This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering

Federico Cocchi*[1,2], Nicholas Moratelli*[1], Marcella Cornia[1], Lorenzo Baraldi[1], Rita Cucchiara[1,3]
[1]University of Modena and Reggio Emilia, Italy
[2]University of Pisa, Italy [3]IIT-CNR, Italy
{name.surname}@unimore.it

Abstract

Multimodal LLMs (MLLMs) are the natural extension of large language models to handle multimodal inputs, combining text and image data. They have recently garnered attention due to their capability to address complex tasks involving both modalities. However, their effectiveness is limited to the knowledge acquired during training, which restricts their practical utility. In this work, we introduce a novel method to enhance the adaptability of MLLMs by integrating external knowledge sources. Our proposed model, Reflective LLaVA (ReflectiVA), utilizes reflective tokens to dynamically determine the need for external knowledge and predict the relevance of information retrieved from an external database. Tokens are trained following a two-stage two-model training recipe. This ultimately enables the MLLM to manage external knowledge while preserving fluency and performance on tasks where external knowledge is not needed. Through our experiments, we demonstrate the efficacy of ReflectiVA for knowledge-based visual question answering, highlighting its superior performance compared to existing methods. Source code and trained models are publicly available at https://aimagelab.github.io/ReflectiVA.

1. Introduction

While the past few years have seen a surge of Large Language Models (LLMs) with increasing fluency and reasoning capabilities [6, 57, 66], thanks to the availability of large-scale training data and novel training techniques [14, 31, 54], the Computer Vision community has recently started extending the capabilities of such models beyond pure text, with the inclusion of additional modalities like images, video, and 3D data. The resulting emergence of Multimodal Large Language Models (MLLMs) has been characterized by the development of models targeting multiple tasks [8, 24, 64] – ranging from visual dialogue to image generation –, architectural innovations [2, 40, 72], and novel training recipes [34, 46].

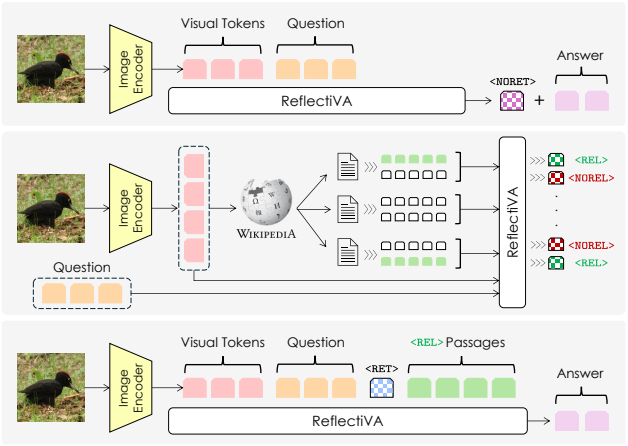


Figure 1. Overview of ReflectiVA, which employs reflective tokens for knowledge-based visual question answering. Our model learns to predict the need of retrieving data from an external knowledge source (top), classifies the relevance of each retrieved item (middle) and generate the final answer based on relevant items (bottom).

What most existing MLLMs share, though, is their exclusive reliance on the knowledge learned at training time – an issue that severely limits their practical applicability to cases that adhere to the training distribution. While this issue is also common to LLMs, it becomes more pressing in the case of MLLMs, where obtaining high-quality and large-scale multimodal data becomes even more difficult. Ideally, indeed, an MLLM should be capable of engaging in dialogues concerning specific visual details, long-tail knowledge, fine-grained categories, and instances [12, 51]. However, this type of knowledge makes it hard for MLLMs to encode in their parameters because such long-tail information occurs rarely in the training data. Additionally, this lack of precise knowledge can also lead the MLLM to generate incorrect answers, thus again limiting their usage in practical cases.

A viable solution to this issue is to rely on a non-parametric approach, where content from external knowledge sources is incorporated into the MLLM context, con-

*Equal contribution.

ditioning the generation on accurate and relevant information [9, 71]. However, building retrieval-augmented MLLMs presents unique challenges, ranging from the need to retrieve appropriate content, to the limited reasoning capabilities of MLLMs compared to LLMs. Further, an MLLM should also be capable of identifying *when* external knowledge is needed, as opposed to answering questions that simply do not need external knowledge, *e.g.* purely visual questions. Lastly, retrieving relevant items from a multimodal knowledge source is an open problem, with state-of-the-art embedding spaces [56] exhibiting limited performance [43, 68]. This further outlines the need to discover *which* retrieved items may be relevant to answering a given user query.

Drawing inspiration from these challenges, in this paper, we propose a multimodal model for knowledge-based visual question answering which can jointly determine the need for accessing external knowledge and the relevance of items retrieved from an external knowledge base. Our model termed Reflective LLaVA (ReflectiVA), employs *reflective tokens* to augment the capabilities of a pre-trained MLLM [46] for knowledge-based generation. In particular, the vocabulary of the model is extended to generate additional tokens with which the model can decide whether retrieval is needed or not (Fig. 1, top), and if a retrieved sample is relevant or not for the input query (Fig. 1, middle). Training is conducted with a two-stage procedure employing two learnable models. We first train an in-article discriminator that can discriminate relevant passages from irrelevant ones found inside the same article. We then employ data annotated synthetically with this model, together with a mixture of existing datasets, to train the final model on all reflective tokens.

Experimentally, we evaluate the performance of the proposed approach on the Encyclopedic-VQA [51] and InfoSeek [12] datasets, which contain question-answer pairs linked with two knowledge bases derived from Wikipedia pages. Additionally, we assess the zero-shot generalization capabilities to two other VQA datasets [29, 36] which may require external knowledge to answer questions correctly. With extensive experiments, we demonstrate that our model outperforms previous works, and provides increased answer accuracy on all considered datasets and settings. Further, we demonstrate that the proposed approach maintains high performance on standard MLLM benchmarks [19, 41, 74] as well as on traditional VQA datasets [26, 61] that do not require external knowledge during generation.

2. Related Work

Multimodal Large Language Models. Large Language Models (LLMs) [18, 30, 66] have shown remarkable performance and adaptability across diverse tasks [13, 65], with recent extensions incorporating multimodal capabilities, particularly integrating vision and language [8, 15, 39, 50, 64]. A primary challenge in this domain is effectively combining LLMs with visual features and creating multimodal datasets for robust training. Architectural solutions for vision-to-language integration vary, including single linear projections or MLPs, as in the LLaVA models [45, 46], as well as Q-Former [16, 40] or Perceiver [35] modules to extract fixed-dimensional visual features. Models like Flamingo [2, 4] utilize cross-attention layers to directly integrate multimodal information. For training, multiple stages often leverage image captions for visual-text alignment [11, 20], and specialized datasets are developed for visual instruction tuning [16, 34, 45], with data quality and annotation specificity shown to significantly influence performance [17].

