

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

Supplementary Material

In the following, we provide additional details and analyses for the proposed Reflective LLaVA (ReflectiVA) model. First, we outline key distinctions between ReflectiVA and existing approaches, focusing on architectural design and the knowledge bases used. Next, we present a comprehensive overview of the implementation details, training setup, and the data mixture employed in our two-stage, two-model training strategy. Finally, we supplement the main paper with extended experimental evaluations and qualitative results, further validating the effectiveness of our approach.

A. Detailed Comparison with Existing Methods

Positioning. Knowledge-based VQA has been widely studied by the Computer Vision community, particularly with the emergence of retrieval-augmented architectures [22, 25, 42, 44] tailored for small-scale datasets such as OK-VQA [49, 59]. More recently, the introduction of larger-scale and challenging datasets like Encyclopedic-VQA [51] and InfoSeek [12], alongside advances in LLMs, has shifted the focus towards developing retrieval-augmented solutions leveraging LLMs and MLLMs. In this context, addressing knowledge-based VQA at a Wikipedia-scale remains a relatively unexplored challenge. To the best of our knowledge, only a few methods have attempted to tackle this task effectively. Among these, Wiki-LLaVA [9], RORA-VLM [55], and EchoSight [71] are the most comparable to our approach. Table 7 presents a detailed comparison of ReflectiVA with these state-of-the-art LLM-based methods specifically designed for this task.

Architectural Design. As detailed in the main paper, both RORA-VLM and Wiki-LLaVA are built upon the LLaVA-v1.5 MLLM, using Vicuna-7B as LLM. In contrast,

EchoSight employs different LLMs depending on the target dataset (*i.e.* Mistral-7B for Encyclopedic-VQA and LLaMA-3-8B for InfoSeek). Our proposed ReflectiVA model, instead, is based on the LLaVA-v1.5 model with LLaMA-3.1-8B as the underlying LLM. To ensure a fair comparison and eliminate any advantage from using a more advanced LLM, we additionally implement reproductions of Wiki-LLaVA and EchoSight using LLaMA-3.1-8B.

Regarding the training phase, EchoSight integrates a re-ranking component to reorder retrieved passages, which is specifically trained on Encyclopedic-VQA while keeping the LLM frozen. In contrast, the other competitors are fine-tuned on the considered datasets. Furthermore, unlike RORA-VLM and Wiki-LLaVA, which train separate models for Encyclopedic-VQA and InfoSeek, ReflectiVA is fine-tuned jointly on both datasets, enabling seamless applicability across both settings.

Knowledge Bases. Each considered dataset is paired with its own knowledge base. Specifically, Encyclopedic-VQA is associated with a controlled knowledge base comprising 2 million English articles, derived from the WIT dataset [62]. Similarly, InfoSeek utilizes a knowledge base of Wikipedia pages, initially consisting of 100k items (as reported in the original paper [12]) and later expanded to 6 million entities⁷. With the exception of RORA-VLM, which uses its own knowledge base directly extracted from WIT, all other approaches adhere to the original evaluation protocols. Specifically, experiments on Encyclopedic-VQA are conducted with the original knowledge base of 2 million entities. For InfoSeek, following [12], evaluations are conducted using a subset of 100k Wikipedia pages. While both Wiki-LLaVA

⁷<https://github.com/open-vision-language/infoseek>

	Wiki-LLaVA [9]	RORA-VLM [55]	EchoSight [71]	Wiki-LLaVA [9]◊	EchoSight [71]◊	ReflectiVA (Ours)
<i>Architectural Design</i>						
LLM	Vicuna-7B	Vicuna-7B	Mistral-7B/LLaMA-3-8B	LLaMA-3.1-8B	LLaMA-3.1-8B	LLaMA-3.1-8B
Underlying MLLM	LLaVA-v1.5	LLaVA-v1.5	-	LLaVA-v1.5	-	LLaVA-v1.5
LLM Fine-tuning	✓	✓	✗	✓	✗	✓
Inherently Multimodal	✓	✓	✗	✓	✗	✓
Dataset Independent	✗	✗	✗	✗	✓	✓
<i>Encyclopedic-VQA</i>						
# KB Items	2M	N/A	2M	2M	2M	2M
From the Original KB?	✓	✗ (WIT KB)	✓	✓	✓	✓
<i>InfoSeek</i>						
# KB Items	100k	N/A	100k	100k	100k	100k
From the Original KB?	✓	✗ (WIT KB)	✗ (E-VQA KB)	✓	✓	✓

Table 7. Comparison with existing methods based on LLMs in terms of architectural design and knowledge bases used for the Encyclopedic-VQA and InfoSeek datasets. The marker ◊ represents our reproductions.

