MuKA: Multimodal Knowledge Augmented Visual Information-Seeking

Lianghao Deng[1], **Yuchong Sun**[1], **Shizhe Chen**[2], **Ning Yang**[3], **Yunfeng Wang**[3], **Ruihua Song**[1*]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Inria, École normale supérieure, CNRS, PSL Research University
[3]ZHI-TECH GROUP
lhdeng@ruc.edu.cn, songruihua_bloon@outlook.com

Abstract

The visual information-seeking task aims to answer visual questions that require external knowledge, such as “On what date did this building officially open?". Existing methods using retrieval-augmented generation framework primarily rely on textual knowledge bases to assist multimodal large language models (MLLMs) in answering questions. However, the text-only knowledge can impair information retrieval for the multimodal query of image and question, and also confuse MLLMs in selecting the most relevant information during generation. In this work, we propose a novel framework MuKA which leverages a multimodal knowledge base to address these limitations. Specifically, we construct a multimodal knowledge base by automatically pairing images with text passages in existing datasets. We then design a fine-grained multimodal interaction to effectively retrieve multimodal documents and enrich MLLMs with both retrieved texts and images. MuKA outperforms state-of-the-art methods by 38.7% and 15.9% on the InfoSeek and E-VQA benchmark respectively, demonstrating the importance of multimodal knowledge in enhancing both retrieval and answer generation. [1]

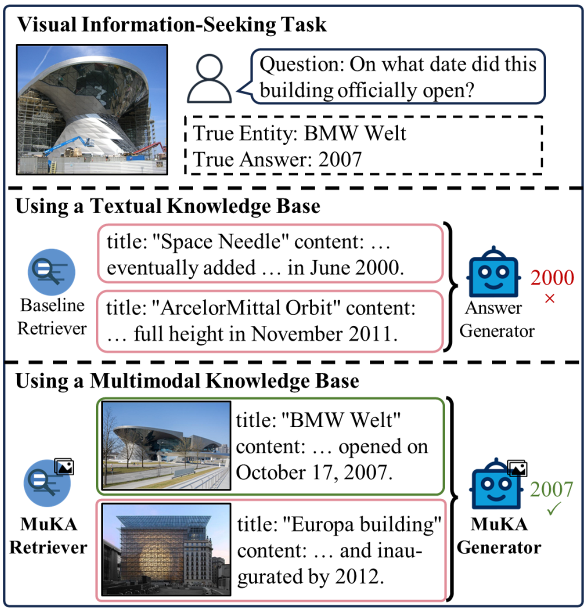

Figure 1: Illustration of the challenge in retrieving documents from a textual knowledge base that match the entity shown in the visual information-seeking question. By utilizing a multimodal knowledge base, our MuKA retriever and answer generator identify the accurate multimodal document, ultimately providing the correct answer.

1 Introduction

Recently, Multimodal Large Language Models (MLLMs) (Alayrac et al., 2022; Liu et al., 2024b; Li et al., 2023) have showcased strong capabilities in vision-language understanding and text generation. Although they have achieved impressive performance in various vision-language tasks such as image captioning and general visual question answering (Goyal et al., 2017; Hudson and Manning, 2019), existing MLLMs still struggle with visual information-seeking tasks (Chen et al., 2023; Mensink et al., 2023) which require knowledge-intensive visual question answering. Figure 1 illustrates an example of such a task, where a user asks about an image of a particular building: “On what date did this building officially open?". To answer such questions, models must not only have general knowledge of object names, colors or quantities, but more importantly, should be equipped with detailed knowledge associated with the specific entity in the image.

Memorizing all the detailed knowledge in MLLMs proves to be challenging (Chen et al., 2023). To address this, previous works have adopted a Retrieval-Augmented Generation (RAG) framework (Caffagni et al., 2024; Yan and Xie,

*Corresponding author.
[1]https://github.com/lhdeng-gh/MuKA

Proceedings of the 31st International Conference on Computational Linguistics, pages 9675–9686
January 19–24, 2025.

2024), which first retrieves relevant documents from a textual knowledge base and then feeds the top-K documents to MLLMs for answer generation. The RAG-based approaches yield promising results by integrating external knowledge with MLLMs' reasoning capabilities. However, existing methods solely utilize the textual knowledge base, making it difficult to retrieve relevant documents given the multimodal query of image and question due to the cross-modal gap. For example, as shown in Figure 1, a baseline system with only textual knowledge base retrieves textual documents with dates of buildings, but fails to find the exact building depicted in the image. Moreover, even if a retriever returns the correct document in its top-K ranking list, simply providing all top-K texts to the MLLM can confuse the model to select the most relevant information for answering the question accurately.

In this work, we propose a novel **Mu**ltimodal **K**nowledge **A**ugmented generation (MuKA) framework to address the limitations of existing methods using pure textual knowledge bases. When we humans seek relevant information for a multimodal query of image and question, we compare the query with not only text but also any associated images with the text to ensure accurate retrieval. Our framework is inspired by this principle. To achieve this, we first construct a multimodal knowledge base by automatically pairing textual documents with their corresponding entity images. Then we design a retriever that matches multimodal queries and multimodal documents by a fine-grained interaction. As illustrated in Figure 1, our MuKA retriever effectively ranks the document about the correct building at the top. To further distinguish the correct document from other similar documents ranked at the top, we propose to enhance the context of the MLLM generator with multimodal documents, allowing the generator to select the most relevant knowledge from the top-ranked documents.