Knowledge-based Visual Question Answering. This task involves answering questions that require external or specialized knowledge beyond the content of the image itself. Early datasets like OK-VQA [49, 59] and KVQA [60] introduced questions that require general and commonsense knowledge, which large-scale architectures such as MLLMs can increasingly handle within their current training scope. Newer datasets, such as Encyclopedic-VQA [51] and InfoSeek [12], present greater challenges by focusing on highly specific, Wikipedia-scale knowledge. These require understanding detailed information about specific entities and nuanced encyclopedic facts. As a result, MLLMs often struggle in these settings, as they lack comprehensive coverage of this in-depth knowledge without relying on external sources.

To tackle this, contrastive image-text encoders [37, 56, 63, 68, 70] are crucial for retrieving semantically aligned content based on image-question queries. Relevant passages from external knowledge sources are then accessed, with entities represented by both textual passages and images. Recent works [9, 55, 71] have combined these retrieval methods with LLMs and MLLMs to enhance knowledge-based VQA tasks. For example, Wiki-LLaVA [9] integrates external multimodal knowledge via a hierarchical retrieval pipeline within a contrastive embedding space [56]. RoRA-VLM [55], instead, introduces a visual token refinement to filter out query-irrelevant visual information from both retrieved and query images. Recently, inspired by advances in NLP [3, 58, 73], EchoSight [71] proposes a Q-Former based re-ranking module to reorder retrieved textual passages before feeding them into the LLM. In contrast, our approach follows a different path enhancing MLLMs with specialized tokens that help to determine when retrieval is needed and assess the relevance of retrieved external knowledge.

Retrieval-Augmented Language Models. In recent years, retrieval-augmented generation has become increasingly popular in the field of LLMs, especially when training or fine-tuning models for specific tasks or domains is impractical. This has led to various techniques [5, 23, 28, 32, 67] for integrating additional information extracted from external sources to enhance the generation quality of frozen LLMs. Recent efforts in the NLP literature [3] align with our pro-

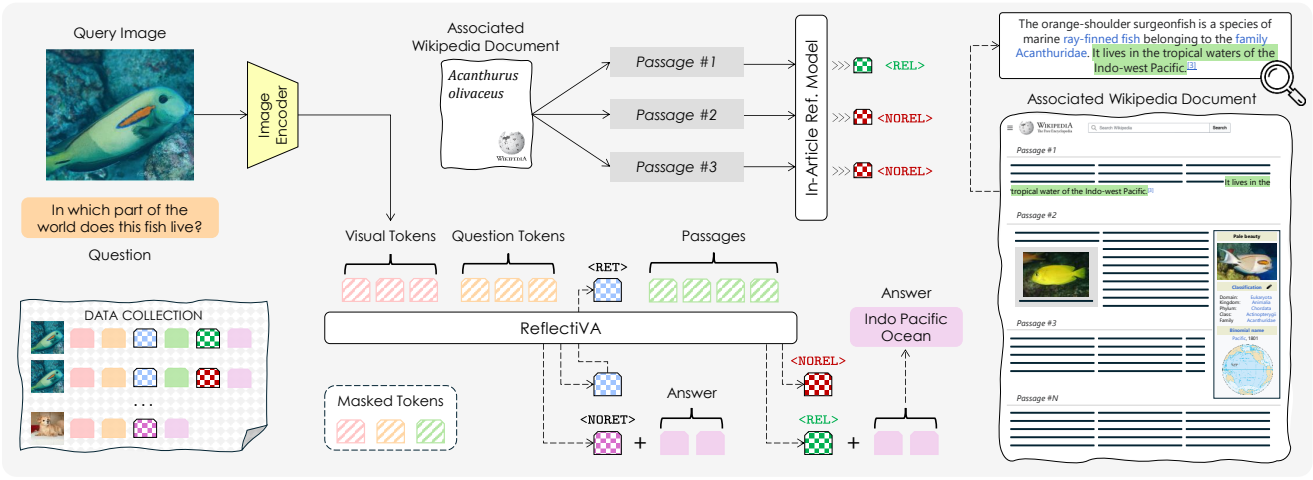


Figure 2. Training approach for ReflectiVA. An in-article model is trained to predict the relevance of passages extracted from the ground-truth document corresponding to an (I, q) pair. The in-article model then generates training data for ReflectiVA, which is trained to predict the need for external knowledge and the relevance of passages, along with the answer, using positive, soft- and hard-negative passages.

posal, introducing special tokens to improve generation and better incorporate retrieved data. To the best of our knowledge, however, we are the first to propose a token-guided retrieval-augmented generation pipeline in MLLMs, where multimodal understanding of both images and text is crucial.

3. Proposed Method

Task Definition. In retrieval-augmented generation, given a textual query q and a query image I, an MLLM is expected to generate an answer y by possibly leveraging additional snippets $\mathbf{S}$ retrieved from an external knowledge source as context. The objective of multimodal retrieval-augmented generation can therefore be written as

$$y = \arg\max_{y} \text{MLLM}(y|I, q, \mathbf{S}). \tag{1}$$

In our setting, the external knowledge source $\mathcal{S}$ is composed of multimodal documents, each endowed with metadata (*e.g.*, title and summary), textual passages organized in sections, and possibly visual content. Formally, the external database can be defined as a collection

$$\mathcal{S} = \{(\tilde{t}_i, \tilde{\mathbf{P}}_i, \tilde{I}_i)\}_{i=1}^{N}, \tag{2}$$

where $\tilde{t}_i$ represents the metadata of a multimodal document, $\tilde{\mathbf{P}}_i$ the set of its textual passages, and $\tilde{I}_i$ its visual content.

Summary of the Approach. To address the limitations of existing retrieval-augmented methods, our approach introduces two innovative strategies. Firstly, we enable the model to determine the optimal timing for retrieval, specifically when generating with an empty retrieval set $\mathbf{S}$ is advantageous because the query does not require external information. Secondly, after the retrieval process, we empower the model to identify the relevance of retrieved passages for generation. Both abilities are enabled through the incorporation of reflective tokens into the vocabulary of the model. These tokens are trained following a two-step two-model training recipe, which ultimately enables the MLLM with the ability to determine whether retrieval is needed and to select pertinent passages from the external database. An overview of our methodology is illustrated in Fig. 2.