	First Stage		Second Stage	
	# Samples	Passages	# Samples	Passages
E-VQA	43.6k	In-Article	2.9M	In- and Cross-Article
InfoSeek	41.0k	In-Article	2.5M	In- and Cross-Article
LLaVA-Instruct	665.3k	-	665.3k	-

Table 8. Training data mixture employed during the two phases of the proposed training strategy.

and ReflectiVA extract subsets directly from the 6M knowledge base provided with the InfoSeek dataset, EchoSight uses 100k entities extracted from the Encyclopedic-VQA knowledge base. To ensure a fair comparison, our reproductions of Wiki-LLaVA and EchoSight are tested using the same knowledge bases employed in our approach.

B. Additional Experimental Details

B.1. Datasets

Training Data Mixture. In Table 8, we summarize the training datasets employed in both stages of the proposed training strategy. At each stage, the samples from each data source are balanced to maintain consistent representativeness. The annotations for relevant and non-relevant passages will be made publicly available.

Additional Knowledge-based VQA Datasets. In addition to the datasets employed in the main paper, we perform zero-shot experiments on two additional knowledge-based VQA datasets, namely ViQuAE [36] and S3VQA [29]. Specifically, we report experiments on the ViQuAE test set, composed of 3,317 image-questions pairs, and on the S3VQA validation set, that contains 750 samples. To facilitate the experiments, for both datasets, we employ the 2M knowledge base of Encyclopedic-VQA, performing image-to-image retrieval with EVA-CLIP as the retrieval model. To compute the results, for ViQuAE samples we follow the official evaluation protocol which includes F1 and exact match scores. Instead, given the high level of difficulty of the questions contained in S3VQA and the absence of an official evaluation protocol, we evaluate generated answers using GPT-4. In detail, following recent literature [10, 53, 69], we prompt GPT-4 with the question and the image description extracted with BLIP-2 [40]⁸, and ask the model to evaluate the alignment between the predicted and ground-truth answers with a score from 0 to 100, where 0 indicates no alignment and 100 indicates a perfect alignment between the two answers⁹.

⁸<https://huggingface.co/Salesforce/blip2-flan-t5-xl>

⁹Specifically, the prompt we used in our evaluation is:

You are trying to evaluate the alignment between a predicted answer and a ground-truth answer for a given question-image pair. To do this, consider the context provided by the question itself and the caption of the query image.

```
# Question: {question}
# Image Caption: {caption}
# Ground-truth Answer: {ground-truth answer}
```

Benchmarks for MLLM Evaluation. Table 6 of the main paper presents the performance of ReflectiVA on a suite of standard benchmarks commonly used to evaluate MLLMs. The evaluation includes results on MMMU [74], the English subset of MMBench (MMB) [47], POPE [41], the image-specific subset of SEED-Bench (SEED-Img) [38], and MME [19]. These benchmarks comprehensively assess MLLM capabilities across task-oriented and instruction-following scenarios. Additionally, we report results on established VQA datasets, including GQA [26], TextVQA [61], ScienceQA [48], and AI2D [33]. All evaluations were performed using a dedicated library tailored for benchmarking large multimodal models¹⁰.

B.2. Additional Implementation Details

In our experiments, we adopt the prompt formats used by LLaMA-3.1 and utilize three of the four supported roles: system, user, and assistant. The instructions provided are consistent with the style of the LLaVA-v1.5 MLLM. Specifically, for each image-question pair, the prompt we employ is as follows:

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
  You are a helpful language and vision
  assistant. You are able to understand the
  visual content that the user provides, and
  assist the user with a variety of tasks
  using natural language.<|eot_id|>

