Multimodal Hypothetical Summary for Retrieval-based Multi-image Question Answering

Peize Li[1, 2*], **Qingyi Si**[3], **Peng Fu**[2, 4†], **Zheng Lin**[2, 4], **Yan Wang**[5, 1†]

[1]School of Artificial Intelligence, Jilin University, Changchun, China
[2]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[3]Huawei Technologies Co., Ltd., Beijing, China
[4]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[5]Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun, China
lipz21@mails.jlu.edu.cn, siqingyi@huawei.com, {fupeng, linzheng}@iie.ac.cn, wy6868@jlu.edu.cn

Abstract

Retrieval-based multi-image question answering (QA) task involves retrieving multiple question-related images and synthesizing these images to generate an answer. Conventional "retrieve-then-answer" pipelines often suffer from cascading errors because the training objective of QA fails to optimize the retrieval stage. To address this issue, we propose a novel method to effectively introduce and reference retrieved information into the QA. Given the image set to be retrieved, we employ a multimodal large language model (visual perspective) and a large language model (textual perspective) to obtain *multimodal hypothetical summary* in question-form and description-form. By combining visual and textual perspectives, MHyS captures image content more specifically and replaces *real* images in retrieval, which eliminates the modality gap by transforming into text-to-text retrieval and helps improve retrieval. To more advantageously introduce retrieval with QA, we employ contrastive learning to align queries (questions) with MHyS. Moreover, we propose a coarse-to-fine strategy for calculating both sentence-level and word-level similarity scores, to further enhance retrieval and filter out irrelevant details. Our approach achieves a 3.7% absolute improvement over state-of-the-art methods on RETVQA and a 14.5% improvement over CLIP. Comprehensive experiments and detailed ablation studies demonstrate the superiority of our method.

Introduction

Visual question answering (VQA) is an engaging multimodal task that involves analyzing both images and natural language. Unlike conventional VQA tasks, retrieval-based multi-image QA (Penamakuri et al. 2023) is a newly proposed and more challenging VQA task. It requires retrieving and integrating multiple question-related images to generate an answer. Existing vision-language models often rely on extensive pre-training (Li et al. 2022a; Cho et al. 2021; Si et al. 2023) or sophisticated attention mechanisms (Kim, Jun, and Zhang 2018; Anderson et al. 2018; Zhu et al. 2020) to tackle multimodal tasks. However, they face difficulties with retrieval-based multi-image QA due to the limited exposure to multi-image question-answering pairs during pre-training and the added complexity of the retrieval process.

To advance research in this area, recent work has proposed several solutions. For example, REALM (Guu et al. 2020), RAG (Lewis et al. 2020) and RETRO (Borgeaud et al. 2022) utilize extensive world knowledge from language models to improve QA performance, though they are limited to textual data. To broaden the scope of knowledge, MuRAG (Chen et al. 2022) extends these approaches to mul-

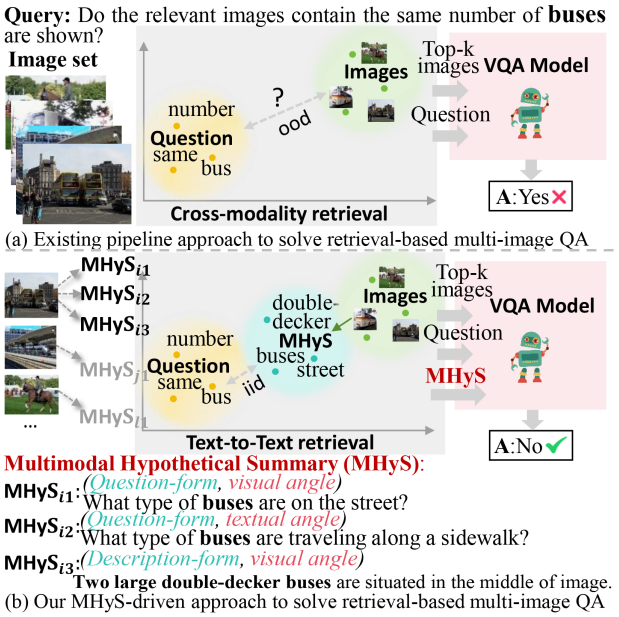

Figure 1: An illustration of our motivation. Compared to the "retrieve-then-answer" pipeline, our approach leverages multimodal hypothetical summary (MHyS) to transform cross-modality retrieval into text-to-text retrieval, effectively introducing and referencing retrieval into QA.

*Work done during an internship at Institute of Information Engineering, Chinese Academy of Sciences.
†Corresponding Authors.

timodal corpora. Besides, Solar (Yu et al. 2023) transforms images and tables into a unified text format, utilizing the advantages of language models more effectively. Despite enhancing knowledge and content generation, these methods are still part of the "retrieve-then-answer" pipeline, and face the challenge of misaligned training objectives between retrieval and QA. To address this issue, our method employs multimodal hypothetical summary to replace images during retrieval, optimizing the network end-to-end using contrast enhancement loss and VQA loss. We illustrate this idea with the example shown in Figure 1.

This example reflects the two purposes of our designed MHyS, encompassing both question-form and description-form. First, for the query, "*Do the relevant images contain the same number **buses** are shown?*", MHyS captures important details effectively, such as "***buses***". Besides, the question-form MHyS aligns with query in terms of semantic overlap (same words) and structural similarity (question format). Second, the description-form MHyS provides description-relevant visual information, such as "***Two large double-decker buses***", which not only aids in retrieval but also helps the model identify question-attended evidence.

The "retrieve-then-answer" pipelines (Penamakuri et al. 2023; Chen et al. 2022; Yu et al. 2023) first utilize cross-modality retrieval method (e.g. BLIP (Li et al. 2022b), ALBEF (Li et al. 2021) and mPLUG (Li et al. 2022a)) to rank images based on question. Then, these question-related images are combined with question into VQA model (e.g. BAN (Kim, Jun, and Zhang 2018), LXMERT (Tan and Bansal 2019), VL-BART (Cho et al. 2021)). Instead of decomposing retrieval-based multi-image QA task, our MHyS-driven approach uses multimodal hypothetical summary to connect retrieval and QA. Concretely, we replace real images with MHyS to calculate the similarity score with query, effectively transforming text-to-text retrieval. To retain more information and contact retrieval with QA, we combine the selected real images (based on similarity scores) with their MHyS to generate answers.

