling long-tail knowledge. For example, some studies find that regions of interest (ROIs) corresponding to salient objects are helpful to knowledge retrieval (Lin et al., 2022; Sun et al., 2023), and utilize existing object detectors to identify ROIs and employ them for retrieval.

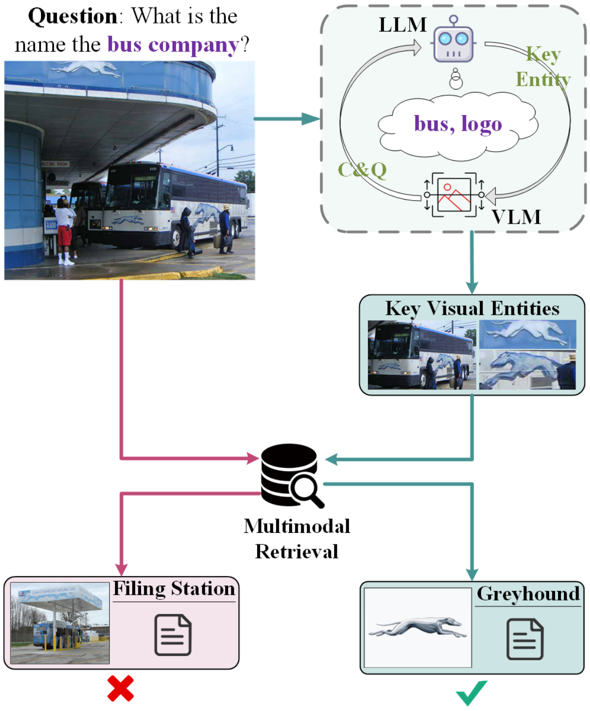


Figure 1: Schematic illustrates how LLM assists multimodal retrieval for VQA. Due to redundant visual information like "person" and "building", undesired knowledge like "filling station" is retrieved, which matches the high-level semantics of the image. Leveraging LLM's reasoning capability, key visual entities like "bus" and "logo" can be extracted, which is significant for obtaining the desired knowledge. C&Q: The caption and question which is parsed as a prompt.

In RA-VQA context, there are frequently key visual entities in the images, which are important for retrieving necessary knowledge. However, in existing RA-VQA systems, they are not sufficiently taken into account for the following rea-

[*] Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 10939–10956
November 12-16, 2024 ©2024 Association for Computational Linguistics

sons: 1) key visual entities are overwhelmed by redundant visual information irrelevant to the question. Taking Figure 1 as an example, the question focuses on the "bus company", thus the key visual entities being "bus" and "logo". In contrast, other redundant visual information, such as "person", "road", and "building", are irrelevant to the question and interfere with retrieval. As illustrated in Figure 1, it could lead to undesired retrieval results. 2) Encoding of key visual entities is prone to cross-entity interference. Following Karpukhin et al. (2020), previous works encode all visual representations into a one-dimensional query for retrieval. It causes mutual interference between key visual entities, making the retriever insensitive to fine-grained visual information.

To address the aforementioned problems, we propose LLM-RA, which leverages the reasoning capabilities of LLMs to assist VLMs in key visual entity extraction during retrieval, thereby enhancing the RA-VQA system. Broadly, the extraction of key visual entities involves two components: reasoning and grounding. With crafted prompts, the LLM infers key visual entities essential for answering the question based on the image description. These key visual entities are then parsed into reference expressions (Liu et al., 2023b; Li et al., 2022b) for visual grounding. In this process, LLMs interact with VLMs, compensating for the VLM's lack of reasoning capability. Furthermore, inspired by recent advanced multimodal retrieval works (Lin et al., 2024a,b), LLM-RA independently encodes different key visual entities when constructing the query for the subsequent multimodal joint retrieval (Khattab and Zaharia, 2020). This design enables different key visual entities to independently contribute to the retrieval, making LLM-RA sensitive to key visual entities.

Leveraging the reasoning capabilities of LLM to emphasize question-relevant key visual entities during retrieval, LLM-RA achieves state-of-the-art (SOTA) performance compared to previous RA-VQA systems in two challenging knowledge-intensive VQA benchmarks, OK-VQA (Marino et al., 2019) and Infoseek (Chen et al., 2023d). LLM-RA also achieves comparable or even better performance than systems with 1-2 orders large parameters ($\geq$ 50B). We summarize our contributions as follows:

- We propose LLM-RA, an LLM-assisted multimodal retrieval method emphasizing question-relevant key visual entities during retrieval, to enhance RA-VQA systems.
- We systematically explore how to better exploit the visual feature to retrieve desired knowledge. The empirical results suggest that emphasizing question-relevant key visual entity performs the best, compared to other approaches.

2 Related Work

Our work is related to several recent studies in the VLMs and LLMs field.

Knowledge-intensive VQA. Different from traditional VQA tasks focusing on grasping high-level image semantics (Goyal et al., 2017; Hudson and Manning, 2019), certain visual questions like examples in (Marino et al., 2019; Schwenk et al., 2022) necessitate VLMs to comprehend world knowledge embedded within images, i.e., knowledge-intensive VQA (KI-VQA). In recent years, numerous studies have focused on exploring knowledge-based VQA tasks with LLMs. Some approaches treat LLMs as extensive knowledge sources, employing methods like in-context learning to harness internal knowledge sources within the model (Hu et al., 2023; Shao et al., 2023; Xenos et al., 2023). However, these efforts suffer from LLMs' limitations in modeling long-tail knowledge. Furthermore, some works in the KI-VQA domain highlight the importance of retrieving information from external knowledge sources (Gui et al., 2022; Si et al., 2023; Lin et al., 2024a), to tackle long-tail knowledge, i.e., RA-VQA systems. For example, datasets like Infoseek and Encyclopedic VQA (Chen et al., 2023d; Mensink et al., 2023) demand models to identify specific species, industrial products, etc. Leveraging multimodal retrievers to capture information from external knowledge bases (Ren et al., 2023; Tang et al., 2023) can compensate for the limitations of mainstream VLMs in modeling long-tail knowledge in the real world.

Multimodal Retrieval. When tackling visual questions with retrieval, the majority of works utilize mature text retrievers (Karpukhin et al., 2020; Khattab and Zaharia, 2020). These approaches employ VLMs to convert images into text descriptions, leveraging pre-trained retrievers on large-scale corpora to achieve desired performance (Lin and Byrne, 2022; Si et al., 2023). However, such processes result in a significant loss of visual information and represent a sub-optimal approach. CLIP and its variants are frequently employed for image-text retrieval to compensate for unimodal retrievers' limitations. However, some studies indicate that the corpora used to train CLIP may lack long-tail knowledge (Li et al., 2023a; Yu et al., 2023a), like well-known figures, places and events. Recent works integrate text retrievers with CLIP-extracted visual features for joint information retrieval, illustrating their complementary nature (Salemi et al., 2023; Yu et al., 2023b). In this work, we extend existing multimodal retrievers by employing LLMs to filter visual information irrelevant to the question, thus enhancing retrievers' performance in the VQA field.

LLM-assisted Visual Reasoning. Some researchers have attempted to leverage the strong reasoning and instruction-following capabilities of LLMs to assist VLMs in completing complex visual tasks. LLMs serve various roles, including task dispatchers, reasoners, or language refiners in these works. Woodpecker (Yin et al., 2023) utilizes LLMs' information extraction abilities to extract entities from image descriptions for hallucination correction. REPARE (Prasad et al., 2023) leveraging LLMs' command-following capabilities to add the image's relevant details in the question, for prompt refinement. GPT4Tool (Yang et al., 2024) utilizes open-source LLMs like LLaMA to schedule tools such as object detectors and image captioners, executing complex visual tasks. These studies inspire us to leverage LLMs' reasoning capabilities to extract key visual entities from images based on questions, enhancing multimodal retrieval in RA-VQA systems.

3 Method

As shown in Figure 2, LLM-RA enhances the RA-VQA system by improving multimodal retrieval. It comprises two stages: 1) Key visual entities extraction; and 2) Multi-modal joint retrieval. In the first stage, the reasoning capability of LLM is harnessed to discern all potentially key visual entities for visual grounding. In the second stage, key visual entities are independently encoded in the query for joint retrieval. In principle, any VQA answer generator concerning RA-VQA can utilize the knowledge retrieved by LLM-RA. We employ existing methods that use retrieval documents to generate VQA answers (Lin and Byrne, 2022) for outputting the final answers.

3.1 Key Visual Entity Extraction

To eliminate redundant information and highlight key visual entities during retrieval, we leverage the strong reasoning capabilities of LLMs to determine potentially question-relevant visual entities.