<|start_header_id|>user<|end_header_id|>
  <image>
  What color is the car?<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>
  -----
  <NORET>
  Black<|eot_id|>
```

In practice, <image> is replaced with the visual tokens generated by the visual encoder, after being projected into the space of the LLM using the vision-to-language adapter of the model (*i.e.* an MLP). For completeness, we report the output of the model below the dashed line. In this case, the query does not require retrieval; therefore, ReflectiVA generates the <NORET> token and directly provides the answer.

When instead the image-question pair requires retrieval to effectively produce the final answer, the model first generates the <RET> token. After performing retrieval and allowing

```
# Predicted Answer: {predicted answer}
You have to determine the alignment between the
predicted answer and the ground-truth on a scale from 0
to 100, where 0 indicates no alignment and 100 indicates
perfect alignment. Your response should be in JSON
format, outputting a list where each element is a
dictionary representing a candidate with:
  ``score``: a numeric value between 0 and 100 indicating
  the alignment level,
  ``reason``: a string explaining the rationale for the
  given score.
```

¹⁰<https://github.com/EvolvingLMs-Lab/lmms-eval>

the model to identify the relevant passages, the answer is finally generated. Specifically, the complete prompt after the retrieval stage is as follows:

```
<|begin_of_text |>
<|start_header_id|>system<|end_header_id|>
  You are a helpful language and vision
  assistant. You are able to understand the
  visual content that the user provides, and
  assist the user with a variety of tasks
  using natural language.<|eot_id|>

<|start_header_id|>user<|end_header_id|>
  <image>
  How big can this plant become?<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>
  <RET>