We expect the multimodal hypothetical summary (MHyS) to proficiently capture key object words (overlapping with query words) and provide more specific information about real images, enhancing both retrieval and QA. To achieve this, we design question-form and description-form MHyS. To effectively utilize vision-language alignment knowledge and filter out irrelevant details, we employ the frozen CLIP (Radford et al. 2021) (pre-trained on 400M image-text pairs using contrastive learning) to calculate sentence-level similarity and a multimodal encoder (Cho et al. 2021) (rich in multimodal knowledge) to compute word-level similarity. We adopt a coarse-to-fine strategy for accurate similarity calculation. To seamlessly introduce and effectively utilize retrieval and QA, we use contrastive enhancement loss to align query with MHyS. This approach also enables the model to capture more generalized representation.

The main contributions are summarized as follows: (1) We propose an innovative MHyS-driven paradigm that effectively transforms into text-to-text retrieval and seamlessly connects retrieval with QA. (2) We design the question-form MHyS to closely align with query through semantic overlap and similar structure, while the description-form MHyS not only facilitates retrieval but also helps in capturing question-attended evidence. (3) The proposed multi-granularity retrieval aims to precisely filter out irrelevant details in coarse-to-fine strategy. (4) Our approach achieves a 3.7% absolute accuracy improvement over the state-of-the-art on the RETVQA dataset and a 14.5% enhancement over CLIP, demonstrating its effectiveness.

Related Work

Visual Question Answer and Retrieval-based Multi-image QA. In recent years, various interesting studies have been proposed for visual question answering (VQA). Conventional VQA tasks, such as knowledge-based VQA (Marino et al. 2019; Schwenk et al. 2022), visual reasoning VQA (Hudson and Manning 2019; Zellers et al. 2019) and language bias VQA (Agrawal et al. 2018; Si et al. 2022b), have been thoroughly studied. Retrieval-based QA tasks currently include the WebQA (Chang et al. 2022) and RETVQA (Penamakuri et al. 2023). WebQA relies on retrieving textual knowledge and has been addressed by various LLM-based methods (Chen et al. 2022; Yu et al. 2023; Gu et al. 2024). However, RETVQA (retrieval-based multi-image question answering) is newly proposed and more challenging in comparison. It requires retrieving multiple question-related images from a collection of relevant as well as irrelevant images. The challenge lies in accurately retrieving question-relevant images and providing the correct answer. Existing solution MI-BART (Penamakuri et al. 2023) for this task uses a two-stage "retrieve-then-read" pipeline. Unlike separating the task into two sub-tasks, our method uses MHyS to introduce multi-image retrieval into QA, which gets rid of the cascading error in two-stage pipelines.

Image-to-Text Retrieval. Previous methods typically improve retrieval performance in two ways: one is to pretrain on larger datasets and the other is to optimize the retrieval components. For example, widely used multimodal models like BLIP (Li et al. 2022b) and mPLUG (Li et al. 2022a) are trained on 14M caption-based datasets including MSCOCO (Lin et al. 2014) and Flickr30K (Young et al. 2014), using Image-Text Contrastive (ITC) and Image-Text Matching (ITM) losses. These models, which operate at the word level, excel in aligning entity semantics. Benefit from contrastive learning, CLIP (Radford et al. 2021) contains vast vision-language alignment knowledge (pre-trained on 400 million image-text pairs) to select important visual information. Therefore, CLIP excels at sentence-level encoding. Our approach combines the strengths of both methods by integrating sentence-level and word-level similarity for more accurate image-text matching.

Typical retrieval methods mainly focus on query expansion or re-ranking scheme. For example, DQU-CIR (Wen et al. 2024) creates a unified multimodal query by merging visual and textual queries, while LeaPRR (Qu et al. 2023) follows the latter way and uses graph reasoning to explore higher-order intra- and inter-modal relationships. These approaches require time-consuming pre- and post-processing. Without requiring any extra annotations, our approach uti-

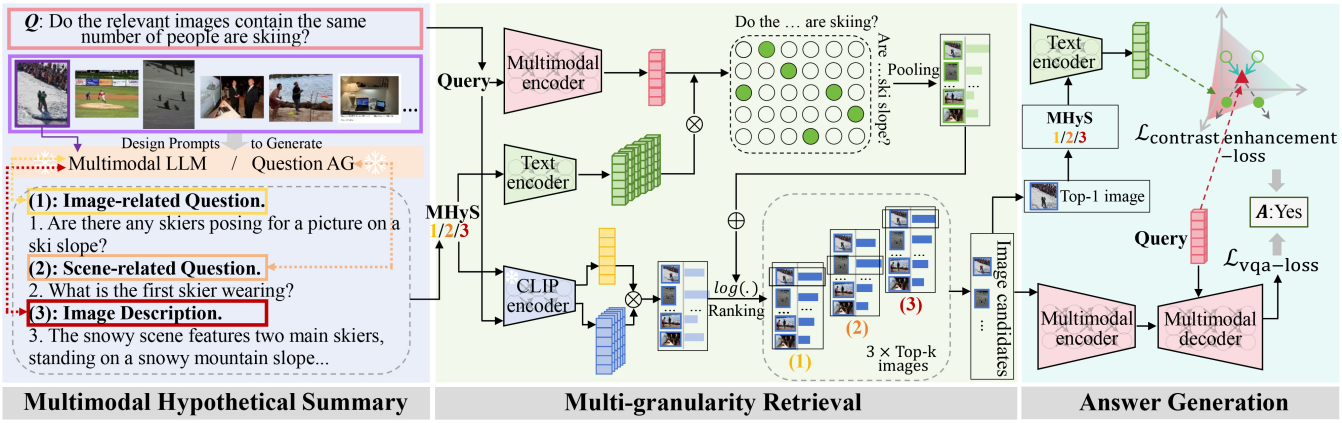

Figure 2: The overview of our approach. **Multimodal Hypothetical Summary** (MHyS) employs multimodal large language model (visual perspective) and language large model (textual perspective) to obtain both question-form and description-form *hypothetical* summary, which replaces *real* images during retrieval and eliminates the modality gap by transforming into text-to-text retrieval. **Multi-granularity Retrieval** calculates sentence-level and word-level similarities to rank images. To capture more information, the selected *real* images (based on similarity scores) are combined with their MHyS to generate the answers.

lizes MHyS (question-form and description-form) to transform into text-to-text retrieval, effectively introducing and referencing retrieved information into QA.