Experimental results on two public benchmarks InfoSeek and E-VQA show that our MuKA retriever outperforms the best baseline by 9.3 and 4.0 points in terms of R@5 performance, and our MuKA generators consistently outperform their counterparts that read textual knowledge, implying the effectiveness of multimodal knowledge in answer generation. When used together, our MuKA method improves the state-of-the-art methods by 38.7% and 15.9% on the two datasets respectively.

To summarize, our contributions are three-fold:

- We identify the significance of multimodal knowledge for knowledge-intensive visual information-seeking tasks and construct a multimodal knowledge base to facilitate research in this direction.
- We propose a novel multimodal knowledge augmented generation framework MuKA, which enhances knowledge retrieval by fine-grained multimodal interactions and improves answer generation by enriching contexts with multimodal documents.
- Extensive experiments on the InfoSeek and the E-VQA datasets demonstrate the effectiveness of our MuKA method with multimodal knowledge base.

2 Related Works

Visual Information-Seeking Visual Question Answering (VQA) involves answering questions based on visual context. Traditional VQA benchmarks (Goyal et al., 2017; Hudson and Manning, 2019; Singh et al., 2019) primarily target the assessment of the visual context understanding ability of models. Knowledge-intensive VQA benchmarks (Marino et al., 2019; Schwenk et al., 2022; Wang et al., 2017) elevate this challenge by requiring knowledge related to the visual context. Visual information-seeking, a category of knowledge-intensive VQA, demands more specific and detailed knowledge of the entity presented in the query image. Several datasets have been proposed for this task, including ViQuAE (Lerner et al., 2022), Encyclopedic VQA (E-VQA) (Mensink et al., 2023), and InfoSeek (Chen et al., 2023). To tackle this task, AVIS (Hu et al., 2024) leverages a Large Language Model (LLM) to dynamically strategize the utilization of external tools. RA-VQA-v2 (Lin et al., 2024b) builds a Retrieval-Augmented Generation (RAG) pipeline with its late-interaction knowledge retrieval. In this paper, we build a RAG pipeline over multimodal knowledge bases, and present our results on InfoSeek and E-VQA as the previous work do (Lin et al., 2024c).

Multimodal Large Language Models Multimodal Large Language Models (MLLMs) have demonstrated strong capabilities in visual context understanding and natural language generation. An MLLM typically comprises of a Large Language Model (LLM), a vision encoder and vision-language integration modules. Open-source

LLMs (Raffel et al., 2020; Touvron et al., 2023; Jiang et al., 2023) greatly contribute to the development of MLLMs. Vision encoders are typically pretrained visual backbones (Radford et al., 2021; Sun et al., 2023; Zhai et al., 2023) to encode visual inputs into features. As for the vision-language integration modules, Flamingo (Alayrac et al., 2022) inserts cross-attention layers within the LLM. Recent MLLMs adopt a simpler approach by projecting visual features into the embedding space of LLMs. These projectors may comprise of fully-connected layers (Liu et al., 2024b; Zhu et al., 2023a) and cross-attention blocks (Li et al., 2023; Ye et al., 2023). LLaVA-1.5 (Liu et al., 2024a) employs a MLP module as the projector. VILA (Lin et al., 2024a) adopts a similar approach but reduces the number of visual features through down-sampling. VILA is pre-trained on interleaved image-text corpus (Zhu et al., 2023b) and image-text pairs (Byeon et al., 2022). In this paper, we develop answer generators reading textual or multimodal documents using the LLaVA-1.5 and VILA models.

Retrieval-Augmented Generation Retrieval-Augmented Generation (RAG) augments the inputs of LLMs with retrieved documents (Guu et al., 2020; Lewis et al., 2020), thereby improving performance in knowledge-intensive tasks. Fine-tuning LLMs on document-reading examples facilitates the utilization of retrieved information (Luo et al., 2023; Zhang et al., 2024; Asai et al., 2024). Recent studies have successfully applied RAG to knowledge-intensive vision-language tasks (Lin and Byrne, 2022; Qiu et al., 2024). However, their knowledge retrieval targets are textual, and using multimodal queries to retrieve textual documents poses difficulty in matching long-tail entities with their knowledge. The retrieval for multimodal queries can be conducted in stages (Caffagni et al., 2024), sequentially performing visual (Radford et al., 2021) and textual (Karpukhin et al., 2020; Izacard et al., 2021) retrieval. EchoSight (Yan and Xie, 2024) performs multimodal re-ranking after visual retrieval, but still only passes textual documents for answer generation. Recent studies have developed multimodal retrievers to handle multimodal queries (Wei et al., 2023; Lin et al., 2024b,c). PreFLMR (Lin et al., 2024c), built upon the late-interaction architecture, demonstrated strong performance on a variety of retrieval tasks. We build the MuKA retriever based on PreFLMR to retrieve multimodal documents for downstream generators.