General Information from Image Captions. We leverage the unique capabilities of large visual-language models (LVLMs), which excel at image captioning (Li et al., 2022a, 2023b). Such captions encompass the significant visual details, like entities relevant to corresponding questions (Zhu et al., 2023; Dai et al., 2024). LLM utilizes these caption and input questions to perform inference and determine key visual entities. For instance, in Figure 2, the LVLM generates the following caption: `"A sunlit churchyard features a white church with twin towers and a statue amid graves..."`

LLM-assisted Key Entity Extraction. LLM-RA utilizes the remarkable capabilities of LLMs in reasoning and generative information extraction (Xu et al., 2023). Specifically, we employ in-context learning to prompt open-source LLMs to generate structured entities and their attributes. The utilized instruction is `"Given a description of an image, output the entities that the question might focus on; {examples}; {question}; {caption}; output:"`, as detailed in the prompt template provided in Appendix A. With carefully crafted instructions and in-context examples, we instruct the LLM to generate entities along with their attributes potentially relevant to the question, in structured formatting as `{"entity": "attribute"}`. For example, in Figure 2, the question focuses on the `"region"`. With the generated caption, LLM would deduce that landmarks like `{"statue": "amid graves"}` are key entities helpful to identify the `"region"`.

Visual Grounding. After obtaining key entities and their attributes relevant to given questions, we utilize an existing referring expression comprehension model (Liu et al., 2023b) to determine the regions of interest (ROIs) corresponding to these entities. Specifically, we employ the template `"The {entity} that {attribute}"` to parse the structured entities into referring expressions. For example in Figure 2, structured entities `{"statue": "amid graves"}` is parsed as `"The statue that amid graves"`, used as referring expression for visual grounding.

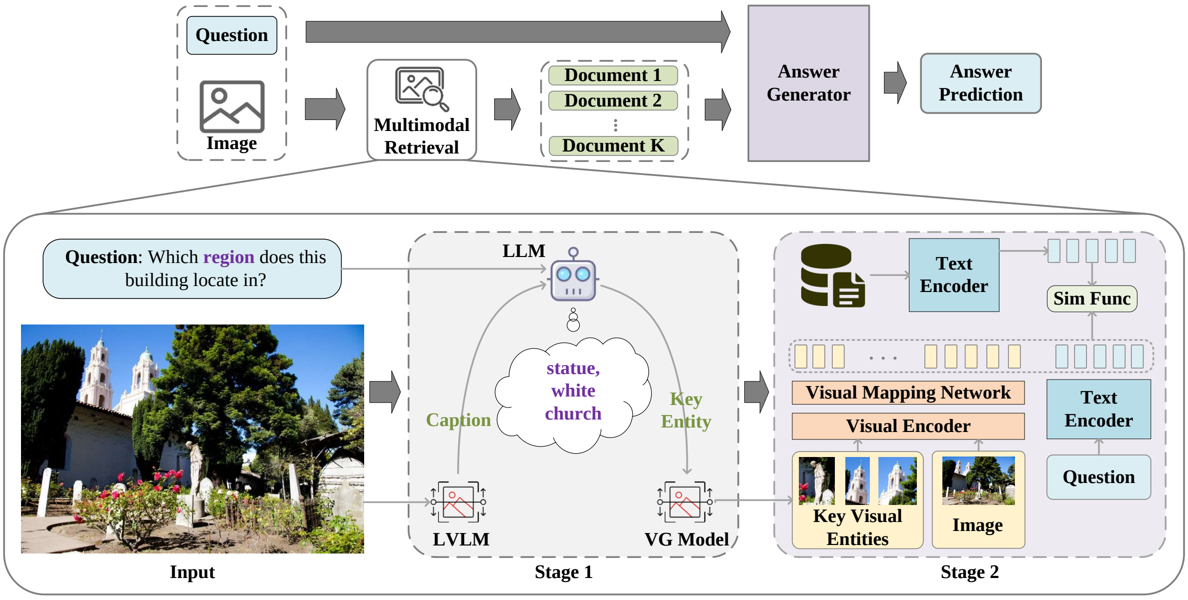


Figure 2: Schematic of LLM-RA. In stage 1, with image caption "`A sunlit churchyard features a white church with twin towers and a statue amid graves...`", LLM would deduce that landmarks like "`statue`" and "`church`" are key visual entities that help to identify the "`region`". Then visual grounding is conducted to determine ROIs. In stage 2, ROIs of key visual entities are independently encoded for joint retrieval.

3.2 Multimodal Joint Retrieval

To eliminate cross-entity interference during retrieval, inspired by recent advanced multimodal retrieval works (Lin et al., 2024a,b), we independently encode the extracted key visual entities and perform joint retrieval.

Independent visual representation. For the image I and ROIs of key visual entities, we employ a CLIP Encoder $\mathcal{H}_v(\cdot)$ to extract global features and key entity-centric visual features (Radford et al., 2021). For the questions Q_s, we utilize a Transformer-based text encoder $\mathcal{H}_l(\cdot)$ (Santhanam et al., 2022) to extract text representations. Furthermore, we pretrain a visual mapping network $\mathcal{H}_p(\cdot)$ with contrastive loss on a large scale of Wikipedia image-text pairs following Lin et al. (2024a) and Wei et al. (2023), to model fine-grained knowledge mappings between images and documents in the context of knowledge retrieval. This network maps input visual features to embeddings with the same depth as the text encoder. Subsequently, we concatenate all visual and textual embeddings to obtain the final query

$$E_Q = \begin{bmatrix} \mathcal{H}_l(Q_s) \\ \mathcal{H}_p(\mathcal{H}_v(I)) \\ \mathcal{H}_p(\mathcal{H}_v(I_1)) \\ \vdots \\ \mathcal{H}_p(\mathcal{H}_v(I_{N_r})) \end{bmatrix} \in \mathcal{R}^{N_Q \times d_L} \tag{1}$$

for retrieval, where I_r is the ROI of r-th key visual entities, and N_r represents the number of ROIs. With text encoder $\mathcal{H}_l(\cdot)$, we extract token level knowledge representation

$$E_D = \mathcal{H}_l(D) \in \mathcal{R}^{N_D \times d_L} \tag{2}$$

from the documents D in the knowledge base.

Joint retrieval. We adopt the similarity in (Khattab and Zaharia, 2020) to measure the relevance between the question-image pair (Q_s, I) and document D in the knowledge base, given by

$$S((Q_s, I), D) = \sum_{i=1}^{N_Q} \max_{j=1}^{N_D} E_{Q_i} \cdot E_{D_j}. \tag{3}$$

Unlike DPR (Karpukhin et al., 2020), which compresses E_Q and E_D into one-dimensional embeddings, this similarity models their relevance at the word, phrase, and visual entity levels. It prevents interference between different key visual entities and textual information.

4 Experimental Setup

4.1 Implementations

The implementations of the LLM-RA and the training setup are briefly described. More detailed implementations are shown in Appendix C.

Multimodal retriver implementation. For detailed caption generation, MiniGPTv2 (Chen et al., 2023a) is adopted. And the frozen LLM of MiniGPTv2 is utilized for key entity extraction. Then we select Grounding-Dino (Liu et al., 2023b) for visual grounding. ColBERTv2 (Santhanam et al., 2022) and CLIP ViT-base (Radford et al., 2021) are adopted to initialize the text encoder and vision encoder in the multimodal retriever. For executing retrieval, all documents in the knowledge base are indexed using FAISS (Johnson et al., 2019), an off-the-shelf library that enables fast vector-similarity search. We use contrastive loss (Radford et al., 2021) for multimodal retriever training. In-batch negative sampling is adopted during training following Santhanam et al. (2022).

Answer generator. We adopt BLIP2-Flan-T5-XL (Li et al., 2023b) as the answer generator. It takes the concatenation of text embeddings (question and retrieved documents) and visual embeddings (image) as input. Following Lin and Byrne (2022), the generator outputs an answer for each retrieved document and selects the best candidate answer by the joint probability of retrieving and the generated answer. We use LoRA (Hu et al., 2021) to finetune the answer generator, while the retriever is frozen.

4.2 Datasets

We conduct our experiments on OK-VQA (Marino et al., 2019) and Infoseek (Chen et al., 2023d), both of which are knowledge-intensive VQA datasets. Details are the following.

OK-VQA. OK-VQA is a relatively large-scale dataset that emphasizes the necessity of external knowledge (commonsense or domain-specific) to answer questions. The questions in OK-VQA are not annotated with ground truth documents. Therefore, we leverage the Google Search (Luo et al., 2021) for OK-VQA as the knowledge base. Google Search is a passage corpus crawled by Luo et al. (2021) from high-frequency Google search web pages. The advantages of this corpus include its large scale ($\sim$ 170K passages), public availability, and wide coverage of knowledge types. Previous research (Luo et al., 2021; Lin and Byrne, 2022) has indicated that this corpus encompasses a substantial amount of knowledge required to answer questions in OK-VQA.