3.1. Adding Reflective Tokens

Given a pre-trained MLLM, we augment its vocabulary $\mathcal{V}_0$ with a set of four additional tokens, *i.e.* {<RET>, <NORET>, <REL>, <NOREL>}, which will enable the model to distinguish whether retrieval is needed (<RET>, <NORET>) and whether a retrieved sample is relevant to the input query (<REL>, <NOREL>).

Generation Protocol. At test time, the MLLM is prompted with an input image I and a query q, and is asked to produce either the <RET> or the <NORET> token. If the <NORET> token is sampled, the MLLM will be asked to directly generate the answer y without relying on additional snippets. In this case, the generation process follows a schema

$$\begin{aligned} \texttt{"<NORET>"} &\sim \text{MLLM}\left([I, q], \{\texttt{<RET>}, \texttt{<NORET>}\}, 1\right), \\ y &\sim \text{MLLM}([I, q, \texttt{<NORET>}], \mathcal{V}_0), \end{aligned} \tag{3}$$

where $y \sim \text{MLLM}(p, V, t)$ indicates that a sequence of tokens y is sampled (*e.g.* through beam search) from the MLLM when prompted with an input p and constrained to emit tokens belonging to a vocabulary V and up to a length of t tokens[1]. Finally, $[\cdot, \cdot]$ indicates concatenation.

If instead the <RET> token is sampled, we firstly retrieve a set of candidate textual passages $\mathbf{S}_0 = \{s_0, ..., s_k\}$ from the external knowledge base, and then ask the MLLM to

[1]When t is omitted, we let the model generate until an "end of sequence" token is sampled.

evaluate the relevance of each of them through the emission of <REL> and <NOREL> tokens. In this case, the generation follows a protocol in the form

$$\texttt{"<RET>"} \sim \text{MLLM}([I, q], \{\texttt{<RET>}, \texttt{<NORET>}\}, 1), \quad (4)$$
$$r_i \sim \text{MLLM}([I, q, \texttt{<RET>}, s_i], \{\texttt{<REL>}, \texttt{<NOREL>}\}, 1),$$

where the second generation step is repeated for each item in $\mathbf{S}_0$ and r_i indicates the relevance token sampled for each retrieved snippet.

After sampling relevance tokens, the MLLM is then asked to generate the final answer using the set of snippets that have been judged relevant, *i.e.* $\mathbf{S} = \{s_i \in \mathbf{S}_0, r_i = \texttt{<REL>}\}$. Formally, this generation stage is defined as

$$y \sim \text{MLLM}([I, q, \texttt{<RET>}, \mathbf{S}], \mathcal{V}_0), \quad (5)$$

where, for readability, the concatenation of all items in $\mathbf{S}$ is not made explicit.

Coarse-Grained Retrieval. To obtain the set of candidate passages $\mathbf{S}_0$, we utilize the input query image I as an anchor to retrieve a set of candidate documents from the knowledge source. We encode either the metadata or the image[2] of each document in the database using a CLIP-based textual or visual encoder [56, 63] and build a dense vector-search database. The encoded features $\mathbf{z}_i \in \mathbb{R}^d$, where d is the dimensionality of the CLIP embedding, act as a search index $\mathbf{Z} = \{\mathbf{z}_i\}_i$. The retriever then employs a non-parametric function to compute the cosine similarity between the embedding of the query image and all search indexes. Based on this similarity search, the coarse-grained retriever retrieves the top-k articles that are most similar to the query image. The set of candidate passages $\mathbf{S}_0$ is then built as the union of the passages belonging to the top-k articles.

3.2. Training an In-Article Reflective Model

Clearly, a coarse-grained retrieval at the document level is not enough precise to retrieve the exact passage containing the answer. Indeed, this retrieval step is expected to have limited recall at lower values of k. Further, a more detailed examination of relevant documents is necessary to identify passages that can be utilized by the MLLM.

To this aim, we train the MLLM to emit the relevance tokens <REL> and <NOREL>. This is done with a two-stage, two-model training pipeline. Initially, we train an in-article reflective MLLM capable of distinguishing between relevant passages and negative passages from the same article. Subsequently, we employ predictions from that model to train the final MLLM with the ability to cope with negative passages taken from the same articles and from different articles.

Automatic Data Construction. As most of the datasets for knowledge-based VQA do not provide human-labeled annotations of ground-truth textual passages, we employ a proprietary LLM to automatically annotate positive and negative passages. Given a query image I, a question q and the set of passages from the associated article $\mathbf{P}$, we caption the image I with three captioning models (*i.e.* LLaVA-v1.5 [46], BLIP-2 [40], and InstructBLIP [16]) and prompt the LLM to assess whether a passage $s \in \mathbf{P}$ can answer q given the textual description of I. To help identify positive passages, for each sample, we select the two passages that have the highest similarity with their respective question according to the Contriever embedding space [27]. We also ensure to have at least one positive and one negative passage for each (I, q) sample in the dataset.

Model Training. Having collected positive and negative passages, the in-article reflective model is trained on a mixture of samples associated with positive and negative passages, using sequences in the form

$$[I, q, \texttt{<RET>}, s, \texttt{<REL>}, y] \text{ and}$$
$$[I, q, \texttt{<RET>}, \tilde{s}, \texttt{<NOREL>}, y], \quad (6)$$

where s refers to a passage predicted as positive, and $\tilde{s}$ to a negative passage from the same page. The model is trained using a time-wise cross-entropy loss over the reflective tokens and over the answer.

3.3. Training the Overall Model

In the second stage, predictions from the in-article reflective model are employed to construct the dataset for training the overall model. The capabilities of the in-article model are indeed used to automatically annotate textual passages from existing datasets that require external knowledge. This is also complemented by negative passages taken from other pages of the knowledge base, plus samples that do not need an external knowledge base. The ultimate result of this stage is an MLLM capable of both answering questions and generating special tokens to assess whether additional information retrieval is necessary and whether it would be beneficial for answering.

Data Curation. The training split of the Encyclopedic-VQA [51] and InfoSeek [12] datasets are employed to create the data collection to train the second stage. Each sample (I, q) is expanded with three distinct passages: a positive, a hard negative coming from the ground-truth page, and a soft negative coming from a different page.