<|start_header_id|>user<|end_header_id|>
  Consider this paragraph:
  <paragraph>
  Prunus laurocerasus is an evergreen shrub
  or small to medium-sized tree, growing to
  5 to 15 metres (16 to 49ft) tall, rarely to
  18 metres (59ft), with a trunk up to 60cm
  broad. The leaves are dark green, leathery,
  shiny, with a finely serrated margin. The
  leaves can have the scent of almonds when
  crushed. The flower buds appear in early
  spring and open in early summer in erect
  7 to 15cm racemes of 40 flowers, each
  flower 1cm across, with five creamy-white
  petals and numerous yellowish stamens with
  a sweet smell.
  </paragraph>
  Give a short answer.<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>
  -----
  <REL>
  16 to 49ft<|eot_id|>
```

For simplicity, here we include a single textual passage. However, as detailed in Sec. 3.1 of the main paper, the final answer is generated using the set of passages that have been judged relevant. As shown, we also introduce two additional special tokens, `<paragraph>` and `</paragraph>`, to mark the beginning and end of each passage. These tokens are incorporated into the original LLaMA-3.1 vocabulary, as described for the proposed reflective tokens (Sec. 4.1).

B.3. Baselines and Competitor Reproductions

Vanilla LLMs and MLLMs. To evaluate text-only and multimodal LLMs, we adopt the original system prompts provided by the models, appending the instruction “Give a short answer” to enforce concise responses. For text-only LLMs, we supplement the input with automatically generated image descriptions extracted using BLIP-2 [40].

Wiki-LLaVA [9]. Following the original paper, all experiments involving the Wiki-LLaVA model are conducted with the number k of retrieved documents equal to 1, where the

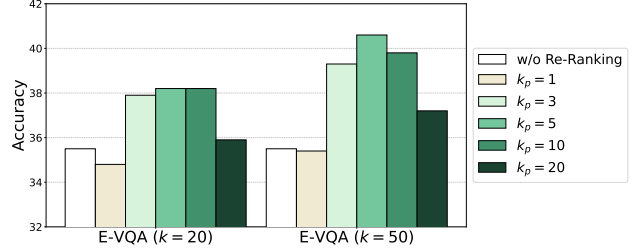


Figure 4. Re-ranking performance on the single-hop question split of Encyclopedic-VQA when varying the number k of retrieved documents and the number k_p of passages given to our model after re-ranking. The “without re-ranking” bars refer to the best results achieved without using the re-ranking component, as reported in Table 2 of the main paper.

Model	LLM	ViQuAE		S3VQA
		F1	EM	GPT-4
LLaVA-v1.5 [46]	Vicuna-7B	15.1	26.6	23.9
LLaVA-v1.5 [46]	LLaMA-3.1-8B	15.0	25.6	24.4
Wiki-LLaVA (E-VQA) [9]◊	LLaMA-3.1-8B	10.5	16.7	22.7
Wiki-LLaVA (InfoSeek) [9]◊	LLaMA-3.1-8B	12.7	21.8	21.8
ReflectiVA (w/o KB)	LLaMA-3.1-8B	16.6	27.6	26.9
ReflectiVA (Ours)	LLaMA-3.1-8B	23.2	38.1	29.3
		(52.0%)		(16.8%)

Table 9. Zero-shot performance on additional knowledge-based VQA datasets. The percentage of samples in which our model incorporates external knowledge is highlighted in green.

LLM is provided with the top-3 passages retrieved via the Contriever model [27]. To evaluate Wiki-LLaVA with the LLaMA-3.1 LLM, we train two separate models as described in the original work: one on the Encyclopedic-VQA training set and the other on the InfoSeek training data, adhering to the same experimental settings.

EchoSight [71]. Experiments with LLaMA-3.1 are conducted by replacing the original LLM (*i.e.* Mistral for Encyclopedic-VQA and LLaMA-3 for InfoSeek) while keeping all other experimental settings unchanged. The LLM is prompted as outlined in the original paper, employing a one-shot example for InfoSeek experiments.

C. Additional Experimental Results

C.1. Further Ablation Studies and Analyses

Results on Other Knowledge-based Datasets. In addition to the results on the Encyclopedic-VQA and InfoSeek datasets, we also validate the generalization capabilities of ReflectiVA to zero-shot settings which always require knowledge retrieval. Specifically, we report the results on two additional knowledge-based VQA datasets, *i.e.* ViQuAE [36] and S3VQA [29]. From Table 9, it can be seen that even in these challenging settings, ReflectiVA achieves the best results, outperforming competitors by a significant margin and demonstrating the usefulness of predicting reflective tokens.

Model	E-VQA		InfoSeek		
	Single-Hop	All	Unseen-Q	Unseen-E	All
<i>Textual Retrieval Mode</i>					
Self-RAG [3]	17.9	17.6	5.0	5.5	4.5
ReflectiVA (Ours)	28.0	29.2	40.4	39.8	40.1
<i>Visual Retrieval Mode</i>					
Self-RAG [3]	18.8	18.4	5.1	4.3	4.6
ReflectiVA (Ours)	35.5	35.5	28.6	28.1	28.3

Table 10. Experimental comparison with Self-RAG on the Encyclopedic-VQA test set and the InfoSeek validation set.