Infoseek. Infoseek is custom-built for information retrieval questions that cannot be addressed solely with commonsense. This dataset emphasizes identifying fine-grained entity classes and their detailed attributes in images. The performance of existing multimodal models on this dataset is subpar (Chen et al., 2023d). Different from OK-VQA, Infoseek offers a knowledge base extracted from Wikipedia, supplemented by ground truth document annotations. The released knowledge base contains 6M Wikipedia passages. Following the experimental setup in the original paper (Chen et al., 2023d) and Lin et al. (2024a), we retain the Wikipedia passages annotated as ground truth knowledge in Infoseek VQA samples ($\sim$ 7K), and randomly sampled entities from the remaining 6M Wikipedia passages to form a 100K Wikipedia knowledge base.

4.3 Evaluation Metrics

The following metrics are adopted to evaluate the proposed method.

Retrieval metrics. We employ Recall@K (Lin and Byrne, 2022) to assess the retriever's performance. This metric measures the probability of a positive document within the top K retrieved documents. Given the absence of ground truth document annotations in OK-VQA, following Lin and Byrne (2022) and Salemi et al. (2023), we consider documents containing the correct answers as positive documents.

VQA metrics. To assess the final accuracy of VQA, we employ the VQA Score and Exact Match (EM) metrics for OK-VQA (Marino et al., 2019). For Infoseek, we utilize the built-in evaluation metric, Infoseek Score (Chen et al., 2023d).

5 Results

5.1 Overall Performance

The comparison between our method and other baselines is shown in Table 1 and Table 2. Detailed analyses are the following.

Results on OK-VQA. As illustrated in Table 1, LLM-RA outperforms other retrieval-augmented VQA (RA-VQA) systems. It achieves the highest VQA Score of 63.29 and an exact match (EM) of 68.31, which is an improvement of 1.43 VQA Score over the previous state-of-the-art (SOTA) system, RAVQAv2 (Lin et al., 2024a). Additionally, based on a 4B multimodal model, LLM-RA demonstrates performance on KI-VQA tasks comparable to many sophisticated systems utilizing very large base models ($\geq$ 50B) such as PromptCap and Prophet, employing the GPT-3 with in-

Method	Knowledge Source	Finetune	R@5	EM	VQA
VRR (Luo et al., 2021)	Google Search	✓	80.40	47.61	45.08
TRiG (Gao et al., 2022)	Wikipedia	✓	-	54.73	50.50
RA-VQA (Lin and Byrne, 2022)	Google Search	✓	81.25	55.77	51.22
KAT (Gui et al., 2022)	Wikipedia + GPT-3	✓	-	57.85	54.41
TWO (Si et al., 2023)	VQAv2 + Wikipedia	✓	-	61.32	56.67
REVIVE (Lin et al., 2022)	Wikipedia + GPT-3	✓	-	62.38	58.00
RA-VQAv2 (Lin et al., 2024a)	Google Search	✓	89.32	67.10	61.86
Systems based on very large models ($\geq$50B parameters)					
PICa (Yang et al., 2022)	-	✗	-	-	48.00
Flamingo-80B (Alayrac et al., 2022)	-	✗	-	-	57.80
PromptCap (Hu et al., 2023)	-	✗	-	66.07	60.40
Prophet (Shao et al., 2023)	-	✗	-	66.74	61.11
LLM-RA (RA-VQA system with LLM-assisted multimodal retrieval)					
LLM-RA	Google Search	✓	**90.37**	**68.31**	**63.29**
w/o Key visual entity	Google Search	✓	87.24	66.43	60.98
w/o Independent VR	Google Search	✓	85.83	65.05	60.37
w/o Key visual entity & Independent VR	Google Search	✓	83.76	63.81	59.17
w/o External knowledge	Google Search	✓	-	59.45	55.49

Table 1: Performance of the proposed method on OK-VQA. VR stands for visual representation. EM stands for Exact Match. VQA stands for VQA Score. R@5 stands for Recall@5.

context learning paradigm (Hu et al., 2023; Shao et al., 2023; Yang et al., 2022), and the extensive multimodal model like Flamingo-80B (Alayrac et al., 2022). This notable performance improvement can be attributed to the innovative multimodal retrieval method implemented in LLM-RA, which enhances retrieval recall by focusing on key visual entities, in contrast to existing methods such as CLIP (Lin et al., 2022), DPR text retrievers (Wu and Mooney, 2022), and Wikipedia data pretrained multimodal retrievers (Lin et al., 2024a).

Results on Infoseek. The performance of LLM-RA on the Infoseek dataset is presented in Table 2. Given that the questions are derived from Wikipedia knowledge, answering Infoseek necessitates fine-grained and often long-tail knowledge. Consequently, even meticulously engineered large-scale multimodal models exhibit subpar performance on this dataset. Our proposed approach, LLM-RA, achieves a notable Infoseek Score of 23.14, surpassing the previous SOTA (22.1 achieved by PaLI-X (Chen et al., 2023c)) by a margin of 1.04. Remarkably, LLM-RA demonstrates comparable performance to PaLI-X on questions with unseen entities. Noteworthy is that equally requiring fine-tuning, LLM-RA shows generalization capabilities on previously unseen knowledge, comparable to models with 1-2 orders larger parameters.

5.2 Ablation Study

We elaborate on the contributions of each design in LLM-RA to the final performance.

Effects of key visual entity. In our experimentation, we remove the regions of interest (ROIs) associated with key visual entities in LLM-RA during retrieval. This setup leads to a 3.13% decrease in Recall@5 on OK-VQA dataset, as depicted in Table 1. Consequently, the final VQA Score decreases by 2.31, and the EM decreases by 1.88. Similarly, for Infoseek dataset, as indicated in Table 2, the removal of ROIs of key visual entities results in a 7.53% decrease in Recall@5, leading to a reduction of 2.29 in the final Infoseek Score. These results underscore the significance of leveraging LLM's reasoning capability to eliminate redundant visual information and extract key visual entities, thereby significantly enhancing retrieval performance and enhancing RA-VQA systems. The results indicate that compared to OK-VQA, key visual entities have a more significant impact on improving the retriever's recall on the Infoseek dataset, although the overall improvement in VQA metrics is approximately similar. We believe the actual enhancement in retrieval performance on OK-VQA dataset might surpass what is reported in Table 1. Unlike Infoseek, OK-VQA does not provide golden knowledge documents but uses pseudo-relevant documents with

Method	Knowledge Source	Finetune	R@5	Unseen-Q	Unseen-E	Overall
OFA (Wang et al., 2022)	-	✓	-	14.8	9.7	11.9
LLaVA-1.5 (Liu et al., 2023a)	-	✓	-	19.4	16.7	17.9
PaLI (Chen et al., 2022)	-	✓	-	20.7	16.0	18.1
CLIP + FID (Chen et al., 2023d)	Infoseek	✓	-	20.7	18.1	19.3
Systems based on very large models (≥50B parameters)						
CLIP + PaLM-540B (Chen et al., 2023d)	Infoseek	✗	-	21.9	18.6	20.1
PaLI-X-55B (Chen et al., 2023c)	-	✓	-	23.5	20.8	22.1
LLM-RA (RA-VQA system with LLM-assisted multimodal retrieval)						
LLM-RA	Infoseek	✓	**47.31**	**26.12**	**20.90**	**23.14**
w/o Key visual entity	Infoseek	✓	39.78	24.68	18.05	20.85
w/o Independent VR	Infoseek	✓	37.52	24.07	18.08	20.65
w/o Key visual entity & Independent VR	Infoseek	✓	32.47	22.40	17.25	19.49
w/o External knowledge	Infoseek	✓	-	18.23	14.10	15.95

Table 2: Performance of the proposed method on Infoseek. VR stands for visual representation. R@5: Recall@5. Unseen-Q stands for Infoseek Score on samples with unseen questions. Unseen-Q stands for Infoseek Score on samples with unseen entity categories.

target answers, which may not be truly informative for answering questions. Leveraging key visual entities increases the likelihood of the retriever capturing genuinely relevant knowledge. Given that this singular comparison may not fully reflect the advantages of LLM-RA, we further compare different ROI extraction methods below.

Effects of independent visual representation. We examine the impact of independent visual representation on multimodal retrieval performance. In contrast to the independent visual representation utilized in LLM-RA, we sum up all visual and textual embeddings into a one-dimensional query for retrieval, like DPR (Karpukhin et al., 2020). As depicted in Table 1 and Table 2, the absence of independent visual representation results in a notable decline in performance across both benchmarks. Specifically, on OK-VQA and Infoseek, there is a reduction of 4.24% and 9.79% in Recall@5, respectively, even falling below the baseline system that does not incorporate ROIs of key visual entities. These results highlight the importance of independently representing key visual entities extracted based on the reasoning capability of LLM. The lack of such independent representation leads to mutual interference between different fine-grained visual details, resulting in degraded retrieval performance.

