RA-BLIP: Multimodal Adaptive Retrieval-Augmented Bootstrapping Language-Image Pre-training

Muhe Ding, Yang Ma, Pengda Qin, Jianlong Wu, Member, IEEE, Yuhong Li, Liqiang Nie, Senior Member, IEEE

Abstract—Multimodal Large Language Models (MLLMs) have recently received substantial interest, which shows their emerging potential as general-purpose models for various vision-language tasks. MLLMs involve significant external knowledge within their parameters; however, it is challenging to continually update these models with the latest knowledge, which involves huge computational costs and poor interpretability. Retrieval augmentation techniques have proven to be effective plugins for both LLMs and MLLMs. In this study, we propose multimodal adaptive Retrieval-Augmented Bootstrapping Language-Image Pretraining (RA-BLIP), a novel retrieval-augmented framework for various MLLMs. Considering the redundant information within vision modality, we first leverage the question to instruct the extraction of visual information through interactions with one set of learnable queries, minimizing irrelevant interference during retrieval and generation. Besides, we introduce a pre-trained multimodal adaptive fusion module to achieve question text-tomultimodal retrieval and integration of multimodal knowledge by projecting visual and language modalities into a unified semantic space. Furthermore, we present an Adaptive Selection Knowledge Generation (ASKG) strategy to train the generator to autonomously discern the relevance of retrieved knowledge, which realizes excellent denoising performance. Extensive experiments on open multimodal question-answering datasets demonstrate that RA-BLIP achieves significant performance and surpasses the state-of-the-art retrieval-augmented models.

Index Terms—Retrieval-augmented model, vision-language pre-training, multimodal retrieval, open question answering.

I. INTRODUCTION

THE birth of the Internet has triggered an unprecedented information revolution, catapulting humanity into the era of information explosion. It is a great challenge to efficiently find answers from a vast amount of information based on our questions. Open Multimodal Multihop Question Answering (MMQA) [1]–[7] can help alleviate this problem of information overload by retrieving external knowledge based on questions and generating correct answers. In recent years, several advanced LLMs and MLLMs like FlanT5 [8], LLaMA [9],

This work was supported in part by the National Natural Science Foundation of China under Grant 62376069, in part by Young Elite Scientists Sponsorship Program by CAST under Grant 2023QNRC001, and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515012027.

Muhe Ding, Jianlong Wu and Liqiang Nie are with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China (e-mail: dmh1216380870@gmail.com, jlwu1992@pku.edu.cn, nieliqiang@gmail.com).

Yang Ma is with the School of Computer Science, University of Sydney, Sydney, NSW 2006, Australia (e-mail: yama5878@uni.sydney.edu.au).

Pengda Qin and Yuhong Li are with the Security Department, Alibaba Group, Hangzhou 311121, China (e-mail: qinpengda0406@163.com, daniel.yuhong@gmail.com).

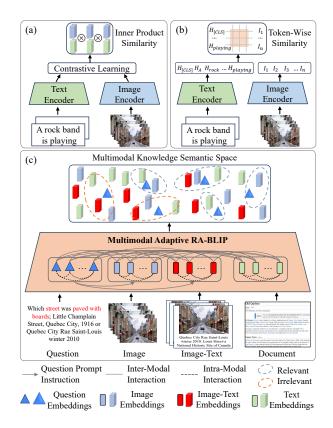


Fig. 1. Illustration of different multimodal retrieval approaches. (a) Cross-modality retrieval. (b) Late-interaction retrieval. (c) RA-BLIP multimodal adaptive retrieval. For RA-BLIP, questions, documents, images, and imagetext pairs are projected into a unified multimodal space.

BLIP-2 [10], GPT-4 [11], etc., have been notably explored to enhance their performance by implicitly encoding a substantial amount of external knowledge within their parameters, which now scale into the hundreds of billions [12]. While these models have yielded exciting results on various multimodal tasks, they have also encountered high computational costs and significant challenges in terms of interpretability.

To alleviate the challenge, many researchers proposed retrieval augmentation techniques that divide the model into two key components: the retriever and the generator [13]–[16]. The retriever accesses relevant knowledge from a knowledge base based on the posed question, while the generator leverages this information to create textual output in response. In earlier stages, text-modality retrieval-augmented models, such as REALM [13], RAG [17], and so on [18], [19], have been

proposed to solve text-only question answering. They build the dense index as a non-parametric document memory from extensive textual sources like Wikipedia for effective knowledge retrieval, and the generator produces answers based on the retrieved knowledge. More recently, multimodal retrieval-augmented models, such as MuRAG [14], SKURG [20], and so on [16], [21], have emerged one after another. These models extend the knowledge memory across various modalities, employing pre-trained visual language models to retrieve relevant evidence and support reasoning for answers.

However, existing methods exhibit certain limitations. The first limitation is the insufficient integration and interaction between vision and language. On the one hand, existing methods lack explicit integration of multimodal information, hindering the alignment of questions and multimodal knowledge in the semantic space. As shown in Fig. 1(a), some methods [14], [22], [23] have employed separate visual encoder and text encoder for individual modality encoding and adopted contrastive learning [24] for multimodal alignment to retrieve. This may lead to an unbalanced and biased multimodal retrieval and reasoning process towards specific modalities. Besides, late-interaction retrieval approaches [25]-[27] in Fig. 1(b), retain dual-encoder independent encoding architecture and perform token-wise interactions only in the late scoring stage, which sacrifices the retrieval efficiency for the benefits of fine-grained feature learning. On the other hand, existing methods [14], [20], [28], [29] do not utilize questions to instruct the image encoder in selectively extracting visual features, and thus suffer from interference and noise caused by redundant information in images. Moreover, such methods lack mutual instruction when encoding different modal features, making it challenging to model relationships between multiple sources. The second limitation is that existing approaches do not inspect the correctness of the retrieved relevant knowledge at the generation stage. However, the retrieved knowledge contains significant noise, resulting in poor model anti-interference and robustness. The generator assumes all retrieved relevant knowledge is correct, potentially leading to the utilization of irrelevant or confusing information.

To address the above issues, we propose multimodal adaptive Retrieval-Augmented Bootstrapping Language-Image Pretraining (RA-BLIP). RA-BLIP consists of two key components: a multimodal adaptive retrieval-augmented framework and an adaptive selection knowledge generation (ASKG) strategy. To tackle the first limitation, RA-BLIP is based on the InstructBLIP architecture [30] and adopts Q-Former to implement instruction-aware visual feature extraction that uses questions as instructions. The question instruction interacts with the query embeddings through shared self-attention layers and encourages the extraction of question-relevant visual features. Additionally, we incorporate a pre-trained multimodal adaptive fusion module to fuse vision and text information, obtaining multimodal features [31]-[33]. As a result, RA-BLIP achieves question text-to-multimodal retrieval by aligning the questions and multimodal knowledge bases in the semantic space of three modalities: text, image, and imagetext [34], as shown in Fig. 1(c). For the second limitation, we leverage the implicit capabilities of LLMs and introduce an adaptive selection knowledge generation strategy, which gives the generator the capability of selecting knowledge by data enhancement to make the model automatically judge the relevance of knowledge. ASKG strategy allows the generator to not simply rely on the word similarity between the question and knowledge, but to understand the semantic information of question and know which knowledge contains the answer. Furthermore, the parameters of the image encoder and LLM of our framework are frozen, significantly reducing computational costs. Extensive experiments on three representative QA datasets demonstrate the effectiveness of our methods.