3 Method

3.1 Overview

For the task of visual information-seeking, given a multimodal query (q, I_q), where q is the textual question and I_q is the query image, a model is expected to generate a textual answer a. A knowledge base can be utilized during generation, comprising of candidate multimodal documents (d, I_d), where d represents the document text and I_d represents the document image. Previous works leverage the multimodal query to retrieve textual documents but the exact textual documents can be hard to find due to difficulties in recognizing long-tail entities within the query images. As the example in Figure 1 shows, it is difficult to judge which one is the most relevant entity to the query image based on the texts of two entities. However, we can recognize which building is more similar to the query image by observing their corresponding images. Therefore, we construct multimodal knowledge bases (See Sec. 3.2) and propose leveraging multimodal documents in the knowledge bases in retrieving relevant documents and generating an answer to the given multimodal query.

We then propose a new framework called MuKA, which adopts an RAG framework based on a multimodal knowledge base to solve the problem. In the stage of retrieval, we add document images as a source to match with query text and images and propose a masked fine-grained multimodal interaction mechanism. (See Sec. 3.3.) In the stage of generation, we propose fine-tuning a foundation multimodal large language model with multiple interleaved retrieved documents, consisting of image and text, as our generator model. (See Sec. 3.4.) During inference, given a multimodal query, we first retrieve top-K relevant multimodal documents from the knowledge base by using our proposed MuKA retriever. Then we compose a prompt, including the image and text of question, a list of interleaved image and text of top documents, and the instruction, to generate a short answer. By following the instruction, our generator finally generates answers.

3.2 Multimodal Knowledge Base Construction

To make fair comparisons with previous works, we choose two widely used benchmarks: InfoSeek (Chen et al., 2023) and Encyclopedic

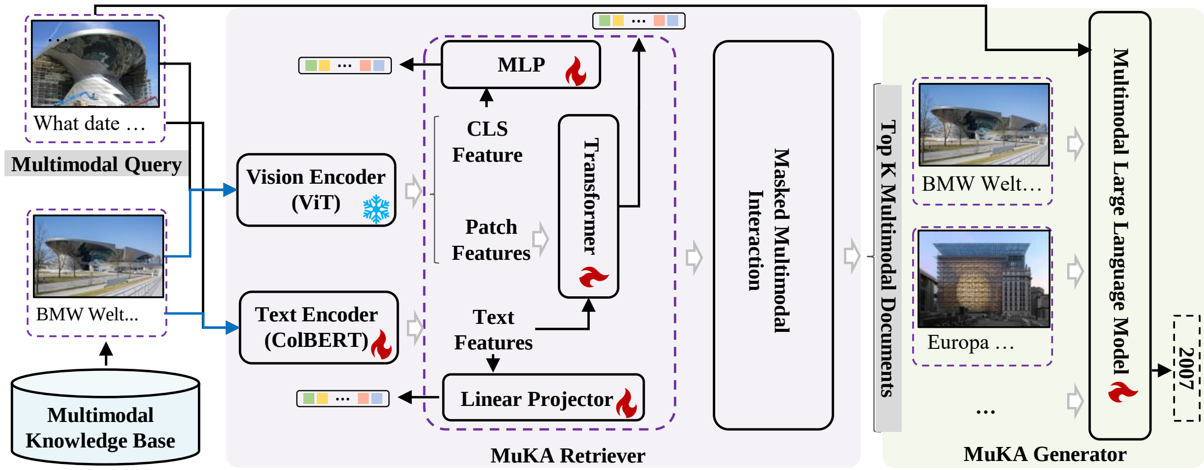

Figure 2: An overview of our MuKA framework, which consists of a multimodal knowledge base (Sec. 3.2), a MuKA retriever (Sec. 3.3) and a MuKA generator (Sec. 3.4).

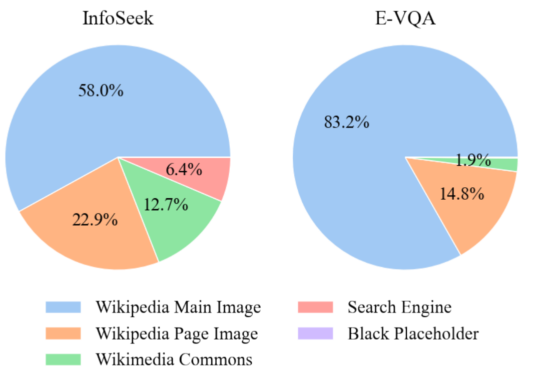

Figure 3: Distribution of the sources to collect images for multimodal knowledge bases.

VQA (E-VQA) (Mensink et al., 2023), which has only text knowledge bases. Therefore, we collect a corresponding image for each entity and upgrade the two datasets with a new constructed multimodal knowledge bases.

Specifically, we adopt a cascaded strategy to collect the entity images from multiple sources. We first request the main image using the Wikipedia API. If the main image is unavailable, we then select the first non-trivial image on the Wikipedia page. If previous methods fail, we search on Wikimedia Commons and select the top-ranked image. As a fallback, we use a common search engine to find the top-ranked image. Finally, for the very few entities where all methods fail, we use a black image as a placeholder.