To construct the first two cases, each sample is processed by the in-article reflective model to label each passage as either relevant or not relevant. After this, the passage with the highest probability of containing the <REL> token is taken as positive. Instead, the hard negative sample is randomly chosen from one of the sections predicted as <NOREL>. For the soft negative case, the image I is used to retrieve inside the top-1 page, excluding the ground-truth page. From

[2] Depending on the test case, see Sec. 4.1 for details.

the retrieved page, then, a random passage is considered a soft negative. Additionally, data from the LLaVA-Instruct dataset [46] are also included. These samples are labeled as cases where no retrieval is necessary.

Model Training. Finally, the model is trained using a balanced mixture of sequences in the form

$$\begin{gathered}[I, q, \texttt{<NORET>}, y], \\ [I, q, \texttt{<RET>}, s, \texttt{<REL>}, y], \\ [I, q, \texttt{<RET>}, \tilde{s}, \texttt{<NOREL>}, y] \text{ and} \\ [I, q, \texttt{<RET>}, \bar{s}, \texttt{<NOREL>}, y], \end{gathered} \tag{7}$$

where $\bar{s}$ refers to a soft-negative passage. We then employ a time-wise cross-entropy loss over all ground-truth tokens, with the exception of those related to I, q, and retrieved passages $s, \tilde{s}, \bar{s}$.

4. Experiments

4.1. Experimental Setup

Datasets. Our experiments are conducted on Encyclopedic-VQA [51] and InfoSeek [12], which contain question-answer pairs linked to documents from an external knowledge base (*e.g.* Wikipedia). Encyclopedic-VQA consists of 221k pairs, each paired with up to 5 images and associated with 16.7k fine-grained entities (*i.e.* Wikipedia pages). Questions are divided into single-hop and two-hop types: the former indicates that a single Wikipedia page is required to answer them, while the latter requires a sequential retrieval process across multiple documents. Dataset samples are split into training, validation, and test sets with 1M, 13.6k, and 5.8k items respectively. Experiments are reported on the test set, where single-hop questions correspond to 4.8k samples. The InfoSeek dataset, instead, contains 1.3M image-question pairs associated with around 11k Wikipedia pages. The dataset comprises 934k training, 73k validation, and 348k test samples. Following existing literature [9, 71], experimental results are reported on the validation set which includes questions not contained in the training split and questions associated with unseen entities.

External Knowledge Bases. Both datasets come with an external knowledge base composed of Wikipedia documents. In particular, Encyclopedic-VQA contains a knowledge base of 2M Wikipedia pages. Each page includes the Wikipedia title, the corresponding textual sections, and associated images. InfoSeek, instead, provides a knowledge base composed of 6M Wikipedia entities. In our experiments, we use the original 2M knowledge base for Encyclopedic-VQA, while we extract a subset of 100k pages[3] from the original 6M for InfoSeek, following recent works [9, 71].

[3]The knowledge base used for InfoSeek contains the same entities as [9].

Model	Ret. Mode	E-VQA			InfoSeek		
		R@1	R@5	R@20	R@1	R@5	R@20
CLIP ViT-L/14	Textual (T)	3.4	8.7	14.0	36.9	59.9	71.9
CLIP ViT-L/14	Textual (T+S)	0.7	2.3	6.4	19.3	40.9	57.1
CLIP ViT-L/14	Visual	9.9	22.0	31.7	22.5	40.4	44.1
EVA-CLIP-8B	Textual (T)	7.5	15.2	20.7	52.5	71.2	79.2
EVA-CLIP-8B	Textual (T+S)	10.1	20.5	29.4	**56.1**	**77.6**	**86.4**
EVA-CLIP-8B	Visual	**15.6**	**36.1**	**49.8**	29.6	41.4	46.6

Table 1. Retrieval performance on Encyclopedic-VQA test set and the InfoSeek validation set. "Textual" refers to image-to-text retrieval using either title only (T) or title with summary (T+S), and "Visual" corresponds to image-to-image retrieval.

Evaluation Metrics. We follow the evaluation protocol provided along with the datasets. Generated answers for Encyclopedic-VQA are evaluated according to the BERT matching score (BEM) [7] between predicted and ground-truth answers. Instead, when evaluating answers for image-question pairs from InfoSeek, we use VQA accuracy [21] and relaxed accuracy [52] depending on the question type.

Architectural and Training Details. Our model is based on the LLaVA-v1.5 MLLM [15, 46] with LLaMA-3.1-8B [18] as language model[4]. It employs CLIP ViT-L/14@336 as visual encoder and an MLP as the vision-to-language connector. For both training phases, we fine-tune the LLaVA architecture to learn how to generate the introduced reflective tokens. To do that, we modify the original vocabulary of the LLaMA-3.1 LLM, replacing the final four reserved special tokens with our custom tokens. During training, we employ a learning rate of 2×10^{-5} and a global batch size of 128, updating the weights of both the MLP and LLM. The training is stopped based on the loss value on a separate validation split extracted from the training dataset.

Training Data Collection. We train our model on a collection of data from different sources. For both training phases, we include samples from the Encyclopedic-VQA and InfoSeek training splits, as well as data from LLaVA-Instruct [46] to retain the generative capabilities of the MLLM. To automatically annotate positive and negative passages used to train the in-article reflective model, we employ GPT-4, prompted with the query question, the textual description of the associated image, the corresponding answer, and the passage to annotate. We also employ few-shot examples in the prompt to ease the task. Additional details on the training data mixture are reported in the supplementary material.

Coarse-Grained Retrieval Details. To identify the set of documents most relevant to the query image and associated question, we evaluate two CLIP-based retrieval models, namely CLIP ViT-L/14@336 [56] and EVA-CLIP-8B [63]. We explore two retrieval configurations for both models: (i) image-to-text retrieval, which computes similarity between