Recently, advanced RAG (Eibich, Nagpal, and Fred-Ojala 2024) method uses large language models (LLMs) to rewrite queries. The multimodal RAG RA-CM3 (Yasunaga et al. 2023) can retrieve and generate both text and images. In addition to deriving the corpus from external documents, the corpus can also be artificially constructed. Hypothetical Document Embedding (Gao et al. 2023) enhances document retrieval by using LLM to generate a hypothetical answer to a query. However, it may produce hallucinated queries that introduce noise and is limited to text retrieval. On the contrary, our method uses multimodal hypothetical summary (MHyS) to replace images being retrieved rather than queries, helping align queries with target images and capture question-attended valuable evidence for QA.

Methodology

Given the retrieval-based multi-image dataset $\mathcal{D} = \{(\boldsymbol{Q}_i, \mathcal{I}_i, \boldsymbol{A}_i)\}_{i=1}^{N}$ with N samples, where $\boldsymbol{Q}_i$, $\mathcal{I}_i$ and $\boldsymbol{A}_i$ denote the question, image set and ground-truth answer of i-th sample respectively. The image set $\mathcal{I}_i = \{\boldsymbol{I}_{i1}, \boldsymbol{I}_{i2}, ..., \boldsymbol{I}_{ij}\}_{j=1}^{N_e}$ includes both relevant and irrelevant images to the question, the task is to select multiple question-related images from the image set $\mathcal{I}_i$ and use these selected images, along with the question, to generate the answer $\boldsymbol{A}_i$. Figure 2 illustrates an overview of our method, encompassing multimodal hypothetical summary, multi-granularity retrieval and answer generation.

Multimodal Hypothetical Summary

Multimodal hypothetical summary (MHyS) aims to obtain question-form and description-form multimodal hypothetical summary, which aids in aligning with query and provides question-attended valuable evidence for QA. By replacing real images in retrieval, MHyS mitigates inter-modal out-of-domain issue by transforming into text-to-text retrieval.

Concretely, given the i-th sample ($\boldsymbol{Q}_i$, $\mathcal{I}_i$, $\boldsymbol{A}_i$), we obtain MHyS for each image $\boldsymbol{I}_{ij} \in \mathcal{I}_i$ using the following approach: (1) Image-related Question: We utilize the *Multimodal Large Language Model* mPLUG-Owl2 (Ye et al. 2024) with the prompt "Generate a question based on the image:" to create MHyS $\boldsymbol{Q}_{ij}^{m}$ in the form of question. This approach directly generates MHyS based on images (visual perspective), which focuses more on the visual content. mPLUG-Owl2 (Ye et al. 2024) is a multimodal LLM with strong multimodal generation capabilities and proficiency in visual reasoning. The multimodal LLM, which is further trained to follow instructions, can zero-shot generalize to diverse prompts. (2) Scene-related Question: To make MHyS diverse, we propose question-form MHyS from textual perspective. First, we use mPLUG-owl2 (Ye et al. 2024) to generate scene information with the prompt "Generate the scene information based on the image:". The scene information is then fed into *Question and Answer Generation model* QAG (Ushio, Alva-Manchego, and Camacho-Collados 2023) model to generate scene-related questions $\boldsymbol{Q}_{ij}^{d}$. This MHyS enriches the visual perspective with textual perspective. (3) Image Description: Question-form MHyS may not cover specific image information. To address this, we use mPLUG-Owl2 to generate detailed image descriptions $\boldsymbol{D}_{ij}$ based on images with the prompt "Generate a detailed description based on the image:". This process (1)-(3) generates multifaceted MHyS that integrate both visual and textual perspectives, enhancing both retrieval and QA. Following TwO (Si et al. 2023), we use 600-dimensional word embeddings (Pennington, Socher, and Manning 2014) in conjunction with GRU to obtain the base embedding representations ($\boldsymbol{Q}_{ij}^{m}$, $\boldsymbol{Q}_{ij}^{d}$, $\boldsymbol{D}_{ij}$).

The question-form MHyS matches the query in structure (both being in question form) and semantics (sharing the

same keywords). Meanwhile, the description-form MHyS includes not only keywords but also more specific visual details that support QA task. Together, these components greatly enhance the overall performance of the method.

Multi-granularity Retrieval

To more precisely filter out irrelevant details and improve retrieval accuracy, we employ a coarse-to-fine strategy to compute similarity scores at both the sentence and word levels.

To achieve global alignment, we leverage CLIP (Radford et al. 2021) for sentence-level encoding and similarity calculation, which has been pre-trained on 400 million image-text pairs using contrastive learning and possesses a strong implicit cross-modal alignment capability. Specifically, we use CLIP text encoder to process the query (question $\boldsymbol{Q}_i$) and MHyS components (image-related question $\boldsymbol{Q}_{ij}^m$, scene-related question $\boldsymbol{Q}_{ij}^d$ and image description $\boldsymbol{D}_{ij}$), as illustrated below:

$$\boldsymbol{T}_i = \mathrm{CLIP}_{\mathrm{text}}(\boldsymbol{Q}_i) \tag{1}$$

$$\hat{\boldsymbol{Q}}_{ij}^m, \hat{\boldsymbol{Q}}_{ij}^d, \hat{\boldsymbol{D}}_{ij} = \mathrm{CLIP}_{\mathrm{text}}(\boldsymbol{Q}_{ij}^m, \boldsymbol{Q}_{ij}^d, \boldsymbol{D}_{ij}) \tag{2}$$

To facilitate word-level semantic alignment between the query and MHyS while filtering out irrelevant details, we employ a multimodal encoder for word-level encoding and similarity calculation. Concretely, we adopt the multimodal encoder VL-BART (Cho et al. 2021) to encode query tokens:

$$\overline{\boldsymbol{T}}_i = \mathrm{Enc}_{\mathrm{multi}}(\boldsymbol{Q}_i) \tag{3}$$

To encode MHyS at the word level, we use a new textual encoder comprising six transformer layers:

$$\overline{\boldsymbol{Q}}_{ij}^m, \overline{\boldsymbol{Q}}_{ij}^d, \overline{\boldsymbol{D}}_{ij} = \mathrm{Enc}_{\mathrm{text}}(\boldsymbol{Q}_{ij}^m, \boldsymbol{Q}_{ij}^d, \boldsymbol{D}_{ij}) \tag{4}$$

Each transformer layer consists of a self-attention layer followed by a fully connected linear layer with residual connections. We implement this way because the multimodal encoder maintains a cohesive encoding space for questions and images, enhancing the effectiveness of visual question answering.