Effects of external knowledge. We further demonstrate the effectiveness of retrieval augmentation by comparing the baseline model without external knowledge with their retrieval-augmented counterparts. As illustrated in Table 1 and Table 2, when not utilizing external knowledge and solely fine-tuned on the training set, the baseline model achieves 53.74 VQA Score on OK-VQA and 15.95 overall Infoseek Score, respectively. With the integration of LLM-RA, the baseline models experience enhancements of 9.55 in OK-VQA and 7.19 in Infoseek.

Obj. N.	W/ KVE	OK-VQA (GS)		Infoseek	
		R@5	R@10	R@5	R@10
1-3	✗	88.75	94.32	41.59	49.91
	✓	**90.65**	**94.52**	**47.89**	**54.19**
4-9	✗	87.02	92.15	39.85	48.37
	✓	**90.23**	**95.16**	**47.23**	**53.63**
9+	✗	86.13	91.54	37.12	46.94
	✓	**90.12**	**94.13**	**46.54**	**53.31**

Table 3: Performance of the proposed method on a subset of datasets with different numbers of objects. Obj. N.: Object Num; W/ KVE: With key visual entity, i.e., using ROIs of key visual entities extracted by LLM-RA during retrieval; GS: Google Search; R@5: Recall@5.

5.3 Analysis

We conduct several experiments to analyze the efficacy of the proposed method in enhancing retrieval performance, thereby strengthening the RA-VQA system.

Performance under different object number. We analyze the performance improvement of LLM-RA on images with varying degrees of information redundancy. Images that contain a greater

	OK-VQA (GS)		Infoseek	
	R@5	R@10	R@5	R@10
W/o ROIs	87.24	92.78	39.78	48.62
Random ROIs	87.02	92.69	38.29	47.97
Evenly-split ROIs	87.14	92.73	38.63	48.12
All ROIs	88.29	93.76	41.78	49.73
Q-parsed ROIs	88.03	93.50	43.45	50.57
LLM-RA	**90.30**	**94.91**	**47.31**	**53.78**

Table 4: Impact of Different ROIs Extraction Strategies on retriever performance. GS stands for Google Search. R stands for Recall.

number of objects typically present more redundant information in addition to the key visual entities. Therefore, we divide the test dataset into multiple subsets based on the number of objects in the images to assess the improvement in retrieval performance of LLM-RA across different information redundancy levels. OK-VQA, annotated on the MSCOCO dataset, provides object annotations for each image, allowing direct access to the object count. For the Infoseek dataset, we use a well-performing object detector (Carion et al., 2020) to determine the object count. As shown in Table 3, experimental results indicate that LLM-RA yields greater gains for subsets with a larger number of objects. This finding aligns with our expectation that removing redundant visual information from images and emphasizing relevant visual details in the visual embeddings in the query enhances retrieval performance.

Superiority of key visual detail extraction method. To demonstrate that the performance gain from key visual entities results from including question-related finer-grained visual details during retrieval, rather than merely increasing the number of features, We compare the impact of our method on retrieval performance with other ROI extraction approaches. The compared methods include: 1) Randomly cropping patches larger than 100 × 100 pixels from the image as ROIs (random ROIs); 2) Uniformly dividing the image into ROIs (evenly-split ROIs); 3) Obtaining ROIs based on well-performing object detectors (all ROIs); 4) Conducting visual grounding solely based on entities parsed from the question to obtain ROIs (Q-parsed ROIs); 5) Method in LLM-RA (key ROIs). As shown in Table 4, ROIs extracted using our key visual detail extraction methods outperform the other methods. This indicates that LLM-assisted visual details extraction effectively removes redundant visual information from the query while retaining question-relevant finer-grained visual details, thereby enhancing the performance of the RA-VQA system.

	OK-VQA (GS)		Infoseek	
	R@5	R@10	R@5	R@10
Shikra-GCoT	88.17	93.65	41.93	49.87
LLM-RA	**90.30**	**94.91**	**47.31**	**53.78**

Table 5: Comparison of our pipeline-based approach, LLM-RA, and the end-to-end approach Shikra-GCoT, which employs Shikra (Chen et al., 2023b) to generate bounding box of key visual entities as input of multimodal retriever. GS: Google Search. R: Recall.

Comparison of pipeline and end-to-end approach. We explore the superiority of LLM-RA, the pipeline-based approach, over the end-to-end approach which employs VLMs to directly extract the bounding boxes of key visual entities relevant to the problem. We compared our pipeline approach with the end-to-end approach using Shikra (Chen et al., 2023b), which can generate bounding boxes of key visual entities as input (grounding CoT), called Shikra-GCoT. The results in Table 5 show that the pipeline approach performs better compared to the end-to-end approach based on Shikra. To investigate the reasons, we sample examples from OK-VQA and Infoseek to analyze Shikra's output. We find that for some questions, Shikra fails to perform grounding CoT, likely due to the prevalence of fine-grained knowledge questions in these datasets that are out of distribution (OOD) for Shikra. However, there is currently limited research in the multimodal field addressing the OOD problem for grounding CoT. In contrast, The advantage of the pipeline approach is its ability to achieve good performance by leveraging already mature components.

Performance on human judged complex KI-VQA samples. We further investigate the performance of LLM-RA on complex KI-VQA samples. Complex KI-VQA samples are defined as those where: 1) Humans cannot easily answer the questions without knowledge retrieval; 2) Key visual entities required for answering the questions are obscured by significant redundant visual information; 3) Questions do not specify the entities they focus on, necessitating cross-modal reasoning. We randomly select 400 VQA samples from two KI-VQA datasets, requiring human subjects to answer the questionnaire in Ap-

	OK-VQA (GS)		Infoseek	
	Count	Ratio	Count	Ratio
Complex samples	54	13.5%	98	23.5%

Table 6: The count and ratio of human-judged complex KI-VQA samples among 400 randomly selected samples from OK-VQA and Infoseek. GS: Google Search.

	OK-VQA* (GS)		Infoseek*	
	R@5	R@10	R@5	R@10
W/o ROIs	64.81	75.92	21.43	28.57
LLM-RA	81.48	88.89	44.90	51.02

Table 7: LLM-RA's performance on human-judged complex KI-VQA samples. The symbol * indicates that the test set is a subset of OK-VQA and Infoseek selected based on human judgment.

pendix B to identify complex KI-VQA samples. The count and ratio of samples judged as complex in the two datasets are shown in Table 6. The performance of LLM-RA on human-judged complex KI-VQA samples is presented in Table 7. On OK-VQA, Recall@5 increases from 64.81% to 81.48%, while on Infoseek, it increases from 21.43% to 44.90%. These results indicate that for complex visual questions involving key entity reasoning and significant redundant visual information, LLM-RA significantly outperforms baseline systems that do not utilize key visual entities. This suggests that LLM-RA has broader applications in real-world scenarios.

6 Conclusion

In this paper, we propose LLM-RA, an LLM-assisted multimodal retrieval approach for enhancing RA-VQA systems. Leveraging LLM, key visual entities are extracted to highlight question-relevant visual details while removing irrelevant redundant visual information. The independent representation of key visual entities during multimodal joint retrieval ensures there is no mutual interference among key information, thereby enhancing retrieval accuracy. Experimental results demonstrate that our approach outperforms other strong retrieval-enhanced VQA systems and is comparable or even superior to state-of-the-art large-scale multimodal models with 1-2 orders of magnitude more parameters.

Limitations

Firstly, due to computational resource constraints, we do not conduct experiments with LLMs exceeding 13B parameters. Consequently, LLM-RA is designed based on a 7B LLM. Secondly, our approach focuses solely on enhancing the RA-VQA system by improving the multimodal knowledge retriever. Further exploration is needed from other perspectives, such as answer generation. Thirdly, the evaluation datasets, OK-VQA and Infoseek, include a limited number of VQA samples requiring both cross-modal reasoning and long-tail knowledge handling, despite such visual questions being common in real-world scenarios. Consequently, the performance of LLM-RA is not comprehensively evaluated. Therefore, it is essential to explore evaluation methods that involve both cross-modal reasoning and long-tail knowledge handling in multimodal systems.

Acknowledgments

We thank our colleagues Zixuan Ren, Min Xiao, and Dr. Jinliang Lu, who provided valuable insights into the formation of motivation in this work. We thank our colleagues Yaochen Zhu, Jianghao Chen, and Junhong Wu for their helpful suggestions during the writing stage of the paper. Finally, we thank all reviewers for their detailed reviews and insightful comments.

The research work was supported by the National Key R&D Program of China (No. 2022ZD0160602) and the Natural Science Foundation of China (No. 62122088).