Re-Ranking Analysis when Varying k and k_p . As a complement to the experiments with the re-ranking component shown in Table 3, we report in Fig. 4 the performance of ReflectiVA as a function of the number k of retrieved documents and the number k_p of passages provided to the model after re-ranking. For this experiment, we employ the re-ranker module proposed in [71], trained on the Encyclopedic-VQA dataset¹¹. Specifically, we use $k = \{20, 50\}$ and $k_p = \{1, 3, 5, 10, 20\}$, while also reporting the performance of our best configuration without re-ranking. All experiments are conducted using EVA-CLIP as the retrieval model with image-to-image similarity, as it demonstrates superior performance in the Encyclopedic-VQA setting (cf. Table 1). As expected, increasing the number k of retrieved documents generally improves performance, demonstrating that incorporating a re-ranking stage effectively enhances the results. However, it is worth noting that relying solely on the top-1 retrieved passage does not yield the best overall performance. Instead, the highest accuracy scores are achieved with $k_p = 5$ ¹². These results highlight the ability of ReflectiVA to accurately identify the most relevant passages and effectively utilize them to provide more accurate answers.

Comparison with Self-RAG [3]. In addition to task-specific competitors, we also compare our model with Self-RAG [3] that incorporates special tokens to improve retrieval-augmented generation pipelines. In particular, this model has been designed for natural language understanding tasks, like open-ended question answering, reasoning, and fact verification, and has no multimodal capabilities. To conduct the comparison, we adapt Self-RAG to our setting by employing the same retrieval pipeline as our model. After retrieving the top- k documents¹³, we prompt the LLM with all passages from the retrieved documents, allowing the model to identify the most useful information for answering the input question. Since the underlying LLM is not equipped to process visual inputs, we include the image description extracted

¹¹As mentioned in the main paper and shown in Table 7, re-ranking results can not be reported for the InfoSeek dataset, as the original knowledge base used in our experiments differs significantly from the one employed in [71], which was derived from Encyclopedic-VQA data.

¹²All experiments reported in Table 3 of the main paper are conducted using $k = 50$ and $k_p = 5$.

¹³For consistency with our setting without re-ranking, we use $k = 5$.

	<RET>		<NORET>	<REL>	<NOREL>	
	E-VQA	InfoSeek	GQA	E-VQA (Pos)	E-VQA (Soft)	E-VQA (Hard)
GPT-4 [1]	82.5	73.5	94.5	93.8	93.4	91.3
GPT-4V [1]	94.4	96.3	96.6	94.4	92.4	94.3
After LLaVA 1st stage	80.6	99.7	100.0	93.4	96.8	94.8
After LLaVA 2nd stage	88.4	100.0	100.0	94.6	95.9	96.2

Table 11. Accuracy scores achieved by ReflectiVA in predicting the proposed reflective tokens, compared with the performance of GPT-4 and GPT-4V.

using BLIP-2 [40] in the input prompt, as done for vanilla LLMs. Results on both Encyclopedic-VQA and InfoSeek are presented in Table 10. Notably, ReflectiVA consistently outperforms Self-RAG across all settings, further highlighting the effectiveness of our approach and the critical role of incorporating multimodal information for solving the task.

Evaluating Reflective Token Accuracy. To better analyze the performance of our model, we assess its ability to correctly predict the introduced reflective tokens. Specifically, we evaluate the accuracy of the <RET> and <NORET> tokens by extracting a subset of 11k image-question pairs from the validation sets of Encyclopedic-VQA, InfoSeek, and GQA. For this experiment, we assume that all samples from Encyclopedic-VQA and InfoSeek necessitate retrieval, while all image-question pairs in GQA, which contain generic questions about the image content, can be answered without external retrieval. Additionally, we evaluate the prediction accuracy of the <REL> and <NOREL> tokens on a subset of the Encyclopedic-VQA validation set consisting of 500 image-question pairs, where the relevant textual snippet containing the answer is available. For each sample in this subset, we extract the relevant passage containing the snippet with the answer, along with two non-relevant passages. Specifically, we extract a soft negative passage from an unrelated document and a hard negative passage from the same document, ensuring that the relevant textual snippet is not included in the hard negative.

Accuracy scores are shown in Table 11, where we compare the performance of zero-shot models, such as GPT-4 and GPT-4V, with the results obtained by applying our strategy after either the first or second stage of LLaVA training. LLaVA-v1.5 follows a two-stage training process: the first stage pre-trains on image-caption pairs to align image features with the LLM textual space, while the second stage focuses on enhancing multimodal conversational capabilities. We therefore analyze the optimal point in this pipeline to incorporate our strategy for learning new special tokens. As it can be seen, accuracy scores are consistently higher than 85% for all reflective tokens, highlighting the ability of the model to accurately predict when retrieval is necessary and whether the retrieved passages are relevant to the query. Also, applying our training strategy after the second stage of LLaVA training generally yields the best results.