[4]https://huggingface.co/aimagelab/LLaVA_MORE-llama_3_1-8B-finetuning

Model	LLM	Retrieval Mode		E-VQA		InfoSeek		
				Single-Hop	All	Unseen-Q	Unseen-E	All
Zero-shot LLMs								
Vanilla	Vicuna-7B	-		2.1	2.0	0.3	0.0	0.0
Vanilla	LLaMA-3-8B	-		16.3	17.3	1.5	0.0	0.0
Vanilla	LLaMA-3.1-8B	-		16.5	16.6	2.1	0.0	0.0
Vanilla	GPT-4	-		21.9	23.4	7.3	5.0	5.9
Zero-shot MLLMs								
BLIP-2 [40]	Flan-T5$_{XL}$	-		12.6	12.4	12.7	12.3	12.5
InstructBLIP [16]	Flan-T5$_{XL}$	-		11.9	12.0	8.9	7.4	8.1
LLaVA-v1.5 [46]	Vicuna-7B	-		16.3	16.9	9.6	9.4	9.5
LLaVA-v1.5 [46]	LLaMA-3.1-8B	-		16.0	16.9	8.3	8.9	7.8
GPT-4V [1]	-	-		26.9	28.1	15.0	14.3	14.6
Retrieval-Augmented Models								
DPR$_{V+T}$ [37]†	Multi-passage BERT	CLIP ViT-B/32	Visual+Textual	29.1	-	-	-	12.4
RORA-VLM [55]†	Vicuna-7B	CLIP+Google Search	Visual+Textual	-	20.3	25.1	27.3	-
Wiki-LLaVA [9]	Vicuna-7B	CLIP ViT-L/14+Contriever	Textual	17.7	20.3	30.1	27.8	28.9
Wiki-LLaVA [9]♢	LLaMA-3.1-8B	CLIP ViT-L/14+Contriever	Textual	18.3	19.6	28.6	25.7	27.1
EchoSight [71]†	Mistral-7B/LLaMA-3-8B	EVA-CLIP-8B	Visual	19.4	-	-	-	27.7
EchoSight [71]♢	LLaMA-3.1-8B	EVA-CLIP-8B	Textual	22.4	21.7	30.0	30.7	30.4
EchoSight [71]♢	LLaMA-3.1-8B	EVA-CLIP-8B	Visual	26.4	24.9	18.0	19.8	18.8
ReflectiVA (Ours)	LLaMA-3.1-8B	CLIP ViT-L/14	Textual	24.9	26.7	34.5	32.9	33.7
ReflectiVA (Ours)	LLaMA-3.1-8B	EVA-CLIP-8B	Textual	28.0	29.2	**40.4**	**39.8**	**40.1**
ReflectiVA (Ours)	LLaMA-3.1-8B	EVA-CLIP-8B	Visual	**35.5**	**35.5**	28.6	28.1	28.3

Table 2. VQA accuracy scores on the Encyclopedic-VQA test set and the InfoSeek validation set, where all results from retrieval-augmented models are reported without considering any re-ranking stage to reorder retrieved documents. † indicates results that are not directly comparable due to different knowledge bases, and the marker ♢ represents our reproductions with different LLMs.

the query image and document metadata (either the title alone or the title with the summary of the page), and (ii) image-to-image retrieval, which assesses similarity between the query image and images within Wikipedia documents.

Retrieval results for each variant are detailed in Table 1. Notably, EVA-CLIP demonstrates superior results across all configurations. However, the optimal retrieval configuration varies between datasets. Specifically, image-to-image retrieval yields the highest accuracy on Encyclopedic-VQA, while image-to-text retrieval proves the most effective for InfoSeek. This is probably due to the distinct characteristics and structural composition of each dataset and their respective knowledge bases. Consequently, unless specified otherwise, we adopt image-to-image retrieval for Encyclopedic-VQA and image-to-text retrieval (using title-only for CLIP ViT-L and title with summary for EVA-CLIP) for InfoSeek, with the number k of retrieved documents equal to 5.

4.2. Comparison with the State of the Art

Results on Encyclopedic-VQA and InfoSeek. We evaluate our model on the aforementioned datasets, comparing it to various zero-shot LLMs, MLLMs, and retrieval-augmented competitors. Specifically, we report results of four LLMs – Vicuna [13], LLaMA-3, LLaMA-3.1 [18] and GPT-4 [1] – each prompted with both the query question and a description of the query image generated by an image captioning model. Additionally, we assess the performance of BLIP-2 [40], InstructBLIP [16], LLaVA-v1.5 [46] and GPT-4V [1] without external retrieval augmentation and using only the query image and question as input. As direct competitors, we include DPR [37], RORA-VLM [55], Wiki-LLaVA [9], and EchoSight [71], which all leverage external knowledge retrieval. To ensure a fair comparison with the considered methods, for both Wiki-LLaVA and EchoSight we develop a variant based on LLaMA-3.1, employing the same knowledge bases used in our solution.

Model	LLM	E-VQA	
		Single-Hop	All
Textual Retrieval Mode			
EchoSight [71]♢	LLaMA-3.1-8B	26.8	26.0
ReflectiVA (Ours)	LLaMA-3.1-8B	**33.6**	**33.9**
Visual Retrieval Mode			
EchoSight [71]†	Mistral-7B/LLaMA-3-8B	**41.8**	-
EchoSight [71]♢	LLaMA-3.1-8B	36.3	34.2
ReflectiVA (Ours)	LLaMA-3.1-8B	40.6	**39.7**

Table 3. VQA accuracy scores on Encyclopedic-VQA when models are equipped with a document re-ranking component. † indicates results that are not directly comparable due to different knowledge bases, and the marker ♢ represents our reproductions.

Results are shown in Table 2, in which we report for each retrieval-augmented model the details of the retrieval pipeline used (*i.e.* the retrieval model and modality). As it can be seen, both zero-shot LLMs and MLLMs fail to correctly answer the given questions due to the lack of external knowledge during the generation. This is particularly evident on the InfoSeek dataset where LLMs exhibit accuracy scores close to zero, highlighting the need for the

Model	LLM	E-VQA	InfoSeek		
		Single-Hop	Un-Q	Un-E	All
KB Article					
Vanilla	Vicuna-7B	34.1	5.3	4.3	4.7
Vanilla	LLaMA-3-8B	72.9	10.0	7.9	8.8
Vanilla	LLaMA-3.1-8B	73.6	15.2	13.9	14.5
LLaVA-v1.5 [46]	Vicuna-7B	42.9	14.2	13.4	13.8
LLaVA-v1.5 [46]	LLaMA-3.1-8B	54.1	20.1	17.7	18.8
KB Passages					
Wiki-LLaVA [9]	Vicuna-7B	38.5	52.7	50.3	51.5
Wiki-LLaVA [9]$^{\diamondsuit}$	LLaMA-3.1-8B	46.8	51.2	50.6	50.9
ReflectiVA (Ours)	LLaMA-3.1-8B	**75.2**	**57.8**	**57.4**	**57.6**

Table 4. VQA accuracy scores on Encyclopedic-VQA and InfoSeek with oracle Wikipedia entities. The top of the table shows results using the full Wikipedia article as input to the LLM/MLLM, while the bottom of the table shows performance using model-specific strategies to identify relevant text passages.

visual inputs to generate correct answers for this dataset. For knowledge-based models, the proposed ReflectiVA exhibits state-of-the-art results on both the Encyclopedic-VQA and InfoSeek datasets, outperforming all evaluated competitors by a substantial margin and highlighting the benefits of employing reflective tokens for the task.