Finally, we successfully upgrade the knowledge bases for the InfoSeek and the E-VQA datasets into multimodal knowledge bases. The distribution of different sources to collect the images are shown in Figure 3. It shows that most images are obtained via the first step. We will release the data to help reproduce our work and facilitate future research.

3.3 Fine-grained Multimodal Interactions for Retrieval

Given a multimodal query (q, I_q) and candidate multimodal documents of (d, I_d), the retriever aims to identify documents that correspond to the correct entity and contain relevant information. We propose a retriever that performs fine-grained multimodal interactions, derived from the late-interaction mechanism (Khattab and Zaharia, 2020; Lin et al., 2024c).

We start by representing the multimodal query (q, I_q) and a multimodal document (d, I_d) into multi-vector representations, denoted as $\mathbf{Q}$ and $\mathbf{D}$, respectively. Next, we explain how to calculate the relevance between $\mathbf{Q}$ and $\mathbf{D}$ with our masked multimodal interaction mechanism.

Multimodal Query Representation Given the user provided question q along with query image I_q, as shown in Figure 2, the query representation $\mathbf{Q}$ is composed of three categories of features: text features $\mathbf{Q}_\mathbf{T}$, global image features $\mathbf{Q}_\mathbf{I}^{\mathbf{CLS}}$ and image patch features $\mathbf{Q}_\mathbf{I}^{\mathbf{Patch}}$. These categories contain N_q, N_{CLS} and N_{Patch} features, respectively. Each feature vector has d_h dimensions. Therefore, the total number features on the query side l_Q is $N_q + N_{\text{CLS}} + N_{\text{Patch}}$:

$$\mathbf{Q} = \left[\mathbf{Q}_\mathbf{T} \mid \mathbf{Q}_\mathbf{I}^{\mathbf{CLS}} \mid \mathbf{Q}_\mathbf{I}^{\mathbf{Patch}}\right] \in \mathbb{R}^{l_Q \times d_h}. \quad (1)$$

Each category of features is extracted from their respective modality encoders and then projected

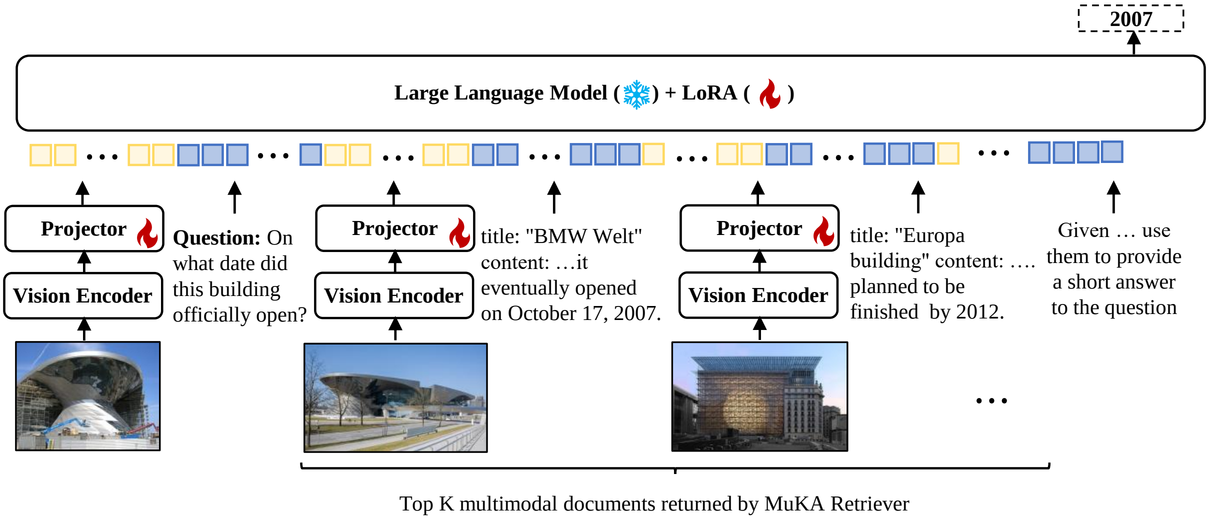

Figure 4: Architecture of our proposed MuKA Generator. Based on an MLLM pre-trained on interleaved image and text, our generator takes multimodal query and top-K multimodal documents returned by our MuKA retriever as the main part of prompt and asks the model to generate short answers.

into the shared dimension. For the text features, a language model F_L first encodes a search instruction and the question q into N_q token features, each with a dimension of d_L. Subsequently, a linear projector F_LIN maps each token feature into the shared dimension d_h.

$$\mathbf{Q_T} = F_\text{LIN}(F_L(q)) \in \mathbb{R}^{N_q \times d_h}. \quad (2)$$

For the query image I, a vision encoder F_V encodes I into both a global image feature and N_Patch image patch features, all with a dimension of d_V. Specifically for Vision Transformers (Dosovitskiy et al., 2020), the global image feature is extracted from the final layer, while the patch features are obtained from the second-to-last layer for better representations. Furthermore, the global image feature is processed through a multi-layer perception module F_MLP, projecting it into N_CLS features of shared dimension d_h:

$$\mathbf{Q_I^{CLS}} = F_\text{MLP}(F_\text{V,CLS}(I)) \in \mathbb{R}^{N_\text{CLS} \times d_h}. \quad (3)$$