Next, we compute similarity scores between the query and MHyS (using image-related questions $\boldsymbol{Q}_{ij}^m$ as an example) at both the sentence-level and word-level:

$$s(\boldsymbol{T}_i, \hat{\boldsymbol{Q}}_{ij}^m) = f_q(\boldsymbol{T}_i)^T f_v(\hat{\boldsymbol{Q}}_{ij}^m) \in \mathbb{R}^{1\times 1} \tag{5}$$

$$s(\overline{\boldsymbol{T}}_i, \overline{\boldsymbol{Q}}_{ij}^m) = g_q(\overline{\boldsymbol{T}}_i)^T g_v(\overline{\boldsymbol{Q}}_{ij}^m) \in \mathbb{R}^{N_l\times N_v} \tag{6}$$

where $s(.)$ represents the cosine similarity function. f_q, f_v, g_q and g_v are the multi-layer perceptron (MLPs) used to encode features. N_l denotes the number of query words, and N_v represents the number of textual words in the MHyS.

Guided by the relevance affinity matrix, we identify the most relevant MHyS content for the query. The multimodal and textual encoders capture the implicit correlations among all MHyS worlds, and the selected query-centric MHyS world incorporates its contextual information, providing crucial clues for retrieval. Specifically, we use max-pooling to assess the relevance of each MHyS world to the query as follows:

$$\widetilde{s}(\overline{\boldsymbol{T}}_i, \overline{\boldsymbol{Q}}_{ij}^m) = \max_{N_v}(\max_{N_l}(s(\overline{\boldsymbol{T}}_i, \overline{\boldsymbol{Q}}_{ij}^m))) \in \mathbb{R}^{1\times 1} \tag{7}$$

To enhance retrieval accuracy, we integrate both sentence-level and word-level similarities:

$$h(\boldsymbol{Q}_i, \boldsymbol{Q}_{ij}^m) = \widetilde{s}(\overline{\boldsymbol{T}}_i, \overline{\boldsymbol{Q}}_{ij}^m) + log(s(\boldsymbol{T}_i, \hat{\boldsymbol{Q}}_{ij}^m)) \tag{8}$$

To address discrepancies in data magnitudes, we apply a $log(.)$ transformation to the word-level similarity before combining it with the sentence-level similarity. Besides, we explore various similarity fusion functions, including direct addition and adaptive methods, which are discussed in detail in the ablation experiments.

We rank the candidate images based on the similarity score of MHyS (image-related question $\boldsymbol{Q}_{ij}^m$), as follows:

$$\mathcal{C}_{qm} = topK(argsort(h(\boldsymbol{Q}_i, \boldsymbol{Q}_{ij}^m))) \tag{9}$$

Similar to equations 5-8, we compute similarity scores and identify candidate images for MHyS (e.g., scene-related questions $\boldsymbol{Q}_{ij}^d$) and MHyS (e.g., image descriptions $\boldsymbol{D}_{ij}$):

$$h(\boldsymbol{Q}_i, \boldsymbol{Q}_{ij}^d) = \widetilde{s}(\overline{\boldsymbol{T}}_i, \overline{\boldsymbol{Q}}_{ij}^d) + log(s(\boldsymbol{T}_i, \hat{\boldsymbol{Q}}_{ij}^d)) \tag{10}$$

$$\mathcal{C}_{qd} = topK(argsort(h(\boldsymbol{Q}_i, \boldsymbol{Q}_{ij}^d))) \tag{11}$$

$$h(\boldsymbol{Q}_i, \boldsymbol{D}_{ij}) = \widetilde{s}(\overline{\boldsymbol{T}}_i, \overline{\boldsymbol{D}}_{ij}) + log(s(\boldsymbol{T}_i, \hat{\boldsymbol{D}}_{ij})) \tag{12}$$

$$\mathcal{C}_d = topK(argsort(h(\boldsymbol{Q}_i, \boldsymbol{D}_{ij}))) \tag{13}$$

By combining candidate images from these three types of MHyS, we generate the final set of candidate images $\mathcal{C}_i$ for each multi-image QA pair ($\boldsymbol{Q}_i$, $\mathcal{I}_i$, $\boldsymbol{A}_i$), as shown:

$$\mathcal{C}_i = \mathcal{C}_{qm} \cup \mathcal{C}_{qd} \cup \mathcal{C}_d \tag{14}$$

Answer Generation

To more effectively introduce and reference retrieved information into QA, we employ contrastive learning to align query with MHyS and VQA loss to align question with question-focused visual content.

To calculate contrast enhancement loss, for the i-th sample, we use the query feature $\overline{\boldsymbol{T}}_i$ processed by multimodal encoder and the top-1 retrieved MHyS feature $\overline{\boldsymbol{Q}}_{ij}^m$ processed by text encoder. We define positive samples $(\overline{\boldsymbol{T}}_i, \overline{\boldsymbol{Q}}_i^{m+})$ and negative sample pairs $(\overline{\boldsymbol{T}}_b, \overline{\boldsymbol{Q}}_b^m)_{b=1}^B$ within the same batch. ($b \neq i$). B denotes the number of negative samples in a batch. Following the MMBS (Si et al. 2022a) study, we use the cosine similarity function to compute the contrastive enhancement loss (in short $\mathcal{L}_{CE}$), formulated as follows:

$$\mathcal{L}_{CE} = -\log \frac{e^{cos(\overline{\boldsymbol{T}}_i, \overline{\boldsymbol{Q}}_i^{m+})}}{e^{cos(\overline{\boldsymbol{T}}_i, \overline{\boldsymbol{Q}}_i^{m+})} + \sum_{b=1}^{B} e^{cos(\overline{\boldsymbol{T}}_b, \overline{\boldsymbol{Q}}_b^m)}} \tag{15}$$