Integrating a Re-Ranking Stage. Recent studies focusing on text-only LLMs [58, 71, 73] have shown that incorporating a re-ranking stage within a retrieval-augmented generation pipeline can enhance performance. Following this approach, we evaluate our model with the re-ranking component proposed in [71] which reorders retrieved textual passages prior to model input. Specifically, in our experiments, we first retrieve the top-k relevant Wikipedia pages from the knowledge base using the retrieval model previously described. The retrieved passages are then processed by the re-ranking component, and we input the top-k_p re-ranked passages into our model to assess their relevance to the query and generate a response[5].

Results are shown in Table 3 comparing our results to those of EchoSight on the Encyclopedic-VQA dataset[6]. As it can be seen, performing a re-ranking step can further improve the results of our model which achieves, in its best configuration, 40.6 accuracy points compared to 35.5 without re-ranking. These results consistently outperform those obtained by EchoSigh, with the same LLM and knowledge base used in our setting, in both retrieval configurations.

Results using Oracle Documents. To thoroughly evaluate the performance of our model, we conduct experiments under an oracle setting, where the ground-truth entity (*i.e.* the Wikipedia page associated with the query) is provided. In this configuration, all text passages from the oracle entity are input to ReflectiVA, which then selects the relevant passages before generating an answer. Table 4 presents the results of this analysis, directly comparing ReflectiVA with Wiki-LLaVA, which leverages a Contriever model [27] to retrieve the most relevant passages within the oracle document. We also report the performance of standard LLMs and MLLMs when prompted with the entire Wikipedia article. Notably, ReflectiVA achieves the highest performance across both the Encyclopedic-VQA and InfoSeek benchmarks, surpassing Wiki-LLaVA and standard models, further highlighting its effectiveness in isolating the passages most pertinent to the given image-question pair. It is also noteworthy that while vanilla LLMs achieve high accuracy on Encyclopedic-VQA when prompted with the entire oracle Wikipedia page, on the InfoSeek having a selection strategy of the most relevant passages lead to significantly better performance, with ReflectiVA always reaching the best results.

[5]A detailed analysis of k and k_p in re-ranking settings is provided in the supplementary material.

[6]For this setting, we do not include results on InfoSeek since the re-ranker proposed in [71] is trained on samples from Encyclopedic-VQA.

	k	E-VQA	InfoSeek		
		Single-Hop	Un-Q	Un-E	All
Effectiveness of Reflective Tokens and Training					
In-Article Reflective Model	5	21.1	25.5	23.8	24.6
Single model (w/ special tokens from LLaMA-3.1)	5	30.7	27.3	28.2	28.5
ReflectiVA (Overall Model)	5	**35.5**	**40.4**	**39.8**	**40.1**
always w/ <RET> token	5	35.3	40.2	**39.8**	40.0
w/ Contriever passages (w/o <REL>/<NOREL>)	5	29.3	30.8	29.1	29.9
w/o <REL>/<NOREL> tokens	5	23.6	32.2	30.6	31.4
w/o KB (always with <NORET> token)	-	21.3	17.7	15.3	16.4
Varying the Number of Retrieved Documents					
	1	29.0	**40.6**	**41.0**	**40.8**
ReflectiVA (Overall Model)	5	35.5	40.4	39.8	40.1
	10	**36.0**	37.2	37.0	37.1
	20	35.7	30.6	31.3	30.9

Table 5. Ablation study results demonstrating the effectiveness of the proposed reflective tokens and training strategy, along with the impact of different numbers of retrieved documents.

Qualitative Results. Fig. 3 provides a qualitative comparison on sample image-question pairs from Encyclopedic-VQA (top row) and InfoSeek (bottom row).

4.3. Ablation Studies and Analyses

Effectiveness of Two-Stage Training. We analyze the impact of our two-stage, two-model training strategy. Specifically, we assess the effectiveness of the in-article reflective model, which is trained using both positive and negative passages drawn from the same Wikipedia pages, on Encyclopedic-VQA and InfoSeek. Table 5 (top) demonstrates that training only the first model alone does not yield high accuracy scores. Likewise, training a single model on full-dataset annotations generated by open-source LLMs such as LLaMA-3.1 fails to achieve competitive performance. In constrast, the complete model (*i.e.* ReflectiVA) consistently achieves higher accuracy.

Effectiveness of Reflective Tokens. To evaluate the effectiveness of the proposed reflective tokens, we conduct ablation studies that isolate the contributions of both retrieval and relevance tokens. We design four variants of the infer-

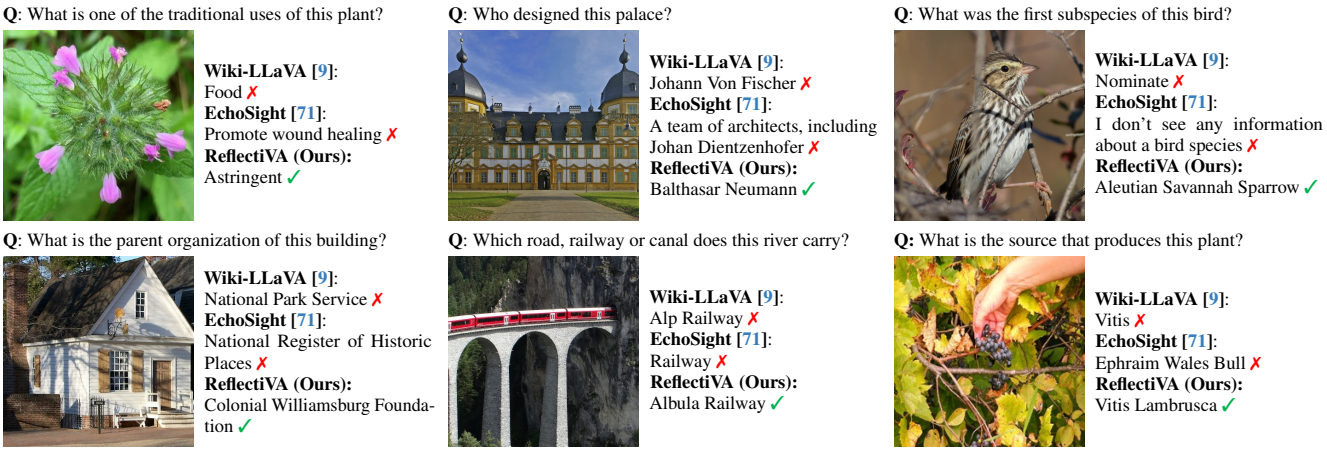


Figure 3. Sample qualitative results on image-question pairs from Encyclopedic-VQA (top row) and InfoSeek (bottom row), where we compare the answers provided by ReflectiVA with those from WikiLLaVA [9] and EchoSight [71].