For each patch feature, a transformer module F_TR incorporates text features into the cross attention mechanism to perform query-aware feature mapping, transforming it into the shared dimension d_h:

$$\mathbf{Q_I^{Patch}} = F_\text{TR}(F_{\text{V},-2}(I), F_L(q)) \in \mathbb{R}^{N_\text{Patch} \times d_h}. \quad (4)$$

Multimodal Document Representation Given a document d along with query image I_d, as shown in Figure 2, the document representation $\mathbf{D}$ is composed of two categories of features: text features $\mathbf{D_T}$ and global image features $\mathbf{D_I^{CLS}}$. We do not use document image patch features because it needs query text in the cross attention modules to calculate patch features online, which is resource-intensive. The representation of a multimodal document is achieved by concatenating of the text features $\mathbf{D_T}$ and the global image features $\mathbf{D_I^{CLS}}$. The number of features for each multimodal document l_D is $N_d + N_\text{CLS}$:

$$\mathbf{D} = \left[\mathbf{D_T} \mid \mathbf{D_I^{CLS}}\right] \in \mathbb{R}^{l_D \times d_h}. \quad (5)$$

Specially, a separate text encoder F'_L encodes the document text d into N_d token features, which are then projected into the shared dimension d_h by its corresponding linear projector F'_LIN:

$$\mathbf{D_T} = F'_\text{LIN}(F'_L(d)) \in \mathbb{R}^{N_d \times d_h}. \quad (6)$$

Regarding the document image I_d, we reuse the vision encoder F_V and multi-layer perception module F_MLP on the query side:

$$\mathbf{D_I^{CLS}} = F_\text{MLP}(F_\text{V,CLS}(I_d)) \in \mathbb{R}^{N_\text{CLS} \times d_h}. \quad (7)$$

The features on the document side can be pre-built and indexed to facilitate efficient retrieval.

Masked Multimodal Interaction As the global document image features and query image patch features are in different levels, interaction between them lacks practical meaning and may introduce interference. We mask out such interaction, implementing a masked multimodal late-interaction mechanism:

$$r((q,I),(d,I_d)) = \sum_{i=1}^{l_Q} \max_{j=1}^{l_D} \text{mask}(\mathbf{Q}_i \mathbf{D_j}^T), \quad (8)$$

where the *mask* operator sets the relevance scores between the patch features of query images and the global features of document images into $-\infty$, thereby excluding them from the max-pooling process of the late interaction mechanism. This prevents the unwanted interaction during relevance calculation.

3.4 Top-K Multimodal Documents Augmented Generator

We propose enriching answer generators with top-K multimodal documents returned by our MuKA retriever, which offers MLLMs a clear view of entities that supply relevant information to the question, leading to more accurate answers.

Similar to VILA (Lin et al., 2024a), our MuKA generates textual responses conditioning on visual and textual contexts, as shown in Figure 4. An image I is first encoded using a pre-trained visual encoder F_V to obtain N_{Patch} patch embeddings, each with a dimension of d_V:

$$\mathbf{E_{I,Patch}} = F_V(I) \in \mathbb{R}^{N_{\text{Patch}} \times d_V}. \quad (9)$$

Second, these patch embeddings are then transformed by a vision-language integration module F_M, which converts them into a sequence of visual tokens of length N_m:

$$\begin{aligned} \langle \text{image} \rangle &= \{v_1, v_2, \ldots, v_{N_m}\} \\ &= F_M(\mathbf{E_{I,Patch}}) \in \mathbb{R}^{N_m \times d_L}, \end{aligned} \quad (10)$$

where each visual token v_i corresponds to an embedding compatible with the language model, hence having a dimension of d_L. Third, we denote a sequence of text tokens as $\langle\text{text}\rangle$ accordingly. The MLLM, with parameter θ, predicts output text sequence $\langle\text{text}\rangle_o$ of length L in an auto-regressive manner:

$$\begin{aligned} &p(\langle\text{text}\rangle_o \mid \langle\text{image}\rangle_q \langle\text{text}\rangle_q \langle\text{image}\rangle_d^1 \langle\text{text}\rangle_d^1 \ldots) \\ &= \prod_{i=1}^{L} p_\theta(\langle\text{text}\rangle_{o,i} \mid \langle\text{image}\rangle_q \langle\text{text}\rangle_q \ldots \langle\text{text}\rangle_{o,<i}), \end{aligned} \quad (11)$$

where $\langle\text{image}\rangle_q \langle\text{text}\rangle_q$ denotes tokens of the query, and $\langle\text{image}\rangle_d^i \langle\text{text}\rangle_d^i$ denotes tokens of the i-th multimodal document.

Dataset	Samples #Train	Samples #Valid	Samples #Test	Knowledge Base #Passages	Knowledge Base #Entities
InfoSeek	100k	-	4,708	98k	34k
E-VQA	167k	9,852	3,750	52k	19k

Table 1: Statistics of the Infoseek and E-VQA datasets used in our experiments. The counts for passages and entities represent unique values across all dataset splits.