C.2. Does ReflectiVA Integrate Built-In Re-Ranking Capabilities?

As a final analysis, we evaluate whether our model can be directly employed to re-rank retrieved passages. To this end, we first retrieve the top- k most relevant documents from the external knowledge base and use the log probabilities of $\langle \text{REL} \rangle$ and $\langle \text{NOREL} \rangle$ tokens to rank the associated textual passages. Specifically, for each passage, we compute the difference between the log probability of the $\langle \text{REL} \rangle$ token and that of the $\langle \text{NOREL} \rangle$ token, ranking all passages in descending order based on the computed difference scores. We then provide the top- k_p passages as input to the model to generate the final answer. This analysis is conducted on the Encyclopedic-VQA dataset, with results reported in Table 12. As it can be seen, increasing the number k of retrieved documents and directly leveraging the probabilities of $\langle \text{REL} \rangle$ and $\langle \text{NOREL} \rangle$ tokens to rank passages leads to the best results. Notably, ReflectiVA with the built-in re-ranking strategy achieves 37.8 accuracy points on single-hop questions using k equal to 50 and k_p equal to 2, outperforming the best result without re-ranking by 2.3 points. While training a dedicated re-ranking module could further improve performance, this analysis demonstrates that ReflectiVA natively integrates effective re-ranking capabilities.

D. Qualitative Results

To comprehensively evaluate the proposed ReflectiVA model, we present additional qualitative results in Fig. 5 and Fig. 6, alongside answers generated by Wiki-LLaVA [9] and EchoSight [71]. These results are based on sample image-question pairs from Encyclopedic-VQA and InfoSeek, respectively. Notably, ReflectiVA effectively handles diverse question types, including those requiring numeric answers such as dates or measurements (*e.g.* the top-right example in Fig. 5 and the top-left example in Fig. 6), as well as questions demanding domain-specific knowledge (*e.g.* the top-left example in Fig. 5 and the bottom-left example in Fig. 6). Conversely, both Wiki-LLaVA and EchoSight often fail to answer the given questions correctly, either generating an incorrect response or a statement indicating their inability to provide an answer. These results qualitatively highlight the superior performance of ReflectiVA compared to existing methods for the task.

E. Limitations and Failure Cases

As the final part of the qualitative analysis, we present some failure cases of our model in Fig. 7. These examples illustrate the challenges the model faces, such as adhering to the expected style of correct answers. This issue is particularly evident in datasets like InfoSeek, where the evaluation is based on exact matches between ground-truth and generated answers. This is evident in the bottom-left example, where

Built-In Re-Ranking	k	k_p	E-VQA	
			Single-Hop	All
\times	5	-	35.5	35.5
\checkmark	5	1	34.7	34.8
\checkmark	5	2	35.0	34.9
\checkmark	5	5	33.4	33.4
\times	20	-	35.7	35.2
\checkmark	20	1	36.3	35.8
\checkmark	20	2	36.6	36.6
\checkmark	20	5	35.9	35.6
\times	50	-	29.6	29.1
\checkmark	50	1	37.6	37.1
\checkmark	50	2	37.8	37.2
\checkmark	50	5	36.7	36.3

Table 12. Experimental analysis of built-in re-ranking capabilities on the Encyclopedic-VQA test set, varying the number k of retrieved documents and the number k_p of selected passages after re-ranking. Best results for each k are in bold, and the best overall results are underlined.

the answer conveys the same semantic meaning but is expressed in a different way. This may lead to discrepancies when evaluating the correctness of the given answer despite its semantic alignment with the ground-truth. Other errors arise from the specificity of the questions. For instance, in the bottom-center example, the question asks about the maximum velocity of a cheetah, requiring high precision. In this case, ReflectiVA provides a reasonable estimate, demonstrating its ability to address such detailed inquiries.

Q: What was designated grade II on the same day as this building?



Wiki-LLaVA [9]:
A of the heritage ✗
EchoSight [71]:
It seems you are asking about the Grade II listing of the building ✗