Model	LLM	MMMU	MMB (EN)	POPE	SEED-Img	MME (P)	MME (C)	GQA	TextVQA	Science-QA	AI2D
LLaVA-v1.5 [46]	Vicuna-7B	34.2	65.3	85.6	66.8	1474.3	314.6	62.4	58.2	69.0	56.4
LLaVA-v1.5 [46]	LLaMA-3.1-8B	39.4	72.4	85.1	69.8	1531.5	353.3	63.6	58.4	76.3	61.8
Wiki-LLaVA (E-VQA) [9]	Vicuna-7B	36.6	70.4	86.6	-	1170.1	290.0	-	-	-	-
Wiki-LLaVA (InfoSeek) [9]	Vicuna-7B	35.6	71.1	84.2	-	1438.9	341.3	-	-	-	-
Wiki-LLaVA (E-VQA) [9]$^{\diamond}$	LLaMA-3.1-8B	32.2	60.9	84.6	59.2	1350.7	306.8	56.6	49.1	67.5	55.1
Wiki-LLaVA (InfoSeek) [9]$^{\diamond}$	LLaMA-3.1-8B	35.9	52.0	85.7	60.5	1417.8	349.6	58.6	50.1	69.1	54.3
ReflectiVA (Ours)	LLaMA-3.1-8B	38.9	69.9	85.1	68.6	1564.5	355.7	62.1	56.8	75.4	60.6

Table 6. Performance preservation analysis on standard benchmarks for MLLM evaluation and traditional VQA datasets.

ence pipeline: (i) always enforcing the <RET> token while using the standard <REL>/<NOREL> pipeline, (ii) without using relevance tokens, instead selecting the top passages retrieved by the Contriever [27] model, (iii) without assessing the relevance of retrieved passages, where two random passages are selected from each of the top-5 documents, and (iv) without retrieval from the knowledge base (*i.e.* always enforcing <NORET> during generation). Results are summarized in the top part of Table 5 for Encyclopedic-VQA and InfoSeek. As shown, consistently performing retrieval causes only a minor performance degradation, given that both datasets are designed to require external knowledge. In contrast, omitting the proposed relevance tokens leads to a substantial accuracy drop (*e.g.* from 35.5 to 23.6 on Encyclopedic-VQA and from 40.1 to 31.4 on InfoSeek), underscoring the critical role of identifying relevant passages before answer generation. Furthermore, the lowest scores occur when retrieval from the external knowledge base is entirely bypassed, emphasizing the need for retrieval. Also, replacing the <REL> and <NOREL> tokens with the top-2 passages retrieved by Contriever results in significantly lower performance on both datasets, underscoring the benefit of explicit relevance modeling over direct retrieval.

Varying the Number of Retrieved Documents. In Table 5 (bottom), we further analyze the effect of varying the number k of retrieved documents. While setting k equal to 10 achieves optimal results on Encyclopedic-VQA, using only the top-1 retrieved document generally performs better for InfoSeek. This discrepancy can be attributed to differences in the knowledge base sizes (*i.e.* significantly larger for Encyclopedic-VQA) and retrieval performance across the two datasets (cf. Table 1). Overall, using the top-5 retrieved documents provides the best trade-off across both datasets, leading us to set k equal to 5 in our experiments.

Performance Preservation on Standard Benchmarks. Finally, we evaluate whether the proposed approach impacts performance on standard MLLM and VQA benchmarks that do not require external knowledge. In Table 6, we compare ReflectiVA with Wiki-LLaVA, which fine-tunes a LLaVA model for similar purposes, and include results from the original LLaVA-v1.5 model, tested with both Vicuna-7B and LLaMA-3.1. ReflectiVA incurs only a minor performance reduction relative to the original LLaVA model, while significantly outperforming Wiki-LLaVA.

5. Conclusion

We proposed ReflectiVA, a multimodal LLM with retrieval-augmented generation. Our method employs reflective tokens, trained in a two-stage two-model pipeline. Extensive experiments, conducted on both VQA datasets requiring external knowledge and standard datasets, demonstrate the efficacy of the proposed solution.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This work has been conducted under a research grant co-funded by Altilia s.r.l., and supported by the PNRRM4C2 project "FAIR - Future Artificial Intelligence Research", funded by the European Commission, by the PNRR project "Italian Strengthening of Esfri RI Resilience" (ITSERR), funded by the European Union - NextGenerationEU (CUP B53C22001770006), and by the "Fortissimo Plus" (FFplus) project, funded by the European High-Performance Computing Joint Undertaking (No. 101163317). We would also like to thank Davide Caffagni and Sara Sarto for their valuable support.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6, 15

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a Visual Language Model for Few-Shot Learning. In *NeurIPS*, 2022. 1, 2

[3] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *ICLR*, 2024. 2, 15

[4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv preprint arXiv:2308.01390*, 2023. 2

[5] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving Language Models by Retrieving from Trillions of Tokens. In *ICML*, 2022. 2

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 1

[7] Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Boerschinger, and Tal Schuster. Tomayto, Tomahto. Beyond Token-level Answer Equivalence for Question Answering Evaluation. *arXiv preprint arXiv:2202.07654*, 2022. 5

[8] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The Revolution of Multimodal Large Language Models: A Survey. In *ACL Findings*, 2024. 1, 2

[9] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs. In *CVPR Workshops*, 2024. 2, 5, 6, 7, 8, 12, 14, 16, 17

[10] David Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. CLAIR: Evaluating Image Captions with Large Language Models. In *EMNLP*, 2023. 13

[11] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In *CVPR*, 2021. 2

[12] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? In *EMNLP*, 2023. 1, 2, 4, 5, 12