Model	InfoSeek	E-VQA
CLIP	17.1	10.4
FLMR	47.1	-
Google Lens	-	62.5
PreFLMR	60.1	73.7
MuKA Retriever (ours)	**69.4**	**77.7**
w/o fine-tuning	66.6	75.6
w/o mask	68.9	77.0

Table 2: Retrieval performance of Recall@5 on InfoSeek and E-VQA datasets. Baseline results for CLIP, FLMR, and Google Lens are sourced from existing literature. PreFLMR and our MuKA Retriever are fine-tuned on respective knowledge bases. *w/o fune-tuning* is the zero-shot version of our MuKA Retriever. *w/o mask* indicates no masking between global image features and patch features.

4 Experiments

4.1 Datasets and Evaluations

Visual information-seeking datasets. We use a sub-split of InfoSeek (Chen et al., 2023) and E-VQA (Mensink et al., 2023) dataset to evaluate the visual information-seeking performance, following the same setup as (Lin et al., 2024c).

Knowledge base. To ensure a fair comparison, we use the knowledge bases introduced in previous literature (Lin et al., 2024c), which consist of textual documents sourced from Wikipedia. Each document belongs to an entity while each entity may corresponds to multiple documents. For each QA pair, the documents that belongs to the correct entity and contain the answer are considered positive items. Our constructed multimodal knowledge bases, as detailed in Sec. 3.2, builds upon the textual knowledge bases, with each textual document paired with the image of its corresponding entity. Table 1 presents the statistics of the two datasets along with their knowledge bases provided.

Evaluation protocol. We report Recall@5 performance for knowledge retrieval. This metric measures whether the correct answer to a ques-

No.	Model	Finetune	RAG		Retriever	Infoseek			E-VQA
			Text	Image		Unseen-Q	Unseen-E	Overall	Overall
Previous SOTA method is RA-VQAv2 w/ PreFLMR (Lin et al., 2024c)									
SoTA Result								30.65	54.45
Baselines: Zero-shot									
(1)	LLaVA-13B	✗	✗	✗	-	11.2	9.0	10.0	17.8
(2)	VILA-13B	✗	✗	✗	-	14.2	11.3	12.6	19.3
Baselines: Fine-tune Without Knowledge Augmentation									
(3)	LLaVA-13B	✓	✗	✗	-	27.5	19.5	22.8	32.7
(4)	VILA-13B	✓	✗	✗	-	28.8	20.9	24.3	32.1
Comparison: Impact of Knowledge Augmentation Modalities									
(5)	LLaVA-13B	✓	✓	✗	baseline	32.5	30.2	31.3	56.3
(6)	VILA-13B	✓	✓	✗	baseline	37.0	30.9	33.7	57.2
(7)	VILA-13B	✓	✓	✓	baseline	42.2	33.0	37.1	59.5
Comparison: Impact of Retrievers									
(8)	VILA-13B	✓	✓	✗	MuKA Retriever	42.1	37.7	39.8	60.2
(9)	VILA-13B	✓	✓	✓	MuKA Retriever	**44.6**	**40.6**	**42.5**	**63.1**
(10)	VILA-8B	✓	✓	✓	MuKA Retriever	39.8	37.3	38.5	60.6

Table 3: Results of LLaVA-1.5 and VILA models on visual information-seeking tasks across different settings. The baseline retriever is the PreFLMR model with ViT-G in a zero-shot manner. All generators are trained using the baseline retriever results to ensure a fair comparison. The best results are highlighted in bold.

tion can be found within the top-5 retrieved documents. To evaluate the generated answers, we use the evaluation provided by InfoSeek and E-VQA. For InfoSeek, each predicted answer is normalized and evaluated based on the question type. An exact match is required for questions expecting an answer in string while a flexible range is allowed for those expecting a time or a number. The InfoSeek results include three scores: one for the subset of unseen questions, another for unseen entities, and an overall score. As for E-VQA, each predicted answer is assessed using BERT Matching (Bulian et al., 2022) against reference answers to determine correctness. We report the average accuracy for E-VQA.

4.2 Implementation Details

We implement our MuKA retriever based on the state-of-the-art PreFLMR (Lin et al., 2024c) that with a ViT-G vision encoder (Cherti et al., 2022). We report the results of PreFLMR and our MuKA retriever after fine-tuning for one epoch on the training split, utilizing textual and multi-modal knowledge bases respectively. During fine-tuning, the vision encoder remains frozen. The learning rate is set to 1e-4 for the mapping networks and 1e-5 for other trainable modules. The parameters are optimized using the Adam optimizer with an in-batch contrastive loss. The training is conducted on 4 GPUs, with a batch size of 8 and gradient accumulation steps set to 8.

For the answer generators, we report results based on two families of MLLMs: LLaVA-1.5 (Liu et al., 2024a) and VILA (Lin et al., 2024a). LLaVA-1.5, designed for using a single image as context, supports textual RAG. VILA, in contrast, trained to understand multiple images, can perform RAG with both textual and multimodal documents.