Overall, our key contributions are as follows:

- We propose a novel multimodal adaptive retrievalaugmented framework, which achieves question text-tomultimodal retrieval and knowledge-intensive multimodal QA by integrating visual and language modalities and projecting them into a unified semantic space.
- We introduce an adaptive selection knowledge generation strategy that leverages the powerful capabilities of LLMs to select the relevant retrieved knowledge for answer reasoning autonomously.
- We conduct extensive experiments on various multimodal and multihop datasets (i.e., WebQA [4], MultimodalQA [5], and MMCoQA [6]). RA-BLIP demonstrates superiority over the existing state-of-the-art retrieval-augmented models.

II. RELATED WORK

A. Vision-Language Pretraining

Vision-language pre-training (VLP) aims to train models on large-scale image-text datasets to capture the relationship between these two modalities. Broadly, VLP methodologies fall into two categories based on their training approach: 1) End-to-end Methods: This category includes methods [35]-[38] that train models end-to-end, backpropagating learned signals to achieve mutual learning between different modalities. 2) Modular Methods: In contrast, modular methods, as seen in works by [39]–[43], involve keeping the parameters of specific pre-trained components (like image encoders or large language models) fixed while focusing on refining other aspects of the model. For instance, LiT [44] utilizes a pre-trained frozen image encoder from CLIP, while Flamingo [45] and BLIP-2 [10] freeze the language model to integrate LLMs into visionlanguage tasks better. Besides, instruction tuning is also an effective approach during VLP. InstructBLIP [30] represents a recent advancement in this area, achieving instruction-aware visual feature extraction and instruction-guided LLM generation. The unique capability of InstructBLIP to extract features based on prompt instructions, combined with its utilization of frozen LLMs and image encoders, positions it as an ideal backbone for our proposed retrieval-augmented framework.

B. Text-modality Retrieval-Augmented Models

Retrieval-augmented techniques have proven to be effective plugins for both LLMs and MLLMs in academia. These techniques extract pertinent world knowledge from

extensive databases, subsequently integrating this information to formulate answers. Pioneering methods like ORQA [3] have employed inverse cloze tasks for retriever pre-training, showcasing their efficacy on open-ended question answering datasets. Following suit, REALM [13] extends this by retrieving and processing documents from comprehensive sources like Wikipedia for logical reasoning, which is pre-trained to reason over a large corpus of knowledge on the fly during inference. Furthermore, RAG [17] adopts a pre-trained model and non-parametric memory for language generation. FiD [18] leverages encoder-decoder transformer models for knowledgeintensive tasks, setting new benchmarks in QA. More recently, RETRO [19] has advanced these methods by handling longer sequences and accessing diverse documents for segmented sequences from expansive retrieval datasets. Text-modality retrieval-augmented models construct dense indices as nonparametric document memories, using extensive textual knowledge bases to align the retrieval process with specific queries. Despite these advancements, a notable limitation remains the need for these methods to effectively leverage vast multimodal knowledge, thereby constraining their applicability in the domain of open multimodal question answering.

C. Multimodal Retrieval-Augmented Models

To overcome the limitations of text-modality retrievalaugmented models, recent research [10], [15], [30], [46], [47] has made strides in integrating multimodal knowledge. Notable efforts, including AutoRouting [5] and MAE [6], involve training distinct models for each modality and using classifiers for task-specific routing, though this approach often hampers cross-modal reasoning. MuRAG [14] seeks to overcome this limitation by employing separate encoders for visual and textual modalities, followed by a joint encoder for multimodal fusion. However, this approach lacks integrated guidance for different modalities and cannot model the relations between knowledge sources during retrieval. SKURG [20] attempts to bridge this gap by using an entity-centered fusion encoder to align modalities, yet faces challenges in computational efficiency and limited interpretability. Besides, Solar [21] transforms multimodal inputs into a unified language format but falls short in handling complex tasks and generalizing visual information. REVAL [16] leverages large-scale knowledge graphs to assist visual language pre-training, but it brings a lot of calculations. In contrast, RA-BLIP distinguishes itself by seamlessly integrating visual and language modalities into a cohesive semantic space, enabling the autonomous selection of relevant knowledge for reasoning, thus addressing these limitations more effectively.

III. METHODOLOGY

In this section, we first formulate the research problem and subsequently elaborate on the model architecture of our retrieval-augmented framework. Then, we describe the learnable query interaction approach, followed by multimodal adaptive fusion module. Subsequently, we show how to train the retriever and rank the relevant knowledge. Lastly, we introduce the adaptive selection knowledge generation strategy.

A. Problem Formulation

This paper presents a multimodal adaptive retrievalaugmented framework called RA-BLIP for open multihop and multimodal QA, integrating retrieval and generation functions. For knowledge-intensive QA, we deconstruct the task into two stages: retrieval and generation, which are implemented by the retriever and generator respectively. The goal of our model training is to learn the distribution $P(y|x_q)$ to generate a textual output y conditioned on input question x_q and multimodal knowledge base KB ($KB = k_1, ..., k_n$). Firstly, the retriever encodes questions, images, and texts from the knowledge base KB. It identifies the most relevant retrieved knowledge, $\mathcal{K}_{ret} \subset \mathcal{KB}$ (\mathcal{K}_{ret} is the retrieved knowledge) for each question x_q , which is modeled as $p(\mathcal{K}_{ret}|x_q)$. Secondly, the generator utilizes an LLM to generate answers y, conditioned on both the question and the retrieved knowledge, which is modeled as $p(y|x_q, \mathcal{K}_{ret})$. We treat multimodal knowledge \mathcal{K}_{ret} as a latent variable from the external knowledge base and marginalize it to increase the overall likelihood of the answer y. The overall process is encapsulated in the equation:

3

$$p(y \mid x_q) = \sum_{\mathcal{K}_{ret} \subset \mathcal{KB}} \underbrace{p(\mathcal{K}_{ret} \mid x_q)}_{\text{Retrieval}} \cdot \underbrace{p(y \mid x_q, \mathcal{K}_{ret})}_{\text{Generation}}. \tag{1}$$

This dual-stage framework effectively addresses the complexities of open multimodal QA by balancing the retrieval of multimodal data and knowledge-based generation, and has been validated by extensive experiments and ablation studies.