To calculate VQA loss, we utilize the multimodal encoder VL-BART to encode the retrieved images $\mathcal{C}_i = \{\boldsymbol{I}_{i1}, \boldsymbol{I}_{i2}, ..., \boldsymbol{I}_{ij}\}$, as illustrated below:

$$\overline{\boldsymbol{I}}_{i1}, \overline{\boldsymbol{I}}_{i2}, ..., \overline{\boldsymbol{I}}_{ij} = \mathrm{Enc}_{\mathrm{multi}}(\boldsymbol{I}_{i1}, \boldsymbol{I}_{i2}, ..., \boldsymbol{I}_{ij}) \tag{16}$$

The encoded images $(\overline{\boldsymbol{I}}_{i1}, \overline{\boldsymbol{I}}_{i2}, ..., \overline{\boldsymbol{I}}_{ij})$ and question $\overline{T}_i$ are fed into VL-BART decoder to generate the final answer according to the prediction probability $P(.)$ over the vocabulary space $|W|$ for each answer token:

$$P(a_i^1), ..., P(a_i^l) = Softmax(\mathrm{Dec}_{\mathrm{multi}}(\overline{T}_i, \overline{\boldsymbol{I}}_{i1}, \overline{\boldsymbol{I}}_{i2}, ..., \overline{\boldsymbol{I}}_{ij})) \tag{17}$$

We minimize the auto-regressive cross-entropy loss for QA:

$$L_{VQA} = \frac{-1}{N \cdot L \cdot |W|} \sum_{i=1}^{N} \sum_{l=1}^{L} \sum_{w=1}^{|W|} A_i^{l,w} log(P(a_i^{l,w})) \quad (18)$$

where l represents the answer length and N denotes the number of samples.

Experiments

Dataset and Experimental Settings

Dataset. We evaluate our approach on the retrieval-based multi-image QA dataset (Penamakuri et al. 2023), which contains 334K samples for training, 41K for validation and 41K for testing. The questions cover various types, including color, shape, counting, object attributes and relations. To further test the generalization of our method, we retrieve the top 1-10 images to evaluate performance.

Experimental settings. We conduct our experiments using an NVIDIA 3090 24GB GPU. We train the network for 20 epochs with the batch size of 100 and an initial learning rate of 1e-4, using AdamW optimizer. The text encoder has a dimension of 768. Accuracy verifies whether the correct answer is among the generated answers, consistent with the original dataset proposed. We adopt the CLIP model with RN50×64 visual encoder backbone. To ensure reliable results, we provide the average performance from three trials.

Comparisons with State-of-the-Arts

Table 1 shows the comparison of our method with state-of-the-art (SoTA) models, categorized into "retrieve then answer" and "introduce retrieval into QA" based on whether the retrieval process is introduced into the QA process. Several observations can be derived: (1) Two-stage "retrieve then answer" approaches generally perform better among the methods compared. However, our approach, which introduces retrieval into the QA process, surpasses the state-of-the-art two-stage method MI-BART by +3.7%. MI-BART uses a cross-modality relevance retriever pre-trained on COCO and fine-tuned on RETVQA, coupled with a transformer-based encoder-decoder framework in the QA phase. By leveraging the Multimodal Hypothetical Summary (MHyS), our approach more effectively introduces and references the retrieved information into the QA process, resulting in improved performance. (2) Compared to "Introduce retrieval into QA" methods, our approach exceeds the SoTA CLIP-BART by +14.5%. By using MHyS to replace visual images in retrieval, we shift to text-to-text retrieval, effectively aligning with queries and capturing question-attended valuable evidence for QA.

Ablation Study

Ablation of Multimodal Hypothetical Summary. We evaluate various aspects and specific forms of MHyS in Table 2 and derive several key findings: (1) Question-form MHyS proves to be the most effective, outperforming Description-form MHyS by +7.58%. It aligns closely with the question-form query and includes key visual object words, such as "bus" in Figure 1. (2) Developed from textual perspective by the QAG model, MHyS-(Scene-Ques) surpasses MHyS-(Image-Ques) by +1.85%. The latter is generated from visual perspective by the MLLM. Combining these two approaches results in a +8.25% performance increase, emphasizing the benefit of obtaining MHyS from both visual and textual perspectives. (3) MHyS-Description performs slightly better than MHyS-Caption because it covers more scene details. Combining MHyS-Question and MHyS-Description results in a +9.99% overall improvement in our approach. (4) Moreover, contrastive enhancement loss provides a +0.49% increase in performance by further aligning the query with MHyS.

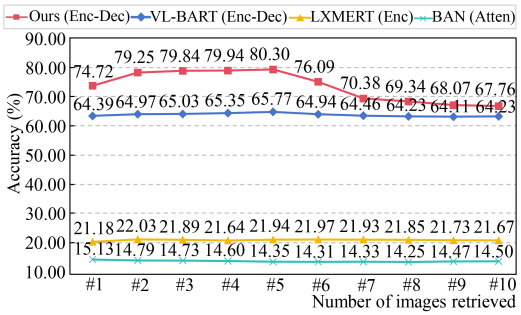

Figure 3: Performance comparison of our method with various baselines under different numbers of retrieved images.

Ablation of Retrieval Similarity. We assess the the effect of various similarity computation strategies in Table 3. An optimal retrieval similarity method should be both simple and effective. We observe the following: (1) *Fine-grained similarity is far more effective than coarse-grained similarity*. For example, "CLIP (sentence-level)" uses global similarity based on entire sentence and image, while "MHyS (word-level)" measures similarity at the object and word level. MHyS (word-level) significantly outperforms CLIP (sentence-level) by +16.4%. (2) *Properly combining coarse-grained and fine-grained methods is crucial*. We experiment with direct addition, $\mathbf{log}(.)$ transformation, and training a fusion adaptive function. The best result is achieved by combining addition with $\mathbf{log}(.)$ transformation, which balances magnitude differences, surpassing both sentence-level and word-level methods (+16.61% $\sim$ +0.21%).