To ensure a fair comparison across answer generators, we use the same retrieval results to construct training examples, obtained from a zero-shot inference using the PreFLMR model aforementioned. For both training and testing, we provide the top-5 retrieved documents. We truncate each E-VQA document to the first 100 words. We apply Low-Rank Adaptation (LoRA) (Hu et al., 2021) to reduce trainable parameters, setting the LoRA rank to 128 and the LoRA alpha to 256. The total batch size is 512, following Wiki-LLaVA (Caffagni et al., 2024). We use the Adam optimizer for fine-tuning, making only the parameters of the multimodal projectors and the LoRA modules trainable, with learning rates set to 2e-5 and 2e-4, respectively.

4.3 Results

Evaluation on Knowledge Retrieval. We compare our proposed retriever with previous baselines and present results in Table 2. The results show that our retriever performs the best in terms of R@5 upon both datasets. It significantly outperforms the best baseline PreFLMR by 9.3 points on the InfoSeek dataset and by 4.0 points on the E-VQA dataset. By an ablation study, as shown in Table 2, we have some findings on what works: 1) Our proposed introducing document images in calculating relevance exhibits immediate performance gain over the baseline PreFLMR, i.e., from 60.1 to 66.6 on InfoSeek and from 73.7 to 75.6 on E-VQA, even without fine-tuning. 2) The performance has a drop of 4% on InfoSeek and 2.7% on E-VQA if without fune-tuning, indicating fine-tuning can continue to improve the performance. 3) Ablating masking matching between patches features of query images and global features of document images brings consistent but slight drops, verifying our idea. These findings strongly support our claims on the well-calculated similarity between query image and document image providing indispensable evidence for judging they are the same entity.

Evaluation on Answer Generation. We conduct extensive experiments to evaluate our proposed MuKA framework in terms of generation results (See 3). Our findings reveal that the MuKA - 13B model, when combined with the MuKA retriever, achieves significant improvements over the state-of-the-art results, with a 38.7% boost on InfoSeek and 15.9% increases on E-VQA. We attribute these gains to several factors. First, fine-tuning on QA pairs enhances performance greatly, as evidenced by comparing method (3) to (1) and (4) to (2). Model augmented with additional knowledge clearly outperform those without, indicating that visual information-seeking questions are highly knowledge-intensive. VILA generally performs better than LLaVA in the same setting. Second, the model augmented with multimodal knowledge, method (7), improves accuracy by about 2 - 3 points over its counterpart with textual knowledge input, method (6). This suggests that visual information in documents aids answer generation. Third, improved retrieval results from our MuKA retriever benefits both forms of augmentation, underscoring the importance of high-quality retrieval in the visual information-seeking task. The combination of the MuKA generator with the MuKA retriever, method (9) clearly outperforms the combination with the baseline retriever, method (7).

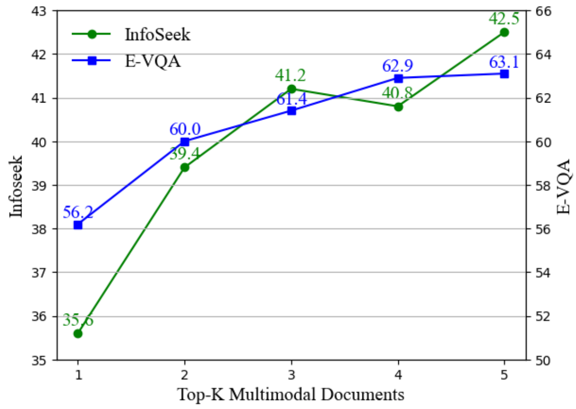

Figure 5: Accuracy of generated answers with different top-K documents on the InfoSeek and E-VQA datasets. The multimodal documents are retrieved by the MuKA retriever and the answers are generated by answer generators based on VILA-13B.

Necessity of Multiple Documents. We analyze the impact of feeding top-K multimodal documents to the final score of generated answers in our MuKA method. As shown in Figure 5, the performance rises up with K increased only except for one dot on two datasets. The best performance is achieved when top five documents are fed into the answer generator. The absolute improvement over the result upon top one document is about 7 points on both datasets. This indicates augmenting multiple documents is necessary and effective because the relevant may not be ranked at the top. Due to the limitation of context length of VILA, we cannot input more documents including images and passages. However, according to the trend on E-VQA where the curve goes flat at K equals to five, we can expect that increasing K cannot bring extra positive gain at some point because more documents may introduce more irrelevant documents to confuse the generator model.

5 Conclusion

In this paper, we tackle the challenging visual information-seeking task by leveraging multiple multimodal documents. We propose MuKA retriever to enable multimodal retrieval from multimodal knowledge bases, and MuKA generator to guide multimodal language models to utilize the multimodal documents for answer generation. We conduct extensive experiments to demonstrate the effectiveness of our approaches, highlighting the significance of multimodal knowledge and multiple documents for this knowledge-intensive task.