References

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

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023b. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. 2023c. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

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

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

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.

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, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2023. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2963–2975.

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

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 Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen Tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 6769–6781. Association for Computational Linguistics (ACL).

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.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

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.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022b. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

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.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2024a. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36.

Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. 2024b. 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. Association for Computational Linguistics.

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in Neural Information Processing Systems*, 35:10560–10571.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

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.

Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2023. Rephrase, augment, reason: Visual grounding of questions for vision-language models. In *The Twelfth International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Zixuan Ren, Yang Zhao, and Chengqing Zong. 2023. Towards informative open-ended text generation with dynamic knowledge triples. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3189–3203.

Alireza Salemi, Juan Altmayer Pizzorno, and Hamed Zamani. 2023. A symmetric dual encoding dense retrieval framework for knowledge-intensive visual question answering. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 110–120.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734.

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.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983.

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.

Zhongfan Sun, Yongli Hu, Qingqing Gao, Huajie Jiang, Junbin Gao, Yanfeng Sun, and Baocai Yin. 2023. Breaking the barrier between pre-training and fine-tuning: A hybrid prompting model for knowledge-based vqa. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4065–4073.

Rongchuan Tang, Yang Zhao, Chengqing Zong, and Yu Zhou. 2023. Multilingual knowledge graph completion with language-sensitive multi-graph attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10508–10519.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2023. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*.

Jialin Wu and Raymond Mooney. 2022. Entity-focused dense passage retrieval for outside-knowledge visual question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8061–8072.

Alexandros Xenos, Themos Stafylakis, Ioannis Patras, and Georgios Tzimiropoulos. 2023. A simple baseline for knowledge-based visual question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14871–14877.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023. Large language models for generative information extraction: A survey. *arXiv preprint arXiv:2312.17617*.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2024. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems*, 36.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3081–3089.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Xinlong Wang, and Jingjing Liu. 2023a. Capsfusion: Rethinking image-text data at scale. *arXiv preprint arXiv:2310.20550*.

Shi Yu, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2023b. Openmatch-v2: An all-in-one multimodality plm-based information retrieval toolkit. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3160–3164.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

A Prompt Templates

In this section, we provide detailed prompt templates for LLM-RA concerning processes involving LLM and LVLM, including image caption and LLM-assisted key entity extraction. The prompt templates are as follows:

Prompt Template For Image Caption

```
<Img>#I</Img> describe this picture in
detail.
```

Prompt Template for Reasoning

```
Given a description of the image, output
the entities along with their attribute
that the question might focus on, based on
the question below. Do not output entities
in the image description that are not
relevant to the problem.

Note: Output common objects and group
them into general categories that are not
duplicated, merging essentially similar
entities. Avoid extracting abstract or
non-specific entities.

Examples:
This image shows a small kitchen with
various appliances, including a microwave,
toaster, and stove. The stove has a burner
on it, and there are several cups and bowls
sitting around the countertops. There is
also a large white refrigerator in the
corner of the room.
questions: How do I open the device located
at the top of the image?
outputs: ["the device": "located at the
top of the image"]

captions: The image depicts a woman in a
tennis outfit and holding a tennis racket,
standing next to a large blue banner with
the words "JPMorgan" and "US Open" written
on it. This is likely a professional tennis
tournament where she is participating or
attending as an attendee
questions: Who won this years championship
in this sport?
outputs: ["Large Blue Banner": "Displaying
the words "JPMorgan" and "US Open.",
"woman": "in a tennis outfit and holding a
tennis racket]

captions: The image shows a dining table
set for breakfast with several bowls and
plates on it. The table is in a cozy
room with chairs surrounding it. A lamp is
placed on a side table near the tablecloth,
adding warmth and light to the space.
questions: What is the square of cloth
under the plate called?
outputs: ["Square of Cloth": "Under the
plate on the dining table"]

Caption: #C
Question: #Q
Outputs:
```

In our prompt template, #I denotes the visual embedding corresponding to the given image. #C denotes caption of the input image, and #Q denotes The input question. For both the OK-VQA and Infoseek datasets, we utilized 5 carefully crafted in-context examples to guide the LLM in extracting key visual entities relevant to the question and their attributes based on the captions. To utilize the multimodal model for generating answers based on questions and retrieved knowledge, we adopt the following prompt template (#Doc denotes the retrieved document from knowledge base):

Prompt Template For Answer Generation

```
Question: #Q Caption: #C Kownledge: #Doc
Answer:
```

After tokenizing the prompt, we concatenated visual embeddings with textual embeddings as the input to the multimodal model. It's noteworthy that, due to the inclusion of captions in the answer prompt template, we consistently used captions as textualized visual information in all ablation experiments concerning VQA performance evaluation to ensure fair comparisons. Additionally, the RA-VQA systems selected for comparison also utilized captions generated by multimodal models, with some even leveraging additional textural visual information such as OCR outputs.

B Questionnaire for Human Judgement of Complex KI-VQA Samples

In this section, we present the questionnaire used for complex KI-VQA sample selection based on human judgment, which is described in Section 5.3. The questionnaire is as follows:

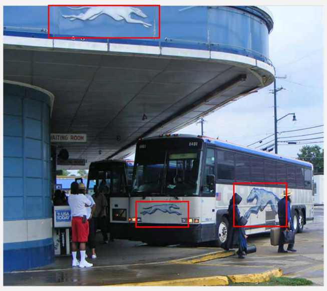
Prompt Template for Reasoning

```
<A VQA Sample>

Given the VQA sample, please answer
the following questions:

1) Without using any external tools
(such as search engines), based on the
given image, can you independently and
confidently answer the provided question?
(yes/no)
```

2) Is the background in the image excessively cluttered? (yes/no)

3) For VQA Samples, there are key visual entities in the image related to the question. For example, the visual entity "brand" in the following image (a) is helpful to answer the corresponding question. Apart from these visual entities, are there three or more redundant visual entities in the image, similar in size or even larger than the key visual entities, unrelated to the question? (yes/no)

Question: *What is the name the **bus company**?*



(a)

4) In the given question, are there unclear references to the key visual entities mentioned in 3) ? (yes/no) For instance, for the visual entity "gun" in the following figure (b), "equipment" is an ambiguous reference.

Question: *In which year was this **equipment** retired?*

(b)

5) For the given question, Is reasoning with the image necessary to determine the key visual entities mentioned in 3) ? (yes/no) For instance, in the above figure (a), the question does not clearly state the visual entity of interest in the image, requiring inference based on "bus company" in the given question to determine that the "brand" in the image is the key visual entity.

Hyper-parameters	Value
CLIP	clip-vit-base-patch32
N_r	3
N_D	512
N_Q	640
d_v	768
d_L	128
Learning rate	10^{-5}
Training steps	10^4
Batch size	30
GPUs	1
Gradient accumulation	2
Optimizer	Adam

Table 8: The hyper-parameters used for the multimodal retriever.

Hyper-parameters	Value
Learning rate	10^{-4}
Training steps	4.8×10^3
Batch size	1
GPUs	1
Gradient accumulation	16
Retriever Top-K	5
Optimizer	Adam

Table 9: The hyper-parameters used for the answer generator.

For the questionnaire described above, if the answer to 1) is yes, and either 2) or 3) is yes, along with one of 4) or 5) being yes, then the sample is considered to meet the definition of a complex VQA sample outlined in Section 5.3

C More Experimental Details

Model Checkpoints. We utilized MiniGPT-v2 (Chen et al., 2023a) for Image Captioning and leveraged its frozen LLM `Llama-2-7b-chat-hf` for inferring key visual entities in images. Then we select Grounding-DINO-L (Liu et al., 2023b) for visual grounding. We adopted `ColBERTv2` and `openai/clip-vit-base-patch32` checkpoints to initialize the text encoder and vision encoder in the multimodal retriever. For training the answer generator that utilizes retrieval knowledge, we employed the `Salesforce/blip2-flan-t5-xl` model, which has approximately 4B parameters. All training processes can be implemented on a single Nvidia A100 (80G) GPU.