B. Model Architecture

RA-BLIP is built on a simple backbone model that is pretrained to encode image-text pairs so that they are suitable for both knowledge base retrieval and answer generation. The overall framework of RA-BLIP is shown in Fig. 2. The backbone model consists of a multimodal encoder $f_{\theta}(\cdot)$ and decoder $q_{\theta}(\cdot)$, which are used as components of the RA-BLIP model to implement retrieval and generation. The multimodal encoder $f_{\theta}(\cdot)$ contains a frozen image encoder ViT [48], Q-Former architecture [30], and the pre-trained multimodal adaptive fusion module. The decoder $g_{\theta}(\cdot)$ is composed of a LLM FlanT5 [8]. Querying Transformer (Q-Former) [10] is a lightweight Transformer consisting of two modules that share the same self-attention layer: one is an image transformer that interacts with the frozen image encoder ViT for visual feature extraction, and the other is a text transformer can act as both text encoder and text decoder for text feature. The visual encoder, composed of ViT and Q-Former image transformer, has instruction-aware visual feature extraction capabilities and can extract visual information based on question instructions. We input N learnable query embeddings into the Q-Former image transformer, which interacts with frozen image features through cross-attention layers to obtain visual representation $f_{\theta}(I) \in \mathbb{R}^{N \times D}$, where D is the hidden dimension of the Q-Former. Additionally, we use the Q-Former text transformer to encode text, taking the [CLS] token as the text representation $f_{\theta}(T) \in \mathbb{R}^{1 \times D}$. To obtain multimodal features combining both image and text, we introduce a pre-trained multimodal adaptive fusion module $M_{\theta}(\cdot)$ to obtain the multimodal representation

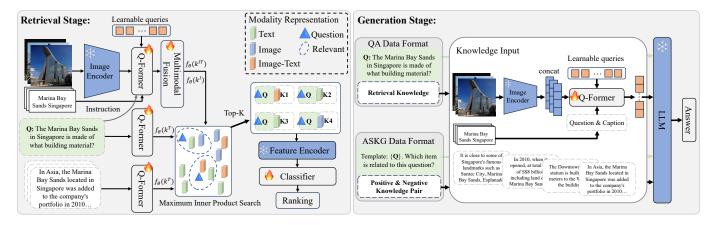


Fig. 2. The overall workflow of RA-BLIP consists of a retrieval stage and a generation stage. We utilize multimodal encoder $f_{\theta}(\cdot)$ to project questions and multimodal knowledge into a unified semantic space to achieve question text-to-multimodal retrieval, and exclude confusing knowledge by ranking. Additionally, we employ ASKG strategy to filter out invalid knowledge, enabling precise reasoning. The parameters of LLM and image encoder are fixed, only the O-Former and classifier are trainable.

 $f_{\theta}(IT) \in \mathbb{R}^{(N+1) \times D}$. Consequently, the multimodal encoder can simultaneously encode image, text, and image-text features. Questions and knowledge are encoded into multimodal information through the multimodal encoder and then input into the decoder for answer generation. In the generation stage, compared with the retriever, the multimodal encoder discards the multimodal adaptive fusion module to reduce the computational cost.

C. Learnable Query Interaction for Multi-images

Questions and image captions are used as instructions to extract visual features to get learnable query embeddings and input them together with text knowledge to LLMs for generation. We employ a novel approach for extracting visual information to alleviate the burden of LLMs in distinguishing knowledge. In the original Q-Former in [30], multiple images are processed by employing multiple separate sets of queries for each, with each set of queries independently extracting visual features. This results in using multiple query features for generation, which can be computationally intensive and less efficient in capturing the interrelations among different images. In contrast, our method innovates by succinctly utilizing one set of learnable queries to directly interact with and extract features from multiple images in a unified manner. This process occurs during the Q-Former stage, enabling more efficient and integrated interaction among multiple visual references. By employing one set of query interaction approach, RA-BLIP not only simplifies the feature extraction process but also enhances the efficiency of information extraction. This unified interaction allows the model to understand better and represent the collective information presented in multiple images, enabling more effective and cohesive feature utilization, especially when dealing with complex scenes or subjects across multiple images.

D. Multimodal Adaptive Fusion Module

The pre-trained multimodal adaptive fusion module $M_{\theta}(\cdot)$ consists of a 3-layer BERT network [31], [32]. Since the

Q-Former has aligned visual and text feature representation, we use Image-Text Matching loss and Image-grounded Text Generation loss for pre-training. As shown in Fig. 3, our approach is to fix the parameters of the image encoder and O-Former, and solely fine-tune the parameters of the multimodal adaptive fusion module. The module concatenates the visual embedding and text embedding with a dimension of $\mathbb{R}^{(N+L)\times D}$, where N is the learnable query embeddings and L is the length of text tokens. Image-text matching loss is used to fuse image and text representations, and the results of the fusion module are fed into a binary linear classifier for each output query to obtain the logits and take the average logits of all queries as the matching score. Given a pre-training dataset $\mathcal{X} = \{I_i, T_i\}_{i=1}^n$, we randomly sample negative texts for each image and randomly sample negative images for each text, to generate negative training data. Therefore, we denote the ground truth label as $y \in \{1,0\}$ for each image-text pair (I_i, T_i) , indicating if the input image-text pair is relevant or not. We use the multimodal encoder $f_{\theta}(\cdot)$ to encode image-text pairs and input it into the multimodal adaptive fusion module $M_{\theta}(\cdot)$. The objective function is defined as follows:

$$\mathcal{L}_{itm} = -\frac{1}{n} \sum_{I_i, T_i \in \mathcal{X}} y \log \left(\rho \left(M_{\theta} \left(f_{\theta}(I_i); f_{\theta}(T_i) \right) \right) \right), \quad (2)$$

where $\rho(\cdot)$ is the softmax function. Image-grounded Text Generation loss trains the fusion module to generate texts, given input images as the condition [10], [30]. For the image-text pairs in the pre-training dataset, each image I corresponds to a text sentence $\mathbf{y}_{1:T} = \{y_1, ..., y_T\}$ of length T. We employ a multimodal causal self-attention mask for multimodal encoder $f_{\theta}(\cdot)$ and multimodal adaptive fusion module $M_{\theta}(\cdot)$ to control the interaction between queries and text. The visual query functions as a prefix causal, ensuring that queries can attend to each other while excluding text tokens. Similarly, each text token y can attend to all visual queries and preceding text tokens. The loss function is defined as:

$$\mathcal{L}_{itg} = -\sum_{t=1}^{T} \log M_{\theta}(f_{\theta}((y_t \mid y_{< t}, I))). \tag{3}$$

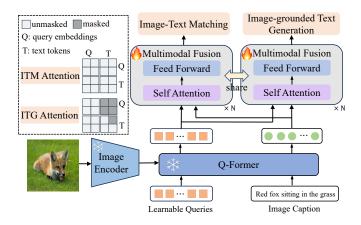


Fig. 3. Schematic diagram of the pre-training process of multimodal adaptive fusion module.