Limitations

While we conducted extensive experiments to validate the significance of multimodal documents in both knowledge retrieval and answer generation, it is important to acknowledge several limitations. Firstly, we consider a single image for each entity. In reality, an entity may have various views under different conditions and perspectives, which our current approach does not account for. Secondly, the sizes of the knowledge bases used in our experiments are moderate. The efficiency and effectiveness still need to be studied on larger knowledge bases in real-world scenarios. Lastly, there is a lack of clarity on how exactly the answer generators provide answers from the retrieved documents. Techniques including chain-of-thought prompting (Wei et al., 2022) could be explored to improve the transparency of the answer generation process. In light of these considerations, our research may be limited, and addressing these limitations could provide valuable insights for future work.

Acknowledgments

This work is supported by the National Key R&D Program of China (2023YFF0905402), Beijing Natural Science Foundation (L233008), National Natural Science Foundation of China (No. 62276268) and Migu Culture Technology Co., Ltd. We acknowledge the anonymous reviewers for their helpful comments.

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.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, Singapore. Association for Computational Linguistics.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2022. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ziniu Hu, Ahmet Iscen, Chen Sun, Kai-Wei Chang, Yizhou Sun, David Ross, Cordelia Schmid, and Alireza Fathi. 2024. Avis: Autonomous visual information seeking with large language model agent. *Advances in Neural Information Processing Systems*, 36.

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.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

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

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. *arXiv preprint arXiv:2004.04906*.

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

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. 2022. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3108–3120.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024a. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26689–26699.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2024b. 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. 2024c. Preflmr: Scaling up fine-grained late-interaction multi-modal retrievers. *arXiv preprint arXiv:2402.08327*.

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

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

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Sail: Search-augmented instruction learning. *arXiv preprint arXiv:2305.15225*.

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.

Jielin Qiu, Andrea Madotto, Zhaojiang Lin, Paul A Crook, Yifan Ethan Xu, Xin Luna Dong, Christos Faloutsos, Lei Li, Babak Damavandi, and Seungwhan Moon. 2024. Snapntell: Enhancing entity-centric visual question answering with retrieval augmented multimodal llm. *arXiv preprint arXiv:2403.04735*.

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.

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

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

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*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

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*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

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

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. 2024. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023b. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.

A Overlap Analysis of Document Images

To ensure that the collected entity images do not overlap with the query images in the test sets of InfoSeek and E-VQA, we utilize the `imagehash` toolkit to determine whether the query image in a test example is identical to its entity image. Specifically, we use the perceptual hash with a hash size of 8, considering images with a distance of 10 or less to be identical. Our analysis reveals that 13 out of 4,708 (0.3%) testing examples in InfoSeek and 47 out of 3,750 (1.25%) testing examples in E-VQA are considered overlapped. Given these low occurrence rates, we do not implement additional processing steps.

B Prompt for Answer Generators

Here, we present the specific prompts used for answer generation across various settings for reference purposes.

Zero-shot & Fine-tune w/o Knowledge Augmentation This prompt is designed to test answer generators without knowledge augmentation. Consequently, the model relies solely on the knowledge stored in its parameters. Fine-tuning without knowledge augmentation ensures that the model provides answers adhering to the format of a specific dataset.

Prompt for Zero-shot & Fine-tuning w/o Knowledge Augmentation
<query image>
Question: {question}
Give a short answer.

Fine-tune w/ Textual Knowledge Augmented This prompt is intended to supply the answer generator with retrieved textual passages. This procedure is similar to Retrieval-Augmented Generation (RAG) in language models, except it includes the input of a query image. We also provide an additional instruction to guide the model in leveraging the matched passages effectively.

Fine-tune w/ Multimodal Knowledge Augmentation This prompt is designed for Multimodal Large Language Models (MLLMs) that accepts contexts with multiple images. We implement our MuKA generators using this prompt. By providing multimodal documents (i.e. documents images and their texts), the MLLMs gains a clear view of the entities while obtains relevant information for answering the questions, thereby leading to more accurate answers.

Prompt for Fine-tuning w/ Textual Knowledge Augmentation
<query image>
Question: {question}
Retrieved passages:
1: <document text 1 >
2: <document text 2 >
3: <document text 3 >
4: <document text 4 >
5: <document text 5 >

Given the question, along with retrieved passages, identify the matched passages and use them to provide a short answer to the question.

Prompt for Fine-tuning w/ Multimodal Knowledge Augmentation
<query image>
Question: {question}
Retrieved passages:
1: <document image 1 ><document text 1 >
2: <document image 2 ><document text 2 >
3: <document image 3 ><document text 3 >
4: <document image 4 ><document text 4 >
5: <document image 5 ><document text 5 >

Given the query image and question, along with retrieved passages and their images, identify the matched passages and use them to provide a short answer to the question.

C Textual RAG with More Documents

We used top-5 multimodal documents in our MuKA generator due to the limitation of the context length of the models. However, for textual RAG, the context length of MLLMs allows reading more textual documents for answer generation.

To find out whether more textual documents could contribute to answer accuracy, we fine-tuned LLaVA-1.5 13B models with more documents following the settings in the paper and tested them with the same number of documents from the MuKA retriever. The final score is 31.0 for top-1, 35.3 for top-3, 37.8 for top-5, 37.7 for top-10, and 37.2 for top-15 passages. Our findings suggest that using top-5 passages is sufficient, as more documents do not necessarily improve performance.