Hyper-paraments. The hyper-parameter settings in the experiments are shown in Table 8 and Table 9.For the CLIP visual encoder of the multimodal retriever, $d_V = 768$. The mapping network used in Section 3.2 is a two-layer multilayer perceptron, projecting the [CLS] token output by CLIP into $N_v = 32$ visual embeddings. The depth of Visual Embeddings, $d_L = 128$, matches that of text embeddings extracted by the text encoder. The number of Key ROIs used for the OK-VQA and Infoseek datasets is $N_r = 3$. Regarding knowledge representation E_D, the number of its embeddings is $N_D = 512$. Similarly, within the multimodal retriever, the number of text embeddings $N_L = 512$ in the query E_Q. Therefore, the total number of embeddings in query E_Q is

$$N_Q = N_L + (1 + N_r) \times N_v = 640. \qquad (4)$$

For both benchmarks, during training, we utilized the Adam optimizer with a learning rate of $lr = 10^{-5}$, a batch size of 30, and trained for 10k steps with gradient accumulation steps set to 2. When training the answer generator, we employed a learning rate of $lr = 10^{-4}$, a batch size of 1, and trained for 4.8k steps with gradient accumulation steps set to 16. During the training of the answer generator, we efficiently fine-tuned the Salesforce/blip2-flan-t5-xl model using default parameters from Lora. It's worth noting that, due to the large size of the Infoseek training set, we randomly sampled 100k samples from the dataset for training. During answer generation, we utilized the top 5 documents with the highest similarity to the query.

D Effect of different Key ROI Numbers

We also investigated the impact of different ROIs numbers on the performance of LLM-RA. As described in Section 5.3, we conducted retrieval using all ROIs obtained from object detectors for comparison. The ROIs were sorted based on the predicted probabilities from visual grounding of key visual entities and object detection in descending order. When the number of ROIs is set to K, it represents selecting the top K ROIs. Padding is applied when the number of ROIs in some images is less than K.

The performance of LLM-RA under different ROI numbers on Infoseek and OK-VQA are illustrated in Figure 3 and Figure 4. The results indicate that our proposed Key ROI extraction method consistently outperforms using all ROIs obtained from object detection in both KI-VQA benchmarks. By extracting key ROIs, redundant information unrelated to the questions in the image is eliminated. Therefore, the performance of LLM-RA peaks when the ROI number is set to 3 compared to using all ROIs. We set the hyperparameter key ROI number to 3 based on these experiments.

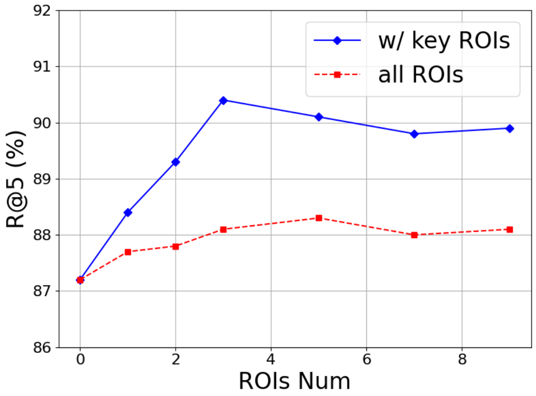


Figure 3: Retrieval Performance of LLM-RA on OK-VQA with different numbers of key ROIs.

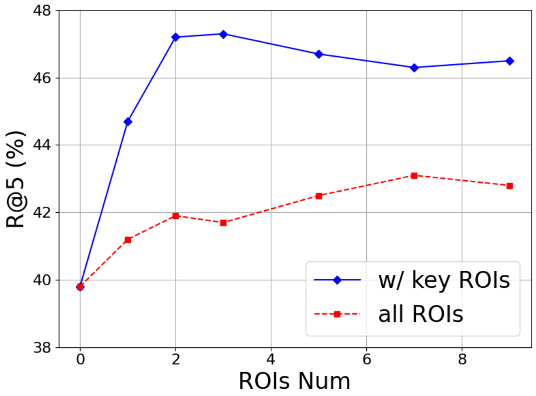
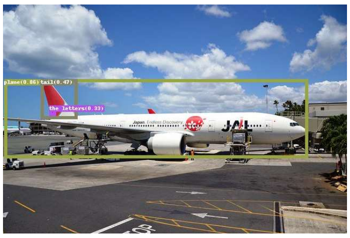


Figure 4: Retrieval Performance of LLM-RA on Infoseek with different numbers of key ROIs.

E Case Study

Figures 5 to 7 illustrate the Case Study for LLM-RA. They compare LLM-RA with a baseline that does not utilize Key Visual Entities. We showcase the VQA outputs and various intermediate results for both LLM-RA and the comparison methods. We also provide explanations for each case. The term "Grounding prompt" in Figures 5 to 7 refers to the prompt used when adopting visual grounding model to extract ROIs of key visual entities.



Dataset: OK-VQA
Question: What airline is this?
Caption: The picture shows a large white airplane parked on a runway...
Grounding prompt: The airplane that parked on a runway.
Answer: Jal, Japan airline

LLM-RA w/o Key ROIs

hawaiian airlines, hawaii's largest and longest-serving airline, offers non-stop service to hawaii from the u. s. mainland and international destinations.

airline summarythe largest airline in europe, frankfurt-based lufthansa (lh) flies non-stop to about 215 destinations. this includes 18 points within germany, as well as 78 countries within europe, ...

lufthansa is certified as a 5-star airline for quality of seats, amenities, catering, ife, cleanliness, and cabin staff and ground staff service standards.

Generation: alitalia ✗

LLM-RA

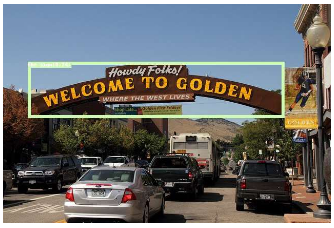
about japan airlines **japan airlines (jal)** is the flag carrier of japan and is the second-largest airline in japan. japan airlines (jal) is the flag carrier of japan and is the second-largest airline in japan...

this is **jal's (japan airlines)** corporate website, where you can view corporate information, safety/ flight information, and sustainability information, etc...

japan airlines (jal) will be dropping \"ladies and gentlemen\" in favor of more inclusive greetings like \"attention all passengers\" and \"welcome, everyone\" from october 1 on flights and in ...

Generation: jal ✓

Explanation: Based on the caption and the term "Airline" in the question, the LLM accurately identified the key visual entity "airplane" and performed visual grounding. Leveraging the ROI associated with "airplane", the relevant document "Japan airline" related to the correct answer was successfully captured.



Dataset: OK-VQA
Question: Where was this taken?
Caption: the picture depicts a busy street scene... and a sign visible above the street....
Grounding prompt: The sign that visible above the street.
Answer: colorado

LLM-RA w/o Key ROIs

phoenix (3tv/cbs5) -- phoenix police are investigating after a man was found shot in a pickup truck outside a circle k store.it happened sunday at around 8:40 p.m.

gunshots can be heard in video shot at the scene of sunday night's shootings near mandalay bay on the las vegas strip.

back to top. taken , 740 dulaney valley road, towson, md, 21204, united states410 337-6856welcome@shoptaken. com. powered by squarespace.

Generation: golden gate ✗

LLM-RA

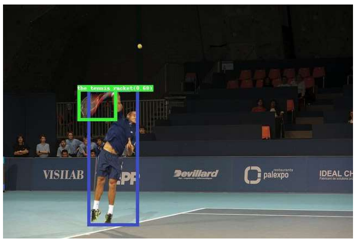
history **colorado**'s photography collection contains approximately 1 million images documenting the history of colorado and the american west from the 1840s to the present day...

may 4, 2020 \u00b7 the photos were taken at the store on mission gorge road in santee saturday afternoon.

san francisco is a great city for photography lovers. of course, the iconic golden gate bridge is a perennial favorite, but there are plenty of other great places to take pictures in sf.

Generation: colorado ✓

Explanation: Utilizing LLM, the key visual entity "sign" was extracted from complex visual scenes, aiding the retriever in capturing the golden knowledge "colorado".



Dataset: OK-VQA
Question: What is the name of the player in this picture?
Caption: the image captures a man in a blue shirt and black shorts playing tennis on an indoor court...
Grounding prompt: The man that playing tennis on an indoor court.
Answer: roger federer

LLM-RA w/o Key ROIs

in football, all you really know about the player is their name and there number. their face is hidden by a facemask. for example, there are guys like ladanian tomlinson, ...

for those unfamiliar with the term, \u201cplayer-coach\u201d refers to a manager role that combines the more traditional tasks of managing a team ...

big name power forwards currently in the nhl include jamie benn, blake wheeler, and aleksander barkov...

Generation: rafael nada ✗

LLM-RA

espn summarizes, \u201cif a player does not dress to participate in a game, he must dress in a manner suitable for a coach...

Roger Federer (born August 8, 1981, Basel, Switzerland) is a Swiss tennis player who dominated the sport in the early 21st century with his exceptional all-around game.