[13] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, 2023. 2, 6

[14] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling Instruction-Finetuned Language Models. *JMLR*, 25(70):1–53, 2024. 1

[15] Federico Cocchi, Nicholas Moratelli, Davide Caffagni, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. LLaVA-MORE: A Comparative Study of LLMs and Visual Backbones for Enhanced Visual Instruction Tuning. *arXiv preprint arXiv:2503.15621*, 2025. 2, 5

[16] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*, 2023. 2, 4, 6

[17] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models. *arXiv preprint arXiv:2409.17146*, 2024. 2

[18] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 5, 6

[19] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*, 2023. 2, 13

[20] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets. In *NeurIPS*, 2023. 2

[21] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA Matter: Elevating

the Role of Image Understanding in Visual Question Answering. In *CVPR*, 2017. 5

[22] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. KAT: A Knowledge Augmented Transformer for Vision-and-Language. In *NAACL*, 2022. 12

[23] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval Augmented Language Model Pre-Training. In *ICML*, 2020. 2

[24] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. OneLLM: One Framework to Align All Modalities with Language. In *CVPR*, 2024. 1

[25] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. REVEAL: Retrieval-Augmented Visual-Language Pre-Training With Multi-Source Multimodal Knowledge Memory. In *CVPR*, 2023. 12

[26] Drew A Hudson and Christopher D Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *CVPR*, 2019. 2, 13

[27] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised Dense Information Retrieval with Contrastive Learning. *arXiv preprint arXiv:2112.09118*, 2021. 4, 7, 8, 14

[28] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *JMLR*, 24(251):1–43, 2023. 2

[29] Aman Jain, Mayank Kothyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. Select, Substitute, Search: A New Benchmark for Knowledge-Augmented Visual Question Answering. In *SIGIR*, 2021. 2, 13, 14

[30] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. 2

[31] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of Experts. *arXiv preprint arXiv:2401.04088*, 2024. 1

[32] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active Retrieval Augmented Generation. In *EMNLP*, 2023. 2

[33] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram is Worth a Dozen Images. In *ECCV*, 2016. 13

[34] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In *NeurIPS*, 2023. 1, 2

[35] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 2

[36] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. ViQuAE, a dataset for knowledge-based visual question answering about named entities. In *SIGIR*, 2022. 2, 13, 14

[37] Paul Lerner, Olivier Ferret, and Camille Guinaudeau. Cross-modal Retrieval for Knowledge-based Visual Question Answering. In *ECIR*, 2024. 2, 6

[38] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 13

[39] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2

[40] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, 2023. 1, 2, 4, 6, 13, 14, 15

[41] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *EMNLP*, 2023. 2, 13

[42] Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained Late-interaction Multi-modal Retrieval for Retrieval Augmented Visual Question Answering. In *NeurIPS*, 2023. 12

[43] Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. PreFLMR: Scaling Up Fine-Grained Late-Interaction Multimodal Retrievers. In *ACL*, 2024. 2

[44] Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. REVIVE: Regional Visual Representation Matters in Knowledge-Based Visual Question Answering. In *NeurIPS*, 2022. 12

[45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *NeurIPS*, 2023. 2

[46] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *CVPR*, 2024. 1, 2, 4, 5, 6, 7, 8, 14

[47] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:2307.06281*, 2023. 13

[48] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *NeurIPS*, 2022. 13

[49] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *CVPR*, 2019. 2, 12

[50] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training. In *ECCV*, 2024. 2

[51] Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic VQA: Visual Questions About Detailed Properties of Fine-Grained Categories. In *ICCV*, 2023. 1, 2, 4, 5, 12

[52] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. PlotQA: Reasoning over Scientific Plots. In *WACV*, 2020. 5

[53] Nicholas Moratelli, Davide Caffagni, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Revisiting Image Captioning Training Paradigm via Direct CLIP-based Optimization. In *BMVC*, 2024. 13

[54] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training Language Models to Follow Instructions with Human Feedback. In *NeurIPS*, 2022. 1

[55] Jingyuan Qi, Zhiyang Xu, Rulin Shao, Yang Chen, Jing Di, Yu Cheng, Qifan Wang, and Lifu Huang. RoRA-VLM: Robust Retrieval-Augmented Vision Language Models. *arXiv preprint arXiv:2410.08876*, 2024. 2, 6, 12

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, 2021. 2, 4, 5

[57] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*, 21(1):5485–5551, 2020. 1

[58] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-Context Retrieval-Augmented Language Models. *TACL*, 11: 1316–1331, 2023. 2, 7

[59] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge. In *ECCV*, 2022. 2, 12

[60] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. KVQA: Knowledge-aware Visual Question Answering. In *AAAI*, 2019. 2

[61] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA Models That Can Read. In *CVPR*, 2019. 2, 13

[62] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. In *ACM SIGIR*, 2021. 12

[63] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. EVA-CLIP-18B: Scaling CLIP to 18 Billion Parameters. *arXiv preprint arXiv:2402.04252*, 2024. 2, 4, 5

[64] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative Pretraining in Multimodality. In *ICLR*, 2024. 1, 2

[65] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford Alpaca: An Instruction-Following LLaMA Model, 2023. 2

[66] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 2

[67] Boxin Wang, Wei Ping, Lawrence McAfee, Peng Xu, Bo Li, Mohammad Shoeybi, and Bryan Catanzaro. InstructRetro: Instruction Tuning post Retrieval-Augmented Pretraining. In *ICML*, 2024. 2

[68] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. UniIR: Training and Benchmarking Universal Multimodal Information Retrievers. In *ECCV*, 2024. 2

[69] Tsung-Han Wu, Joseph E Gonzalez, Trevor Darrell, and David M Chan. CLAIR-A: Leveraging Large Language Models to Judge Audio Captions. *arXiv preprint arXiv:2409.12962*, 2024. 13

[70] Zilin Xiao, Ming Gong, Paola Cascante-Bonilla, Xingyao Zhang, Jie Wu, and Vicente Ordonez. Grounding Language Models for Visual Entity Recognition. In *ECCV*, 2024. 2

[71] Yibin Yan and Weidi Xie. EchoSight: Advancing Visual-Language Models with Wiki Knowledge. *arXiv preprint arXiv:2407.12735*, 2024. 2, 5, 6, 7, 8, 12, 14, 15, 16, 17

[72] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. In *CVPR*, 2024. 1

[73] Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs. *arXiv preprint arXiv:2407.02485*, 2024. 2, 7

[74] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *CVPR*, 2024. 2, 13