E. Retrieval-Augmented Retriever Training

During the retrieval stage, the retriever utilizes the question x_q to retrieve relevant knowledge from multimodal knowledge base KB. To achieve this, we apply the multimodal encoder $f_{\theta}(\cdot)$, which encodes the question x_q along with all latent multimodal knowledge into an embedding space to identify the Top-K most relevant candidates, illustrated in Fig. 2 retrieval stage. We use contrastive learning to construct positive and negative samples for training. The knowledge type consists primarily of three components: image k^{T} , text k^{T} , and image-text k^{IT} . Thus, the l-th example in the dataset is represented as $(x_l,y_l,\{\hat{k}_i^I,\hat{k}_i^T,\hat{k}_i^{IT}\}_l,\{\overline{k}_j^I,\overline{k}_j^T,\overline{k}_j^{IT}\}_l)$, where \hat{k}_i is the *i*-th positive (image, text, image-text) sample and k_i represents j-th negative (image, text, image-text) sample. For a batch of knowledge examples, we gather all associated positive and negative knowledge sources into a batch $\mathcal{K}_B = \{\{\hat{k}_i^I, \hat{k}_i^{IT}, \hat{k}_i^{IT}\}_1, \{\overline{k}_j^I, \overline{k}_j^{TT}, \overline{k}_j^{IT}\}_1, ..., \{\overline{k}_j^I, \overline{k}_j^T, \overline{k}_j^{IT}\}_B\}.$ The multimodal encoder is responsible for encoding the multimodal feature representations and aligning the questions and knowledge within the unified semantic space. This alignment facilitates identifying the proximity between a question and its corresponding knowledge through contrastive learning. The objective function is defined as follows:

$$\mathcal{L}_{con} = -\log \frac{\exp(f_{\theta}(x_q) \cdot f_{\theta}(\hat{k}^I; \hat{k}^T; \hat{k}^{IT}))}{\sum\limits_{k \in \mathcal{K}_B} \exp(f_{\theta}(x_q) \cdot f_{\theta}(k^I; k^T; k^{IT}))}, \quad (4)$$

where $f_{\theta}(\cdot)$ is the multimodal encoder and \mathcal{K}_{B} is a batch of knowledge sources. We use the multimodal encoder trained to encode text, image, and image-text features, and apply Maximum Inner Product Search (MIPS) [49] to select Top-K from knowledge base \mathcal{KB} as the relevant \mathcal{K}_{ret} , as shown in the following:

$$\operatorname{TopK}(\mathcal{K}_{ret} \mid x_q) = \operatorname{TopK}_{k \in \mathcal{KB}} \{ f_{\theta}(x_q) \cdot f_{\theta}(k^I; k^T; k^{IT}) \}. \quad (5)$$

Although the retriever is more efficient for many retrieval tasks, its accuracy is lower on open multimodal question answering. There are instances where certain knowledge is confusing and bears token-wise similarity to the question

at the feature level, yet it fails to understand the question and cannot provide an answer in the semantic space [14]. For instance, consider the question "The Marina Bay Sands in Singapore is made of what building material?", and the relevant text knowledge "Singapore is also the new downtown of Singapore, built on reclaimed land.". The question is about the building materials of a Singapore hotel, but this text is about the location of a Singapore hotel. Despite their tokenwise similarity in the words, they cannot answer this question and cause confusion. To address this issue, we introduce a rank strategy to sort the Top-K candidates and exclude confusing samples, where K is the maximum number of positive samples corresponding to the question. We select the Top-K samples, categorize them into positive and negative samples based on the ground truth, and then perform rank training. Since the Q-Former does not have a classifier, we input the multimodal features output $f_{\theta}(k^I; k^T; k^{IT})$ by the multimodal encoder into the fixed-parameter LLMs encoder and trainable classifier z. The loss function is defined as:

5

$$\mathcal{L}_{cls} = CrossEntropy\left(\boldsymbol{z}(f_{\theta}(k^{I}, k^{T}, k^{IT})), \boldsymbol{y}\right), \quad (6)$$

where y is the ground truth about the knowledge is relevant or not.

F. Adaptive Selection Knowledge Generation

During the generation stage, the retrieved multimodal knowledge is combined with the question x_q as an augmented input $[k_1, ..., k_l, x_q]$, which is fed to the multimodal encoder and LLMs [8] to produce multimodal representation encoding and generate answers. We observe that existing methods [14], [21] directly rely on retrieval results without distinguishing the correctness of the retrieved knowledge, potentially leading to the utilization of incorrect, confusing, or irrelevant information. To address this, we propose an adaptive selection knowledge generation (ASKG) strategy based on a questionand-answer formulation, shown in Fig. 2 generation stage. ASKG strategy enables the generator to go beyond mere word similarity between the question and the retrieved knowledge, allowing it to grasp the semantic information of the question and identify which piece of knowledge contains the answer. Specifically, we manually construct question-and-answer data to enable the model to discriminate the relevance of multimodal knowledge, thereby utilizing the implicit capabilities of LLMs for knowledge filtering. Based on the original dataset, we select relevant knowledge as positive examples and irrelevant knowledge as negative examples. We create an ASKG enhanced dataset and combine the knowledge according to templates, with the identifier of the positive examples serving as the answer. The template for the question \tilde{x}_q is : "We would like to request your feedback on ranking the questions according to their relevance to the references below. Relevance refers to the degree to which the reference can answer the question. The input format is Question: [content], Reference [knowledge ID]: [content]. The output format is: Related content is [knowledge ID].", and the answer \tilde{y} is in the form of "The most relevant reference is Reference [knowledge ID].".

We refer to the above enhanced dataset of questions and answers as \widetilde{x}_q and $\widetilde{\mathbf{y}} = {\widetilde{y}_1, ..., \widetilde{y}_M}$, where M is the text

Algorithm 1 The Training pipeline for RA-BLIP.

Retrieval Training Stage

Input: question $\{x_{qi}\}_{i=1}^{N}$, knowledge base \mathcal{KB} for sampled mini-batch x_q and \mathcal{K}_B do

Compute contrastive loss L_{con} by Eq. (4)

end for

Return retrieval model θ_{ret}

Input: question $\{x_{qi}\}_{i=1}^N$, ground truth \boldsymbol{y} , Top-K

retrieved knowledge from θ_{ret}

for sampled mini-batch x_q , and $\operatorname{Top} \mathcal{K}_B$ do

Calculate crossentropy loss L_{cls} by Eq. (6)

end for

Return ranking model θ_{ran}

Generation Training Stage

Input: question $\{x_{qi}\}_{i=1}^{N}$, answer \mathbf{y} , \mathcal{K}_{ret} from θ_{ran} , ASKG datasets \widetilde{x}_q and $\widetilde{\mathbf{y}}$

for sampled mini-batch x_q , \mathcal{K}_{ret} and \widetilde{x}_q do

Compute generation loss L_{gen} by Eq. (7)

end for

Return generation model θ_{qen}

length. Given the dataset question x_q and ground-turth answer of length T, $\mathbf{y}_{1:T} = \{y_1, ..., y_T\}$, as well as the constructed \widetilde{x}_q and $\widetilde{\mathbf{y}}$, the generator $g_{\theta}(\cdot)$ utilizes attention over question x_q and relevant knowledge \mathcal{K}_{ret} encoded by multimodal encoder $f_{\theta}(\cdot)$ to generate textual outputs token by token. The final generation loss is defined by:

$$\mathcal{L}_{gen} = \sum_{i=1}^{T} -\log g_{\theta} \left(y_{i} \mid y_{1:i-1}, f_{\theta}(x_{q}, \mathcal{K}_{ret}) \right)$$

$$+\alpha \sum_{i=1}^{M} -\log g_{\theta} \left(\widetilde{y}_{i} \mid \widetilde{y}_{1:i-1}, f_{\theta}(\widetilde{x}_{q}) \right),$$

$$(7)$$

where α is the hyperparameter which will be discussed in section IV-D. To give a clear illustration of RA-BLIP, we summarize the training pipeline in Algorithm 1.