Ablation of Top 1-10 Retrieved Images Across Various Baselines. Figure 3 illustrates the performance across different numbers of retrieved images, comparing our approach with various baseline architectures, including attention-based, encoder-only and encoder-decoder visual-language models, all within the "Introduce Retrieval into QA" category. (1) Our method consistently outperforms other baselines across the top 1-10 retrieved images in multi-image QA, demonstrating its broad applicability. Specifically, it exceeds the encoder-decoder baseline VL-BART by +14.53% $\sim$ +3.53% across these images, underscoring the effectiveness of the MHyS-driven paradigm in enhancing both retrieval and QA. (2) Our method's performance peaks at 80.30% with five images, suggesting that both too few

Models	All	Attribute	Color	Count	Relation	Shape
Introduce retrieval into QA						
CLIP (Radford et al. 2021)-BAN (Kim, Jun, and Zhang 2018)	14.4	0.0	17.6	0.3	13.6	37.0
CLIP (Radford et al. 2021)-VisualBERT (Li et al. 2020)	19.2	0.0	26.6	0.4	23.2	44.4
CLIP (Radford et al. 2021)-LXMERT (Tan and Bansal 2019)	21.9	0.0	32.5	0.4	30.8	47.2
CLIP (Radford et al. 2021)-VLBART (Cho et al. 2021)	65.8	78.9	66.4	55.3	15.6	83.9
Retrieve then answer						
mPLUG (Li et al. 2022a)-BAN (Kim, Jun, and Zhang 2018)	15.2	0.0	19.4	0.3	14.8	38.5
mPLUG (Li et al. 2022a)-VisualBERT (Li et al. 2020)	19.1	0.0	28.1	0.3	21.3	44.0
mPLUG (Li et al. 2022a)-LXMERT (Tan and Bansal 2019)	19.7	0.0	32.4	0.3	20.5	42.1
MI-VLBART (Cho et al. 2021) (Question only)	62.4	74.9	58.0	51.6	12.4	86.9
MI-VLP (Zhou et al. 2020)	65.1	76.8	62.0	50.8	36.8	84.0
MI-(LSTM+VGG) (Antol et al. 2015) (Aggregate VQA)	66.6	75.4	60.1	54.6	32.2	91.3
MI-BART (Penamakuri et al. 2023) (Image stitch variant)	72.1	**81.6**	71.8	62.7	52.0	96.2
mPLUG (Li et al. 2022a)-VLBART (Cho et al. 2021)	75.3	76.6	69.6	80.2	33.1	92.3
MI-BART (Penamakuri et al. 2023)	76.5	78.5	72.1	66.0	**69.5**	92.4
Ours *Introduce retrieval into QA*	**80.3**	78.5	**79.5**	**80.5**	40.4	**97.9**

Table 1: Comparison with state-of-the-art approaches on the retrieval-based multi-image QA dataset using the retrieved images. "*retrieve then answer*" pipeline involves two separating stages: first, cross-modality retrieval and then VQA-based answer prediction. In contrast, "*Introduce retrieval into QA*" approach adopts our proposed MHyS-driven paradigm, which replaces images with multimodal hypothetical summary (MHyS) for retrieval, and selects the images (based on similarity scores) along with their MHyS to generate answers. The hyphen "-" before and after denote the retrieval method and VQA model respectively.

Models	All	Att	Col	Cou	Rel	Sha
MHyS-Caption	69.82	78.13	68.60	60.83	25.87	90.47
MHyS-Description	70.87	**79.18**	70.87	61.54	25.11	90.55
MHyS-(Image-Ques)	71.56	79.00	69.45	61.86	26.64	91.51
MHyS-(Scene-Ques)	73.41	77.69	72.52	67.17	32.17	93.15
MHyS-Question	78.45	77.85	77.82	76.01	39.19	96.47
MHyS-(Desc+Ques) w/ RA-loss	79.81	78.41	78.87	78.57	40.25	97.71
MHyS-(Desc+Ques) (**Ours**)	**80.30**	78.51	**79.45**	**80.52**	**40.44**	**97.85**

Table 2: Ablation study of multimodal hypothetical summary.

Models	All	Att	Col	Cou	Rel	Sha
CLIP (*sentence-level*)	63.69	79.15	62.11	52.21	12.76	84.17
log (CLIP)+log(MHyS)	64.23	78.63	64.16	53.70	14.70	81.75
CLIP+log	73.66	79.41	71.42	66.36	33.79	93.32
CLIP+MHyS	73.86	**79.83**	71.99	67.13	34.62	93.15
Adap-Func [CLIP;MHyS]	80.03	78.78	79.22	79.61	39.29	97.55
MHyS (*world-level*)	80.09	78.45	79.45	79.52	39.44	97.85
CLIP+log(MHyS) (**Ours**)	**80.30**	78.51	**79.45**	**80.52**	**40.44**	**97.85**

Table 3: Ablation study of similarity calculation methods.

and too many images are not optimal for answering questions. (3) Simpler models, like attention-based BAN and encoder-only LXMERT, struggle with multi-image QA, indicating the need for more sophisticated models. Our encoder-decoder approach outperforms BAN by +65.17% and LXMERT +58.27%. LXMERT performs better with two images, while BAN is more effective with one, likely due to their design differences. LXMERT excels at modeling multiple object relationships, while BAN focuses on bilinear fusion of a single image and text.

Ablation of Retrieval-based Single-image QA. Table 4 reports the comparison of our method with state-of-the-art approaches using top-1 retrieved image, categorized into "retrieve then answer" and "Introduce retrieval into QA". We can observe two key points: (1) Our method achieves improvements of +18.29% over the "retrieve then answer" SoTA mPLUG-VLBART and +10.33% over the "Introduce retrieval into QA" SoTA CLIP-VLBART, demonstrating its effectiveness and generality. (2) Table 1 indicates that two-stage "retrieve then answer" methods generally surpass "Introduce retrieval into QA" methods in multi-image QA. However, Table 4 reveals the opposite result for using top-1 retrieved image. Concretely, the "Introduce retrieval into QA" SoTA CLIP-VLBART outperforms the "retrieve then answer" SoTA mPLUG-VLBART by +7.96%, suggesting that the "Introduce retrieval into QA" paradigm is more robust with limited visual information and is becoming the dominant trend in tackling this task.