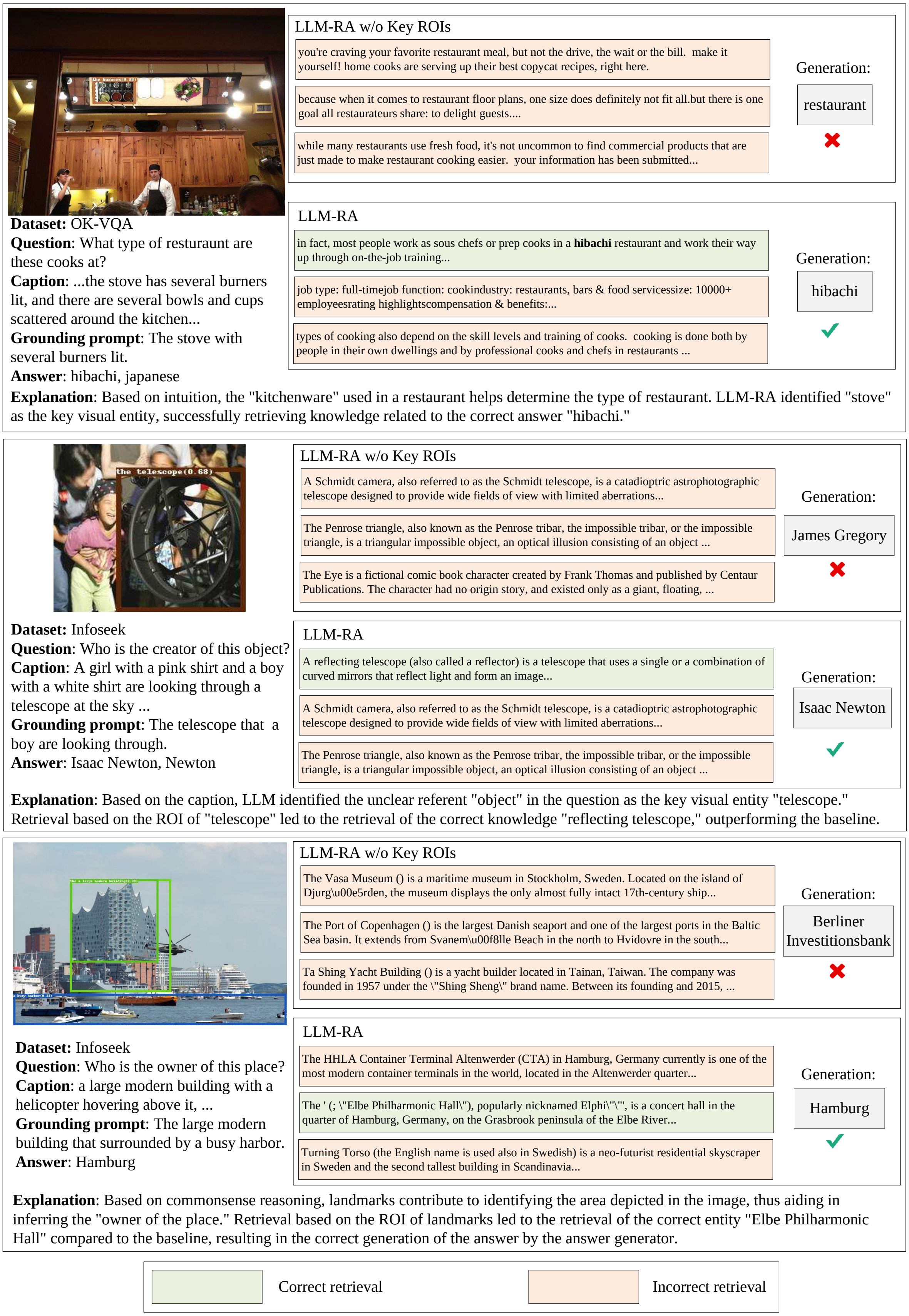
Learn about the life and achievements of Roger Federer, the Swiss tennis legend who holds the record for most Grand Slam titles...

Generation: roger federer ✓

Explanation: Based on the query focus on "player," LLM identified the entity "The man that playing tennis" of interest, leveraging the visual details within the ROI, enabling the retrieval of accurate knowledge related to "Roger Federer."

🟩 Correct retrieval   🟧 Incorrect retrieval

Figure 5: Case study group 1. Each case is accompanied by an explanation. Please zoom in for optimal visual clarity.



Dataset: OK-VQA
Question: What type of resturaunt are these cooks at?
Caption: ...the stove has several burners lit, and there are several bowls and cups scattered around the kitchen...
Grounding prompt: The stove with several burners lit.
Answer: hibachi, japanese

LLM-RA w/o Key ROIs

you're craving your favorite restaurant meal, but not the drive, the wait or the bill. make it yourself! home cooks are serving up their best copycat recipes, right here.

because when it comes to restaurant floor plans, one size does definitely not fit all.but there is one goal all restaurateurs share: to delight guests....

while many restaurants use fresh food, it's not uncommon to find commercial products that are just made to make restaurant cooking easier. your information has been submitted...

Generation: restaurant ✗

LLM-RA

in fact, most people work as sous chefs or prep cooks in a **hibachi** restaurant and work their way up through on-the-job training...

job type: full-timejob function: cookindustry: restaurants, bars & food servicessize: 10000+ employeesrating highlightscompensation & benefits:...

types of cooking also depend on the skill levels and training of cooks. cooking is done both by people in their own dwellings and by professional cooks and chefs in restaurants ...

Generation: hibachi ✓

Explanation: Based on intuition, the "kitchenware" used in a restaurant helps determine the type of restaurant. LLM-RA identified "stove" as the key visual entity, successfully retrieving knowledge related to the correct answer "hibachi."

Dataset: Infoseek
Question: Who is the creator of this object?
Caption: A girl with a pink shirt and a boy with a white shirt are looking through a telescope at the sky ...
Grounding prompt: The telescope that a boy are looking through.
Answer: Isaac Newton, Newton

LLM-RA w/o Key ROIs

A Schmidt camera, also referred to as the Schmidt telescope, is a catadioptric astrophotographic telescope designed to provide wide fields of view with limited aberrations...

The Penrose triangle, also known as the Penrose tribar, the impossible tribar, or the impossible triangle, is a triangular impossible object, an optical illusion consisting of an object ...

The Eye is a fictional comic book character created by Frank Thomas and published by Centaur Publications. The character had no origin story, and existed only as a giant, floating, ...

Generation: James Gregory ✗

LLM-RA

A reflecting telescope (also called a reflector) is a telescope that uses a single or a combination of curved mirrors that reflect light and form an image...

A Schmidt camera, also referred to as the Schmidt telescope, is a catadioptric astrophotographic telescope designed to provide wide fields of view with limited aberrations...

The Penrose triangle, also known as the Penrose tribar, the impossible tribar, or the impossible triangle, is a triangular impossible object, an optical illusion consisting of an object ...

Generation: Isaac Newton ✓

Explanation: Based on the caption, LLM identified the unclear referent "object" in the question as the key visual entity "telescope." Retrieval based on the ROI of "telescope" led to the retrieval of the correct knowledge "reflecting telescope," outperforming the baseline.

Dataset: Infoseek
Question: Who is the owner of this place?
Caption: a large modern building with a helicopter hovering above it, ...
Grounding prompt: The large modern building that surrounded by a busy harbor.
Answer: Hamburg

LLM-RA w/o Key ROIs

The Vasa Museum () is a maritime museum in Stockholm, Sweden. Located on the island of Djurg\u00e5rden, the museum displays the only almost fully intact 17th-century ship...

The Port of Copenhagen () is the largest Danish seaport and one of the largest ports in the Baltic Sea basin. It extends from Svanem\u00f8lle Beach in the north to Hvidovre in the south...

Ta Shing Yacht Building () is a yacht builder located in Tainan, Taiwan. The company was founded in 1957 under the \"Shing Sheng\" brand name. Between its founding and 2015, ...

Generation: Berliner Investitionsbank ✗

LLM-RA

The HHLA Container Terminal Altenwerder (CTA) in Hamburg, Germany currently is one of the most modern container terminals in the world, located in the Altenwerder quarter...

The ' (; \"Elbe Philharmonic Hall\"), popularly nicknamed Elphi\"\"", is a concert hall in the quarter of Hamburg, Germany, on the Grasbrook peninsula of the Elbe River...