IV. EXPERIMENTS

A. Experimental Settings

- 1) Datasets: We evaluate our method on three QA datasets: WebQA [4], MultimodalQA [5], and MMCoQA [6]. The details of these datasets are showcased in Table I.
 - WebQA [4] is a large-scale dataset for multimodal and multihop QA where all questions are knowledge-seeking queries that require two or more knowledge sources. Evaluation metrics are retrieval F1 and QA for assessing answer generation quality, which is measured as both fluency (QA-FL) and accuracy (QA-ACC). We calculate fluency through BARTScore [50] and evaluate accuracy via F1 and recall. The fluency score and accuracy score are multiplied FL * Acc to calculate the overall score.
 - MultimodalQA [5] is a collection of multihop QA pairs that necessitate the fusion of knowledge from text, tables, and images. This dataset requires retrieval and reasoning

TABLE I
OVERALL DETAILS OF DOWNSTREAM DATASETS.

Dataset	Train	Dev	Test
WebQA [4]	34.2K	5K	7.5K
MultimodalQA [5]	23.8K	2.4K	3.6K
MMCoQA [6]	4.6K	0.6K	0.6K

across text, image, and tabular data types. The performance of MultimodalQA is measured by F1 score at the word level and the Exact Match (EM) of the answers.

- MMCoQA [6] is the first dataset constructed for multimodal conversational QA tasks and aims to answer users' questions with multimodal knowledge sources via multi-turn conversations. It comprises multiple supervised signals, including decontextualized questions, answers, and corresponding evidence.
- 2) Compared Methods: For WebQA, MultimodalQA, and MMCoQA, we make comparisons with different baseline methods. The model parameter quantity comparison is shown in Table II. The number of Solar parameters is not published, and both Solar and SKURG use other models to exploit multimodal information without accounting for the parameter counts of other models. We have frozen LLM and ViT, focusing solely on training Q-Former, which has fewer trainable parameters and bfloat16 encoding. In order to verify the scaling law [51], we selected more powerful FlanT5xxl for experiment. To compare LLMs with other methods of similar parameter magnitude, we utilized T5-base and T5-large as benchmarks for a fair comparison. Since T5-base and T5-large are not aligned with the model through pre-training, they need to be fine-tuned during training.
 - VLP [4], [52] pre-trains its transformer-based encoder-decoder with both textual and visual information. They first retrieve knowledge based on the question and feed it into the model to generate answers. In addition, VLP integrates VinVL [41] to improve performance.
 - MuRAG [14] encodes the question and selects Top-K nearest neighbors from multimodal memory. They are then fed into the backbone encoder-decoder to generate textual outputs token by token. The backbone model uses T5-base [53] and ViT-large [48], respectively.
 - SKURG [20] takes multimodal information sources as input and encodes them separately, then utilizes an entity-centered fusion encoder to align the sources of different modalities via the shared entities and structured knowledge. The method adopts OFA-base [54] and BART-base [55]. Besides, it integrates ELMo-based NER [56] and OpenNRE [57] for entity and relation extraction.
 - Solar [21] first converts multimodal inputs into textual data and then utilizes a T5 [53] to generate answers through retrieval, ranking, and decoding. It retrieves and ranks the information using BERT [32]. Additionally, it adopts BLIP [58] for image caption generation and VinVL [41] for image-attribute feature extraction.
- 3) Implementation Details: Our method includes multimodal fusion pre-training, retrieval, ranking, and generation.

TABLE II

COMPARISON OF PARAMETER QUANTITY. PARAMETER QUANTITIES OF
OTHER METHODS REFER TO [20], [21].

Model	#Trainable Params	#Total Params
VLP+VinVL [4]	220M	220M
MuRAG [14]	527M	527M
SKURG [20]	447M	447M
ImplicitDecomp [5]	1310M	1310M
RA-BLIP (T5-base)	387M	1398M
RA-BLIP (T5-large)	902M	1913M
RA-BLIP (FlanT5xl)	109M	4.1B
RA-BLIP (FlanT5xxl)	109M	12.1B

For WebQA and MultimodalQA, we follow all steps as mentioned above. For MMCoQA, we manually include the positive clues in the retrieval results without performing subsequent ranking, following the process used in previous work [21]. We adopt InstructBLIP [30] with frozen EVA-ViT-g/14 [59] as well as LLM (FlanT5xl and FlanT5xxl) [8] for generation. In order to compare LLMs with fewer than 1 billion parameters, we replaced the LLM backbone with T5-large and T5-base. Due to the inconsistent dimensions between T5 and Q-Former, we added a linear layer to align the dimensions and did not perform pre-training. We keep the image encoder and the LLMs frozen, tuning only the Q-Former and the multimodal fusion module during the pre-training and retrieval stage. We adopt the multimodal encoder and LLM encoder as feature encoder at ranking stage. At the generation stage, we froze the image encoder as well as LLMs and only trained the Q-Former with ASKG. We froze the image encoder and FlanT5 during all training processes, as well as used bfloat16 encoding to achieve RA-BLIP with the fewest trainable parameters.

For pre-training, we pre-train the multimodal adaptive fusion module on the SBU dataset [60]. Our approach is to fix the parameters of the image encoder and Q-Former, and solely fine-tune the parameters of the multimodal adaptive fusion module. We use the AdamW optimizer and adopt cosine learning rate of 1e-5, warmup of 1K steps, and batch size 64 for 10 epochs. For fine-tuning, we use the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and a weight decay of 0.01 uniformly for three datasets. For WebQA retrieval, we use a cosine learning rate of 1e-5, warmup of 1K steps, and batch size 4 for 10 epochs. For WebQA ranking, we select the top 10 samples to train the model with a cosine learning rate of 1e-5, warmup of 1K steps and batch size 40 for 5 epochs. For generation, we adopt cosine learning rate of 1e-6, warmup of 1K steps, and batch size 4 for 10 epochs. We set learning rate of 5e-5 for T5-large and T5-base.