Models	All	Att	Col	Cou	Rel	Sha
Retrieve then answer						
mPLUG (2022a)-BAN (2018)	14.19	0.0	18.45	0.31	6.74	38.13
mPLUG (2022a)-VisualBERT (2020)	17.41	0.00	28.30	0.29	10.10	40.50
mPLUG (2022a)-LXMERT (2019)	17.62	0.00	26.62	0.26	12.04	42.32
mPLUG (2022a)-VLBART (2021)	56.43	72.69	47.57	54.37	6.07	73.92
Introduce retrieval into QA						
CLIP (2021)-BAN (2018)	15.13	0.00	18.3	0.32	14.34	39.31
CLIP (2021)-VisualBERT (2020)	19.00	0.00	27.31	0.39	21.89	43.48
CLIP (2021)-LXMERT (2019)	21.18	0.00	33.62	0.37	25.22	45.03
CLIP (2021)-VLBART (2021)	64.39	78.85	63.48	53.35	14.73	84.06
Ours	**74.72**	**79.01**	**74.23**	**67.96**	**33.57**	**94.56**

Table 4: Comparison with state-of-the-art approaches on the retrieval-based multi-image QA dataset using the top-1 retrieved image.

Ablation of MLLMs' Performance. Table 5 represents the comparison of our method with state-of-the-art multimodal large language models (MLLMs). These MLLMs, trained on extensive datasets, exhibit strong generation capabilities and are utilized in the two-stage "retrieve-then-answer" pipeline with both fine-tuning and zero-shot ways. For retrieval, we use mPLUG (Li et al. 2022a), known for its cross-modal skip-connections and top performance in image-text retrieval. We gain: (1) Although our method is part of the"Introduce retrieval into QA" category, it outperforms the fine-tuned MLLM QWENVL (Bai et al. 2024) by +25.87% in multi-image QA and +14.99% in single-image

Models	*FT* or *ZS*	All	Att	Col	Cou	Rel	Sha
Multiple-images Retrieval then answer							
mPLUG (2022a)-QWENVL (2024)	*ZS*	6.70	0.93	13.69	8.75	2.12	4.03
mPLUG (2022a)-mPLUG-Owl2 (2024)	*ZS*	35.32	54.20	42.81	33.08	7.39	25.75
mPLUG (2022a)-QWENVL (2024)	*FT*	54.43	75.71	48.54	45.97	4.57	71.12
Ours *Introduce retrieval into QA*		**80.30**	**78.51**	**79.45**	**80.52**	**40.44**	**97.85**
Single-image Retrieval then answer							
mPLUG (2022a)-QWENVL (2024)	*ZS*	8.57	11.68	12.92	8.04	4.06	4.00
mPLUG (2022a)-LLaVA-1.5 (2024)	*FT*	34.70	51.24	44.58	32.42	9.02	23.76
mPLUG (2022a)-LLaVA-1.5 (2024)	*ZS*	34.85	51.34	44.45	32.48	8.76	24.48
mPLUG (2022a)-mPLUG-Owl2 (2024)	*ZS*	35.09	53.28	42.78	33.28	7.08	25.45
mPLUG (2022a)-QWENVL (2024)	*FT*	59.73	77.73	54.89	57.43	4.91	73.72
Ours *Introduce retrieval into QA*		**74.72**	**79.01**	**74.23**	**67.96**	**33.57**	**94.56**

Table 5: Comparison with the state-of-the-art MLLMs. "*FT*" represents answers generated after fine-tuning the MLLMs. "*ZS*" denotes answers generated directly in a zero-shot manner by MLLMs.

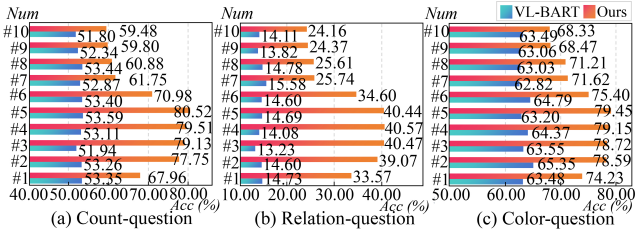

Figure 4: Performance with different question types.

QA, demonstrating the effectiveness of our MHyS-driven paradigm. (2) Both fine-tuned and zero-shot MLLMs perform better on single-image QA than in multi-image QA, which contrasts with typical vision-language pre-trained models. This might be because mPLUG, despite its SoTA performance in single-image caption-based retrieval, struggles with multiple question-related images. MLLMs have not seen multi-image QA datasets during pretraining, making their vision-language alignment less effective for this task. There is an urgent need for a MLLM capable of handling retrieval-based multi-image QA.

Analysis

Performance with Different Question Types. Figure 5 provides the quantitative result across various question categories, from which we can observe: (1) **Our method significantly surpasses the baseline in multiple areas**, including counting, relation and color questions, with varying numbers of retrieved images. Specifically, we find the improvements in count question (+26.93% $\sim$ +7.68%), relation question (+25.75% $\sim$ +10.05%) and color question (+16.25% $\sim$ +4.84%). (2) **Our method is more proficient at solving count question**. (80.52% *vs* 53.59%) This might be because description-form MHyS offers more specific contextual information, while question-form MHyS focuses on key object, enhancing robustness in solving count question.

Qualitative Analysis. To visually demonstrate the effectiveness of our MHyS-driven paradigm ("Introduce retrieval into QA"), we present four examples (count, attribute, relation and shape question) in Figure 5. Figure 5(a) illustrates **question-form MHyS effectively addresses count question**. For the query (question) "Do the relevant images contain the same number of kids?", the question-form MHyS includes "a young boy" and "the little girl", aiding in retrieving (query word "kids") and comparing their numbers. Figure 5(b) clarifies **both question-form and description-form MHyS collaboratively enhance the ability to answer attribute question**. For the query "Is the coat red in the images?", the question-form MHyS contains "red coat", aligning with the query "coat red", while the description-form MHyS in the first image provides a contrasting clue "cyan outerwear", helping determine the correct answer "No". Figure 5(c) demonstrates **the robustness of MHyS in solving relation question**. For the query "who wears a belt?", the description-form MHyS identifies individuals like "people wearing belts" and "woman wearing a belt", despite the presence of noisy objects. Figure 5(d) highlights that **MHyS provides valuable answer-related clues for solving complex open-ended shape question**. For the query "What are the shapes of the hat and pizza?", the description-form MHyS identifies "a conical hat" and the question-form MHyS offers "rectangular pizza", enhancing the model's effectiveness. In comparison, the baseline CLIP-VLBART without MHyS fails to provide correct answers.