ReflectiVA (Ours):
Gate Piers ✓

Q: Is Fr. Emilio Biosca Agüero the current or former pastor of this church?



Wiki-LLaVA [9]:
Former ✗
EchoSight [71]:
I could not find any information about Fr. Emilio Biosca Agüero ✗
ReflectiVA (Ours):
Current ✓

Q: How did this bird fare against the Australian white ibis?



Wiki-LLaVA [9]:
The world's largest ✗
EchoSight [71]:
There is no mention of the Australian white ibis in the provided text ✗
ReflectiVA (Ours):
Narrowly ahead ✓

Q: What portion of the exterior of this railway station is on Belmont street?



Wiki-LLaVA [9]:
North ✗
EchoSight [71]:
I can not verify information about the exterior of the railway station ✗
ReflectiVA (Ours):
Only visible portion ✓

Q: What does the oily plumage of this bird prevent its feathers from getting?



Wiki-LLaVA [9]:
Wet ✗
EchoSight [71]:
Unfortunately, the text does not mention a specific bird with oily plumage ✗
ReflectiVA (Ours):
Waterlogged ✓

Q: How often is the kite festival held at this castle?



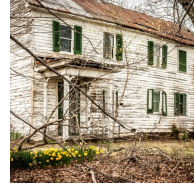
Wiki-LLaVA [9]:
Never ✗
EchoSight [71]:
There is no information in the provided text about the frequency of a kite festival ✗
ReflectiVA (Ours):
Annual ✓

Q: When was the current structure of this church dedicated?



Wiki-LLaVA [9]:
1931 ✗
EchoSight [71]:
The provided text does not mention the dedication date of the current structure ✗
ReflectiVA (Ours):
1922 ✓

Q: What was attached to the western end of the original portion of this building?



Wiki-LLaVA [9]:
The United States ✗
EchoSight [71]:
A kitchen (attached to the house) and living quarters for maids ✗
ReflectiVA (Ours):
Chimney ✓

Q: What kind of insect is this?



Wiki-LLaVA [9]:
Aa wasp, the spider ✗
EchoSight [71]:
The Watercress Darter is actually a type of fish, not an insect ✗
ReflectiVA (Ours):
Beetles ✓

Figure 5. Additional qualitative results on image-question pairs from Encyclopedic-VQA, where we compare the answers provided by ReflectiVA with those from WikiLLaVA [9] and EchoSight [71].

Q: What is the area in square kilometer occupied by this lake?



Wiki-LLaVA [9]:
9.82 ✗
EchoSight [71]:
5.34 ✗
ReflectiVA (Ours):
1.18 ✓

Q: What is the source of energy of this vehicle?



Wiki-LLaVA [9]:
Diesel ✗
EchoSight [71]:
Diesel ✗
ReflectiVA (Ours):
Coal ✓

Q: What is the architectural style of this place?



Wiki-LLaVA [9]:
There is no specific answer to the question about the architectural style in the text ✗
EchoSight [71]:
Georgian architecture ✗
ReflectiVA (Ours):
Greek Revival architecture ✓

Q: Which class of biological feature is this food produced by?



Wiki-LLaVA [9]:
Malt house ✗
EchoSight [71]:
Plants ✗
ReflectiVA (Ours):
Lactobacillus delbrueckii ✓

Q: Where are organization initially formed?



Wiki-LLaVA [9]:
Wright brother's brother ✗
EchoSight [71]:
I cannot provide information or context about the accident of Transbrasil Flight 801 ✗
ReflectiVA (Ours):
Europe ✓

Q: Which crystal system does this material have?



Wiki-LLaVA [9]:
Hexagonal ✗
EchoSight [71]:
There is no crystal system mentioned in the text, so I will say: None ✗
ReflectiVA (Ours):
Trigonal ✓

Q: What is the density (in gram per cubic centimeter) of this place?



Wiki-LLaVA [9]:
100 ✗
EchoSight [71]:
There is no information about the density of this place ✗
ReflectiVA (Ours):
1408 ✓

Q: Which place is this animal endemic to?



Wiki-LLaVA [9]:
North Pacific Ocean ✗
EchoSight [71]:
North America ✗
ReflectiVA (Ours):
New Zealand ✓

Q: Which street is this building located at?



Wiki-LLaVA [9]:
Rue de Rivoli ✗
EchoSight [71]:
There is no street mentioned in the text ✗
ReflectiVA (Ours):
Rue des Francs-Bourgeois ✓

Figure 6. Additional qualitative results on image-question pairs from InfoSeek, where we compare the answers provided by ReflectiVA with those from WikiLLaVA [9] and EchoSight [71].

Q: Terry park became the pirates spring what ballpark?



Ground-truth:
Training home

ReflectiVA (Ours):
Training ground

Q: In what cuisine is this vegetable used?



Ground-truth:
Southeast Asia

ReflectiVA (Ours):
Indian cuisine

Q: What is the fee to see this gate from the outside?



Ground-truth:
No admission fee

ReflectiVA (Ours):
Free

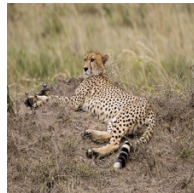
Q: What is the objective of this object?



Ground-truth:
Energy transformation

ReflectiVA (Ours):
Engine tuning

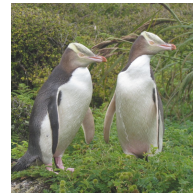
Q: What is the magnitude of the velocity of this animal kilometer per hour?



Ground-truth:
120

ReflectiVA (Ours):
150

Q: What is the conservation status of this bird?



Ground-truth:
Endangered

ReflectiVA (Ours):
Vulnerable

Figure 7. Examples of failure cases on image-question pairs from Encyclopedic-VQA (top row) and InfoSeek (bottom row).