For MultimodalQA [5] and MMCoQA [6], we tested the results on the dev set of MultimodalQA as well as the dev and test sets of the MMCoQA. For MultimodalQA retrieval, we use a cosine learning rate of 1e-5, warmup of 1K steps, and a batch size 4 for 10 epochs. For MultimodalQA generation, we adopt a cosine learning rate of 1e-6, warmup of 1K steps, and a batch size of 8 for 10 epochs. We set learning rate of 1e-5 for T5-large and 5e-5 for T5-base. For MMCoQA, we manually include the positive clues in the retrieval results without performing subsequent ranking, following the process

TABLE III

RESULTS OF WEBQA OFFICIAL TEST-SET. * REPRESENTS LLM IS
FLANT5XL, WHILE † REPRESENTS LLM IS FLANT5XXL. BOLD AND
UNDERLINE DENOTE THE BEST AND PREVIOUS SOTA RESULTS.

Model	Retr-F1↑	QA-FL↑	QA-Acc↑	QA↑
VLP [52]	0.69	42.6	36.7	22.6
VLP+VinVL [41]	0.71	44.2	38.9	24.1
MuRAG [14]	0.75	55.7	54.6	36.1
SKURG [20]	0.88	55.4	57.1	37.7
Solar [21]	0.89	60.9	58.9	40.9
InstructBLIP* [30]	-	51.7	59.0	31.4
InstructBLIP† [30]	-	53.4	62.5	35.0
RA-BLIP (T5-base)	-	62.6	59.7	41.6
RA-BLIP (T5-large)	-	62.9	60.9	42.5
RA-BLIP*	0.83	65.1	65.3	45.8
RA-BLIP†	0.89	65.5	68.7	48.5

used in previous work [21]. We set the learning rate as 1e-5 and batch size as 32 for 10 epochs at the retrieval stage. Then, we use a cosine learning rate of 1e-6, warmup of 1K steps, and a batch size of 16 for 10 epochs at the generation stage. We use the standard evaluation protocol for each dataset and report the same metrics, as well as the random seeds are fixed for reproducibility.

B. Main Results

Results on WebQA. We show the WebQA results in Table III. We can see that RA-BLIP surpasses all baselines in terms of both QA and retrieval F1 scores. RA-BLIP (FlanT5xl) achieves 45.8% accuracy, which is +4.9% higher than the state-of-the-art Solar [21]. Especially the metric OA-Acc is +6.4% higher, proving the model's powerful generation ability. Besides, RA-BLIP (FlanT5xxl) beats SOTA Solar by 7.6% on overall QA accuracy, which shows that RA-BLIP complies with scaling law [51] and can improve the generation accuracy by using more advanced LLM. In order to prove that it is our RA-BLIP framework rather than the advanced MLLM backbone that improves the generative performance, we conducted experiments on InstructBLIP [30] on WebQA. We used RA-BLIP's optimal 0.89 search result for InstructBLIP generation and found that its accuracy was 5.9% lower than Solar, which further proves the effectiveness of RA-BLIP framework and ASKG. RA-BLIP (T5-base) and (T5-large) are not pre-trained to align with Q-Former, but they achieve 41.6% and 42.5% accuracy respectively based on 0.89 search results, surpassing Solar and proving it is the RA-BLIP framework rather than LLM that improves performance. Compared with Solar and SKURG which require additional model assistance, RA-BLIP does not use additional models, but it also achieves very good results in retrieval and is 14% higher than MuRAG, which similarly does not use additional models.

Results on MultimodalQA. We demonstrate MultimodalQA results in Table IV. MultimodalQA contains tables and has many multihop questions that require combining multimodal information. RA-BLIP also improved EM and F1 by 6.0% and 6.6%, respectively, compared to state-of-the-art Solar, which demonstrates the generative ability of our method in incorporating multihop knowledge. In addition,

TABLE IV

MULTIMODALQA DEV-SET RESULTS. * REPRESENTS LLM IS FLANT5XL, WHILE † REPRESENTS LLM IS FLANT5XXL. SINGLE-MODAL AND MUTLI-MODAL RESPECTIVELY INDICATE WHETHER REASONING RELIES ON SINGLE-MODAL OR MUTLI-MODAL KNOWLEDGE. BOLD AND UNDERLINE DENOTE THE BEST AND SOTA RESULTS, RESPECTIVELY.

Model	Single	Single-Modal		Mutli-Modal		All	
	EM	F1	EM	F1	EM	F1	
AutoRouting [5]	51.7	58.5	34.2	40.2	44.7	51.1	
ImplicitDecomp [5]	51.6	58.4	44.6	51.2	48.8	55.5	
SKURG [20]	66.1	69.7	52.5	57.2	<u>59.8</u>	64.0	
Solar [21]	69.7	74.8	55.5	<u>65.4</u>	<u>59.8</u>	<u>66.1</u>	
RA-BLIP (T5-base)	65.4	71.6	59.7	65.7	63.1	69.3	
RA-BLIP (T5-large)	65.2	71.9	62.6	68.4	64.1	70.5	
RA-BLIP*	70.1	77.6	59.3	65.5	65.8	72.7	
RA-BLIP†	69.9	76.4	59.1	65.6	65.6	72.1	

TABLE V

MMCOQA TEST-DEV-SET RESULTS. * REPRESENTS LLM IS FLANT5XL, WHILE † REPRESENTS LLM IS FLANT5XXL. BOLD AND <u>UNDERLINE</u> DENOTE THE BEST AND SOTA RESULTS, RESPECTIVELY.

Model	Dev		Test	
	EM	F1	EM	F1
ORConvQA [61]	1.0	3.0	1.0	1.9
ManyModelQA [62]	0.7	2.3	1.0	1.8
MAE [6]	21.5	30.2	24.9	32.3
Solar [21]	<u>56.8</u>	<u>62.5</u>	<u>57.3</u>	<u>64.6</u>
RA-BLIP*	59.2	67.1	61.0	67.8
RA-BLIP†	58.7	66.7	59.5	66.7

RA-BLIP's accuracy is ahead of SOTA Solar in both single-modality and multi-modality, demonstrating our model can well combine multiple contextual semantic knowledge for cross-modal reasoning. Both RA-BLIP (T5-large) and RA-BLIP (T5-base) surpass Solar, indicating that the performance improvements are due to the RA-BLIP framework rather than the underlying LLM. Notably, the accuracy of more powerful FlanT5xxl is lower than that of FlanT5xl, probably because the powerful LLM is overfitted.

Results on MMCoQA. Our results on MMCoQA are shown in Table V. Compared with WebQA and MultimodalQA, MMCoQA requires the model to correctly incorporate dialog history and develop deep multimodal understanding and reasoning capabilities across multiple conversations. RABLIP achieves a margin of 3.7% enhancement over the best Solar for the EM score and 3.2% for the F1 score in the test split. These results suggest the generalization ability and versatility of our model. Due to the small dataset, RABLIP (FlanT5xxl) has resulted in overfitting, which prevents further performance improvement.