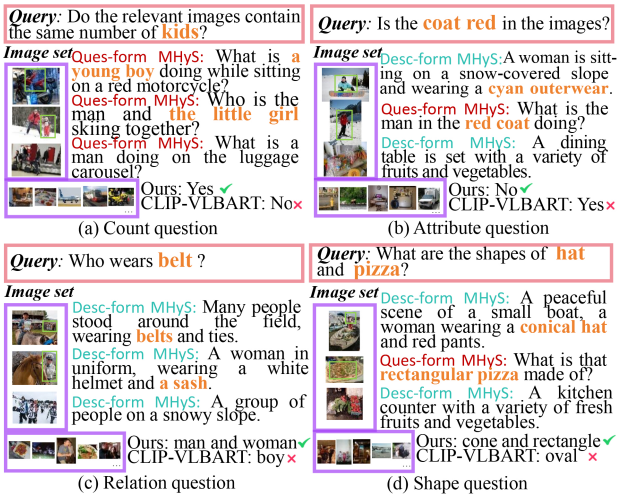

Figure 5: Qualitative comparison between our method and the baseline.

Conclusion

In this paper, we propose using multimodal hypothetical summary (MHyS) to introduce and reference retrieval into QA. This paradigm replaces queried real images with question-form and description-form MHyS in retrieval, enabling transformation into text-to-text retrieval. Incorporating multiple perspectives, MHyS provides question-attended evidence for QA. We employ a coarse-to-fine strategy to filter irrelevant details via sentence- and word-level similarity. Besides, contrastive enhancement loss is used to further align MHyS with the query. Our approach achieves a 3.7% improvement over state-of-the-art methods on RETVQA and a 14.5% gain over CLIP. We hope our work inspires more researchers to explore retrieval-based multi-image QA.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62472419, 62472420, 62072212, 62302218), the Development Project of Jilin Province of China (No. 20220508125RC).

References

Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *CVPR*, 2425–2433.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2024. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. In *ICLR*.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G. B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; et al. 2022. Improving language models by retrieving from trillions of tokens. In *ICML*, 2206–2240.

Chang, Y.; Narang, M.; Suzuki, H.; Cao, G.; Gao, J.; and Bisk, Y. 2022. Webqa: Multihop and multimodal qa. In *CVPR*, 16495–16504.

Chen, W.; Hu, H.; Chen, X.; Verga, P.; and Cohen, W. 2022. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In *EMNLP*, 5558–5570.

Cho, J.; Lei, J.; Tan, H.; and Bansal, M. 2021. Unifying vision-and-language tasks via text generation. In *ICML*, 1931–1942. PMLR.

Eibich, M.; Nagpal, S.; and Fred-Ojala, A. 2024. ARAGOG: Advanced RAG Output Grading. *arXiv preprint arXiv:2404.01037*.

Gao, L.; Ma, X.; Lin, J.; and Callan, J. 2023. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *ACL*, 1762–1777.

Gu, N.; Fu, P.; Liu, X.; Shen, B.; Lin, Z.; and Wang, W. 2024. Light-PEFT: Lightening Parameter-Efficient Fine-Tuning via Early Pruning. In *Findings of ACL*, 7528–7541.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M. 2020. Retrieval augmented language model pre-training. In *ICML*, 3929–3938.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 6700–6709.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear attention networks. In *NeurIPS*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, 9459–9474.

Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; et al. 2022a. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. In *EMNLP*, 7241–7259.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*. PMLR.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2020. What Does BERT with Vision Look At? In *ACL*, 5265–5275.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755. Springer.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *CVPR*, 26296–26306.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 3195–3204.

Penamakuri, A. S.; Gupta, M.; Gupta, M. D.; and Mishra, A. 2023. Answer mining from a pool of images: towards retrieval-based visual question answering. In *IJCAI*, 1312–1321.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Qu, L.; Liu, M.; Wang, W.; Zheng, Z.; Nie, L.; and Chua, T.-S. 2023. Learnable Pillar-based Re-ranking for Image-Text Retrieval. In *SIGIR*, 1252–1261.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge. In *ECCV*, 146–162.

Si, Q.; Liu, Y.; Meng, F.; Lin, Z.; Fu, P.; Cao, Y.; Wang, W.; and Zhou, J. 2022a. Towards Robust Visual Question Answering: Making the Most of Biased Samples via Contrastive Learning. In *EMNLP*.

Si, Q.; Meng, F.; Zheng, M.; Lin, Z.; Liu, Y.; Fu, P.; Cao, Y.; Wang, W.; and Zhou, J. 2022b. Language Prior Is Not the Only Shortcut: A Benchmark for Shortcut Learning in VQA. In *EMNLP*, 3698–3712.

Si, Q.; Mo, Y.; Lin, Z.; Ji, H.; and Wang, W. 2023. Combo of Thinking and Observing for Outside-Knowledge VQA. In *ACL*.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*.

Ushio, A.; Alva-Manchego, F.; and Camacho-Collados, J. 2023. An Empirical Comparison of LM-based Question and Answer Generation Methods. In *ACL*, 14262–14272.

Wen, H.; Song, X.; Chen, X.; Wei, Y.; Nie, L.; and Chua, T.-S. 2024. Simple but Effective Raw-Data Level Multimodal Fusion for Composed Image Retrieval. In *SIGIR*.

Yasunaga, M.; Aghajanyan, A.; Shi, W.; James, R.; Leskovec, J.; Liang, P.; Lewis, M.; Zettlemoyer, L.; and Yih, W.-T. 2023. Retrieval-Augmented Multimodal Language Modeling. In *ICML*, 39755–39769. PMLR.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*, 13040–13051.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Yu, B.; Fu, C.; Yu, H.; Huang, F.; and Li, Y. 2023. Unified Language Representation for Question Answering over Text, Tables, and Images. In *ACL*, 4756–4765.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 6720–6731.

Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; and Gao, J. 2020. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, volume 34, 13041–13049.

Zhu, Z.; Yu, J.; Wang, Y.; Sun, Y.; Hu, Y.; and Wu, Q. 2020. Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *IJCAI*.