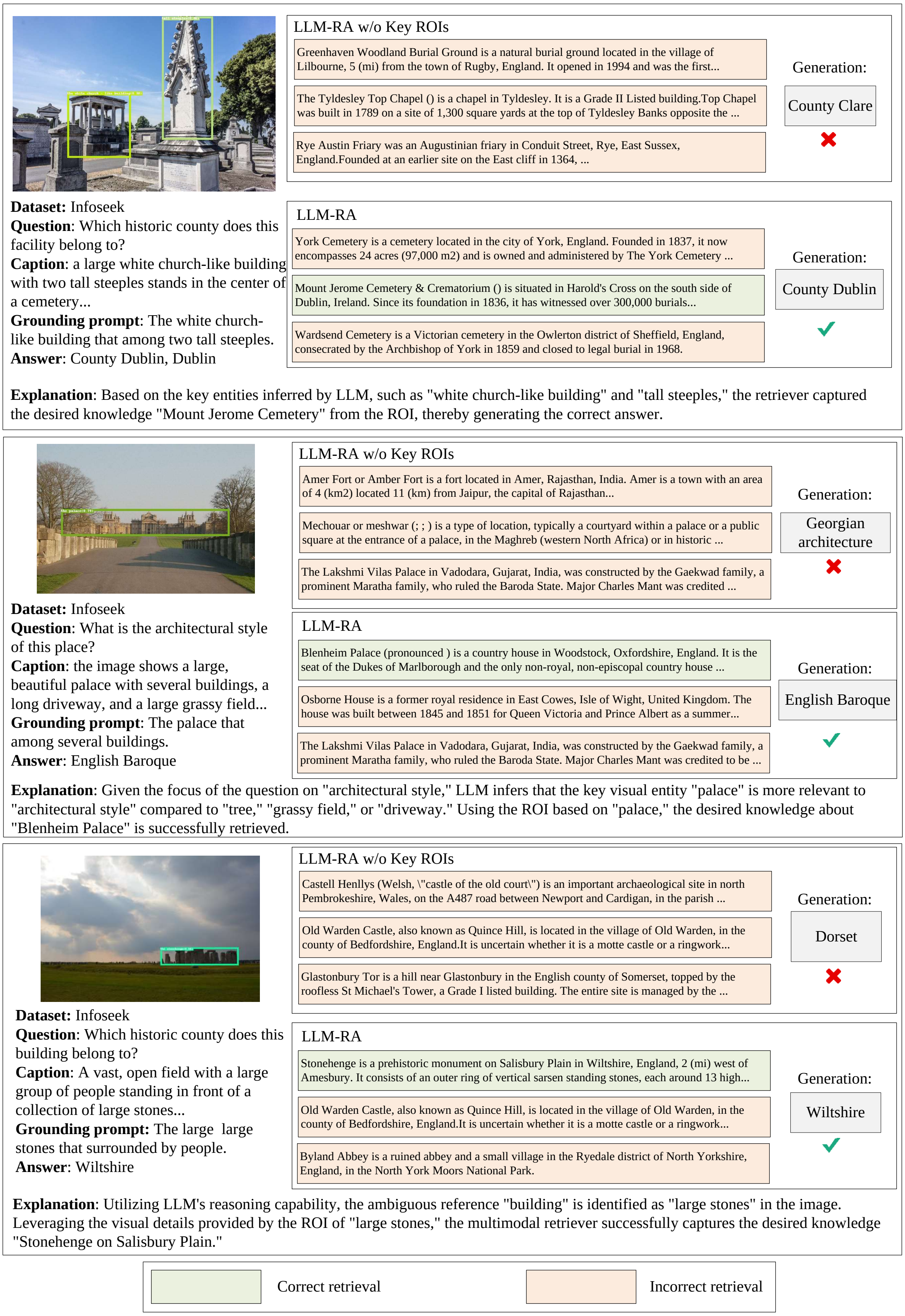
Turning Torso (the English name is used also in Swedish) is a neo-futurist residential skyscraper in Sweden and the second tallest building in Scandinavia...

Generation: Hamburg ✓

Explanation: Based on commonsense reasoning, landmarks contribute to identifying the area depicted in the image, thus aiding in inferring the "owner of the place." Retrieval based on the ROI of landmarks led to the retrieval of the correct entity "Elbe Philharmonic Hall" compared to the baseline, resulting in the correct generation of the answer by the answer generator.

🟩 Correct retrieval 🟧 Incorrect retrieval

Figure 6: Case study group 2. Each case is accompanied by an explanation. Please zoom in for optimal visual clarity.



LLM-RA w/o Key ROIs

Greenhaven Woodland Burial Ground is a natural burial ground located in the village of Lilbourne, 5 (mi) from the town of Rugby, England. It opened in 1994 and was the first...

The Tyldesley Top Chapel () is a chapel in Tyldesley. It is a Grade II Listed building.Top Chapel was built in 1789 on a site of 1,300 square yards at the top of Tyldesley Banks opposite the ...

Rye Austin Friary was an Augustinian friary in Conduit Street, Rye, East Sussex, England.Founded at an earlier site on the East cliff in 1364, ...

Generation: County Clare ✖

LLM-RA

York Cemetery is a cemetery located in the city of York, England. Founded in 1837, it now encompasses 24 acres (97,000 m2) and is owned and administered by The York Cemetery ...

Mount Jerome Cemetery & Crematorium () is situated in Harold's Cross on the south side of Dublin, Ireland. Since its foundation in 1836, it has witnessed over 300,000 burials...

Wardsend Cemetery is a Victorian cemetery in the Owlerton district of Sheffield, England, consecrated by the Archbishop of York in 1859 and closed to legal burial in 1968.

Generation: County Dublin ✓

Dataset: Infoseek
Question: Which historic county does this facility belong to?
Caption: a large white church-like building with two tall steeples stands in the center of a cemetery...
Grounding prompt: The white church-like building that among two tall steeples.
Answer: County Dublin, Dublin

Explanation: Based on the key entities inferred by LLM, such as "white church-like building" and "tall steeples," the retriever captured the desired knowledge "Mount Jerome Cemetery" from the ROI, thereby generating the correct answer.

LLM-RA w/o Key ROIs

Amer Fort or Amber Fort is a fort located in Amer, Rajasthan, India. Amer is a town with an area of 4 (km2) located 11 (km) from Jaipur, the capital of Rajasthan...

Mechouar or meshwar (; ;) is a type of location, typically a courtyard within a palace or a public square at the entrance of a palace, in the Maghreb (western North Africa) or in historic ...

The Lakshmi Vilas Palace in Vadodara, Gujarat, India, was constructed by the Gaekwad family, a prominent Maratha family, who ruled the Baroda State. Major Charles Mant was credited ...

Generation: Georgian architecture ✖

LLM-RA

Blenheim Palace (pronounced) is a country house in Woodstock, Oxfordshire, England. It is the seat of the Dukes of Marlborough and the only non-royal, non-episcopal country house ...

Osborne House is a former royal residence in East Cowes, Isle of Wight, United Kingdom. The house was built between 1845 and 1851 for Queen Victoria and Prince Albert as a summer...

The Lakshmi Vilas Palace in Vadodara, Gujarat, India, was constructed by the Gaekwad family, a prominent Maratha family, who ruled the Baroda State. Major Charles Mant was credited to be ...

Generation: English Baroque ✓

Dataset: Infoseek
Question: What is the architectural style of this place?
Caption: the image shows a large, beautiful palace with several buildings, a long driveway, and a large grassy field...
Grounding prompt: The palace that among several buildings.
Answer: English Baroque

Explanation: Given the focus of the question on "architectural style," LLM infers that the key visual entity "palace" is more relevant to "architectural style" compared to "tree," "grassy field," or "driveway." Using the ROI based on "palace," the desired knowledge about "Blenheim Palace" is successfully retrieved.

LLM-RA w/o Key ROIs

Castell Henllys (Welsh, \"castle of the old court\") is an important archaeological site in north Pembrokeshire, Wales, on the A487 road between Newport and Cardigan, in the parish ...

Old Warden Castle, also known as Quince Hill, is located in the village of Old Warden, in the county of Bedfordshire, England.It is uncertain whether it is a motte castle or a ringwork...

Glastonbury Tor is a hill near Glastonbury in the English county of Somerset, topped by the roofless St Michael's Tower, a Grade I listed building. The entire site is managed by the ...

Generation: Dorset ✖

LLM-RA

Stonehenge is a prehistoric monument on Salisbury Plain in Wiltshire, England, 2 (mi) west of Amesbury. It consists of an outer ring of vertical sarsen standing stones, each around 13 high...

Old Warden Castle, also known as Quince Hill, is located in the village of Old Warden, in the county of Bedfordshire, England.It is uncertain whether it is a motte castle or a ringwork...

Byland Abbey is a ruined abbey and a small village in the Ryedale district of North Yorkshire, England, in the North York Moors National Park.

Generation: Wiltshire ✓

Dataset: Infoseek
Question: Which historic county does this building belong to?
Caption: A vast, open field with a large group of people standing in front of a collection of large stones...
Grounding prompt: The large large stones that surrounded by people.
Answer: Wiltshire

Explanation: Utilizing LLM's reasoning capability, the ambiguous reference "building" is identified as "large stones" in the image. Leveraging the visual details provided by the ROI of "large stones," the multimodal retriever successfully captures the desired knowledge "Stonehenge on Salisbury Plain."

🟩 Correct retrieval 🟧 Incorrect retrieval

Figure 7: Case study group 3. Each case is accompanied by an explanation. Please zoom in for optimal visual clarity.