C. Ablation Study

To analyze the effectiveness of our proposed method, we conducted comprehensive ablations on the WebQA dataset in both retrieval and generation stages. As shown in Fig. 4 (a), RA-BLIP (FlanT5xl) and RA-BLIP (FlanT5xxl) with ASKG strategy improved by 1.9% and 2.7% respectively, which

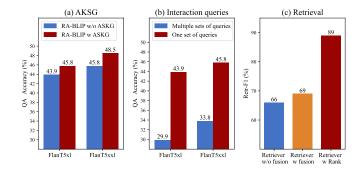


Fig. 4. Ablation study for RA-BLIP generation and retrieval on WebQA.

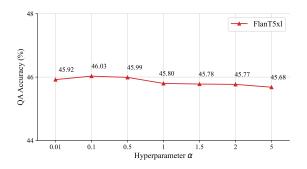


Fig. 5. Influence of varying the hyperparameter α of generation loss.

validates that employing ASKG strategy can assist LLMs in efficiently discerning the relevance of retrieved knowledge and effectively activate the implicit capabilities of more powerful LLMs. In Fig. 4 (b), we compared the effects of one set of queries and multiple sets of queries, where one set of queries has a significant improvement. This proves that compared to multiple sets of queries that simply concatenate visual information from different images, one set of queries can better interact with and extract mixed visual information from multiple images at the feature level. We demonstrated the ablation results of RA-BLIP at the retrieval stage in Fig. 4 (c). The result of the retriever without the multimodal adaptive fusion module exhibits lower performance than that of the complete retriever, which suggests that fusing vision and language in a unified semantic space is essential for multimodal retrieval. By utilizing a retrieval-rank strategy, the retrieval result is significantly improved 20\%, demonstrating the necessity of denoising confusing knowledge through fine-level ranking. The retrieval-rank process employed in our method facilitates precise retrieval among candidate knowledge sources that are primarily relevant but potentially confusing.

D. Sensitivity Analysis

The α is the trade-off hyperparameter of generation loss with ASKG strategy in Eq. (7). We set the range of α from 0.01 to 5. According to Fig. 5, we can see that even for such a large range, the difference between the best and the lowest results is less than 0.04%, indicating our method is robust and insensitive to this parameter.

(1) Q: What is in the man's mouth in L'homme à la Tulipe?

Multimodal Knowledge Base L'homme à la tulipe Lesfemmes(1951), followed by Fanfan la Tulipe (1952), in which she played Madame de Pompadour alongside Gérard Philipe and Gina Lollobrigida. Since then, she has appeared in Italian, French, British and American films. Reference 1 Reference 2

ASKG Output: The relevant reference is Reference 1.

SKURG: In L'homme à la Tulipe, there are flowers in the man's mouth.

RA-BLIP: A cigar is in the man's mouth in L'homme à la Tulipe.

Ground Truth: A cigarette is in the man's mouth in L'homme à la Tulipe.

(2) Q: How many years after Pyramid began airing did Ransom's third season begin airing?



ASKG Output: The relevant reference is Reference 2.

SKURG: It aired 11 years later.

RA-BLIP: It was 45 years after Pyramid began airing that Ransom's third season began airing.

Ground Truth: It aired 45 years after.

Fig. 6. QA Examples. We demonstrate ASKG, RA-BLIP answers, SKURG [20] and ground truth. Relevant knowledge is in green window and irrelevant is in red window. Relevant text in document is underlined.

E. Qualitative Results

Fig. 6 illustrates four examples obtained by RA-BLIP and the baseline SKURG [20]. The retrieved relevant and confusing knowledge is listed, where the green boxes indicate the positive clue and the red boxes represent the negative cue. RA-BLIP outputs the correct answer, while SKURG generates the wrong answer under the identical conditions. This indicates RA-BLIP has more powerful abilities to comprehensively understand the retrieval information, regardless of textual or visual modality. Besides, through ASKG, our model autonomously judges the relevance of retrieved knowledge and selects relevant ones to generate more accurate answers.

V. CONCLUSION

In this paper, we propose a novel multimodal adaptive Retrieval-Augmented BLIP (RA-BLIP), a general retrieval-augmented framework for various classical MLLMs. RA-BLIP utilizes questions as instructions to extract visual features for less irrelevant interference. It incorporates a pre-trained multimodal adaptive fusion module to efficiently integrate information from both visual and textual modalities, thereby achieving question text-to-multimodal retrieval. Additionally, we introduce an adaptive selection knowledge generation strategy to make the generator autonomously discern the relevance of retrieved knowledge. Extensive experiments on multimodal multihop QA and ablation studies verify the effectiveness of

RA-BLIP. In the future, we will explore image-multimodal retrieval and multimodal-multimodal retrieval to realize omnipotent retrieval-augmented models.

9

REFERENCES

- S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra et al., "VQA: visual question answering," in *Proceedings of the IEEE/CVF International* Conference on Computer Vision, 2015, pp. 2425–2433.
- [2] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE/CVF Conference* on Computer Vision and Pattern Recognition, 2017, pp. 6325–6334.
- [3] K. Lee, M. Chang, and K. Toutanova, "Latent retrieval for weakly supervised open domain question answering," in *Proceedings of the* Annual Meeting of the Association for Computational Linguistics, 2019, pp. 6086–6096.
- [4] Y. Chang, G. Cao, M. Narang, J. Gao, H. Suzuki, and Y. Bisk, "Webqa: Multihop and multimodal QA," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16474–16483.
- [5] A. Talmor, O. Yoran, A. Catav, D. Lahav, Y. Wang et al., "Multimodalqa: complex question answering over text, tables and images," in Proceedings of the International Conference on Learning Representations, 2021.
- [6] Y. Li, W. Li, and L. Nie, "Mmcoqa: Conversational question answering over text, tables, and images," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 4220–4231.
- [7] S. Wu, G. Zhao, and X. Qian, "Resolving zero-shot and fact-based visual question answering via enhanced fact retrieval," *IEEE Transactions on Multimedia*, vol. 26, pp. 1790–1800, 2024.
- [8] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay et al., "Scaling instruction-finetuned language models," arXiv preprint arXiv:2210.11416, 2022.
- [9] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux et al., "Llama: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.
- [10] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the International Conference on Machine Learning*, 2023, pp. 19730–19742.
- [11] OpenAI, "GPT-4 technical report," arXiv preprint arXiv:2303.08774, 2023.
- [12] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra et al., "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, pp. 240:1–240:113, 2023.
- [13] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Realm: Retrieval augmented language model pre-training," in *Proceedings of the Interna*tional Conference on Machine Learning, 2020, pp. 3929–3938.
- [14] W. Chen, H. Hu, X. Chen, P. Verga, and W. W. Cohen, "Murag: Multimodal retrieval-augmented generator for open question answering over images and text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 5558–5570.
- [15] Z. Hu, A. Iscen, C. Sun, Z. Wang, K. Chang et al., "Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23369–23379.
- [16] J. Rao, Z. Shan, L. Liu, Y. Zhou, and Y. Yang, "Retrieval-based knowledge augmented vision language pre-training," in *International Conference on Multimedia*, 2023, pp. 5399–5409.
- [17] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 9459–9474.
- [18] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in *Proceedings of the Con*ference of the European Chapter of the Association for Computational Linguistics, 2021, pp. 874–880.
- [19] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford et al., "Improving language models by retrieving from trillions of tokens," in Proceedings of the International Conference on Machine Learning, vol. 162, 2022, pp. 2206–2240.
- [20] Q. Yang, Q. Chen, W. Wang, B. Hu, and M. Zhang, "Enhancing multi-modal multi-hop question answering via structured knowledge and unified retrieval-generation," in *International Conference on Multimedia*, 2023, pp. 5223–5234.

- [21] B. Yu, C. Fu, H. Yu, F. Huang, and Y. Li, "Unified language representation for question answering over text, tables, and images," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 4756–4765.
- [22] F. Liu, J. Liu, Z. Fang, R. Hong, and H. Lu, "Visual question answering with dense inter- and intra-modality interactions," *IEEE Transactions on Multimedia*, vol. 23, pp. 3518–3529, 2021.
- [23] Z. Liu, C. Xiong, Y. Lv, Z. Liu, and G. Yu, "Universal multi-modality retrieval with one unified embedding space," arXiv preprint arXiv:2209.00179, 2022.
- [24] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the International Conference on Machine Learning*, vol. 119, 2020, pp. 1597–1607.
- [25] X. Lu, T. Zhao, and K. Lee, "Visualsparta: An embarrassingly simple approach to large-scale text-to-image search with weighted bag-ofwords," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 5020–5029.
- [26] H. Liu, T. Yu, and P. Li, "Inflate and shrink: Enriching and reducing interactions for fast text-image retrieval," in *Proceedings of the Confer*ence on Empirical Methods in Natural Language Processing, 2021, pp. 9796–9809.
- [27] D. Lin, Y. Ma, Y. Li, X. Song, J. Wu, and L. Nie, "OFAR: A multimodal evidence retrieval framework for illegal live-streaming identification," in Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 3410–3414.
- [28] J. Yu, W. Zhang, Y. Lu, Z. Qin, Y. Hu, J. Tan, and Q. Wu, "Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 22, pp. 3196–3209, 2020.
- [29] J. Zhuang, J. Yu, Y. Ding, X. Qu, and Y. Hu, "Towards fast and accurate image-text retrieval with self-supervised fine-grained alignment," *IEEE Transactions on Multimedia*, vol. 26, pp. 1361–1372, 2024.
- [30] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. C. H. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *Advances in Neural Information Processing Systems*, 2023.
- [31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones et al., "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
- [32] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.
- [33] X. Zheng, Z. Wang, S. Li, K. Xu, T. Zhuang et al., "MAKE: vision-language pre-training based product retrieval in taobao search," in Proceedings of the ACM Web Conference, 2023, pp. 356–360.
- [34] L. Yu, J. Chen, A. Sinha, M. Wang, Y. Chen et al., "Commercemm: Large-scale commerce multimodal representation learning with omni retrieval," in Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 4433–4442.
- [35] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 4904–4916.
- [36] H. Tan and M. Bansal, "LXMERT: learning cross-modality encoder representations from transformers," in *Proceedings of the Empirical Methods in Natural Language Processing*, 2019, pp. 5099–5110.
- [37] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," in Proceedings of the International Conference on Learning Representations, 2022.
- [38] Q. Qi, A. Zhang, Y. Liao, W. Sun, Y. Wang, X. Li, and S. Liu, "Simultaneously training and compressing vision-and-language pretraining model," *IEEE Transactions on Multimedia*, vol. 25, pp. 8194– 8203, 2023.
- [39] Y. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed et al., "UNITER: universal image-text representation learning," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 104–120.
- [40] X. Li, X. Yin, C. Li, P. Zhang, X. Hu et al., "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 121–137.
- [41] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang et al., "Vinvl: Making visual representations matter in vision-language models," arXiv preprint arXiv:2101.00529, 2021.

- [42] M. Tsimpoukelli, J. Menick, S. Cabi, S. M. A. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," in Advances in Neural Information Processing Systems, 2021, pp. 200–212.
- [43] Y. Xing, Q. Wu, D. Cheng, S. Zhang, G. Liang, P. Wang, and Y. Zhang, "Dual modality prompt tuning for vision-language pre-trained model," *IEEE Transactions on Multimedia*, vol. 26, pp. 2056–2068, 2024.
- [44] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers et al., "Lit: Zero-shot transfer with locked-image text tuning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18102–18112.
- [45] J. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr et al., "Flamingo: a visual language model for few-shot learning," in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 23716–23736.
- [46] P. Qi, H. Lee, T. Sido, and C. D. Manning, "Answering open-domain questions of varying reasoning steps from text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3599–3614.
- [47] S. Yavuz, K. Hashimoto, Y. Zhou, N. S. Keskar, and C. Xiong, "Modeling multi-hop question answering as single sequence prediction," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 974–990.
- [48] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations*, 2021.
- [49] R. Guo, P. Sun, E. Lindgren, Q. Geng, D. Simcha et al., "Accelerating large-scale inference with anisotropic vector quantization," in *Proceedings of the International Conference on Machine Learning*, vol. 119, 2020, pp. 3887–3896.
- [50] W. Yuan, G. Neubig, and P. Liu, "Bartscore: Evaluating generated text as text generation," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 27263–27277.
- [51] J. Kaplan, S. McCandlish, and Henighan, "Scaling laws for neural language models," arXiv preprint arXiv:2001.08361, 2020.
- [52] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 13 041–13 049.
- [53] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, pp. 140:1–140:67, 2020.
- [54] P. Wang, A. Yang, R. Men, J. Lin, S. Bai et al., "OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *Proceedings of the International Conference on Machine Learning*, 2022, pp. 23318–23340.
- [55] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed et al., "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7871–7880.
- [56] M. E. Peters, W. Ammar, C. Bhagavatula, and R. Power, "Semi-supervised sequence tagging with bidirectional language models," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 1756–1765.
- [57] X. Han, T. Gao, Y. Yao, D. Ye, Z. Liu, and M. Sun, "Opennre: An open and extensible toolkit for neural relation extraction," in *Proceedings of* the Conference on Empirical Methods in Natural Language Processing, 2019, pp. 169–174.
- [58] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the International Conference on Machine Learning*, 2022, pp. 12888–12900.
- [59] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "EVA: exploring the limits of masked visual representation learning at scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19358–19369.
- [60] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Advances in Neural Infor*mation Processing Systems, 2011, pp. 1143–1151.
- [61] C. Qu, L. Yang, C. Chen, M. Qiu, W. B. Croft, and M. Iyyer, "Openretrieval conversational question answering," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 539–548.
- [62] D. Hannan, A. Jain, and M. Bansal, "Manymodalqa: Modality disambiguation and QA over diverse inputs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 7879–7886.