

ReAG: Reasoning-Augmented Generation for Knowledge-based Visual Question Answering

Supplementary Material

A. Additional Implementation Details

This section contains detailed descriptions of our implementation, including model architectures details, training setup, and evaluation protocols.

Retrieval Details. As mentioned in the main paper, to retrieve candidate documents for each query image, we adopt a two-stage retrieval pipeline based on EVA-CLIP-8B. In the coarse-grained stage, the full query image is encoded and matched against the knowledge base. Instead, in the fine-grained retrieval phase, the subject of the question, corresponding to the visual entity in the image, is extracted using SpaCy⁶. According to the structure of query questions, extraction prioritizes noun phrases starting with demonstratives like “this” or “these”, followed by nouns serving as objects of prepositions (e.g., “of”, “by”, “in”, “from”, etc.). If neither pattern is found, the last noun in the question is used as a fallback. This approach ensures that the most relevant entity is reliably identified for retrieval and reasoning.

Critic Model and Dataset. The critic model builds upon Qwen2.5-VL-3B and is fine-tuned on a curated subset of the ReflectiVA dataset [13]. Specifically, we select 1M samples, balanced between InfoSeek and Encyclopedic-VQA, where each sample is paired with a passage labeled as relevant or irrelevant. To encourage more robust discrimination, irrelevant passages are further divided into *soft negatives* (i.e., semantically related but unhelpful passages) and *hard negatives* (i.e., irrelevant passages that exhibit high similarity to the relevant ones), in proportions of 30% and 70% respectively. During training, we fine-tune only the visual projector and the LLM, while keeping the vision encoder frozen.

At inference time, given (I_q, q, p) (where I_q and q are the input image and question, and p is the passage to evaluate), we apply a *yes*-probability threshold equal to 0.1 (cf. Eq. 4 of the main paper). This conservative threshold ensures that the critic \mathcal{C} filters out only those passages for which it is highly confident they are irrelevant.

Generator Training. Our generator is based on Qwen2.5-VL [4], employing the 3B and 7B scale. As mentioned in the main paper, we employ a multi-stage training strategy: first an SFT stage that serves as cold start for a subsequent RL-based training stage. Specifically, for both the cold-start and RL stages, we apply a cosine learning rate schedule with 450 and 150 warm-up steps, respectively. Weight decay is set to 0.01 during cold-start and removed during RL. SFT

fine-tuning of the 3B and 7B models requires roughly 24 hours on 16 and 64 NVIDIA A100 64GB GPUs, respectively. In the RL stage, completions are generated with a temperature equal to 1.0 and a repetition penalty of 1.05, using vLLM [23].

During training with the custom GRPO loss (cf. Eq. 6 of the main paper), \mathcal{G}_θ and $\mathcal{G}_{\theta_{\text{old}}}$ share the same weights, although they operate as separate models. $\mathcal{G}_{\theta_{\text{old}}}$ runs inside the vLLM worker and remains frozen, and gradients are applied only to \mathcal{G}_θ . After each backward pass, we synchronize $\mathcal{G}_{\theta_{\text{old}}}$ with the updated weights of \mathcal{G}_θ . Fine-tuning the 3B and 7B models with our RL strategy takes roughly 48 hours on 32 and 64 NVIDIA A100 64GB GPUs, respectively. We select the best checkpoint based on the best task-specific accuracy on a held-out validation split. All runs employ DeepSpeed ZeRO-3 [43] and gradient checkpointing. Our training codebase builds on Open-R1 [18] and TRL [49].

B. Reward Design

As discussed in Sec. 3.4 of the main paper, we employ a task-specific accuracy reward. The reward function evaluates only the final answer rather than the intermediate reasoning. To extract the predicted answer, we first search for content enclosed within the `<answer></answer>` tags. If no such content is found, we extract all text following the first `<answer>` tag. If this is unsuccessful, we instead use the text following the `</think>` tag. When none of these patterns appear, the entire model output is used.

In every case, format-specific special tokens are removed. We then apply the same normalization procedure used in the InfoSeek and Encyclopedic-VQA evaluations, including the removal of articles, punctuation, extra whitespace, and capitalization, along with standardization of digits and contractions. The final task-specific reward depends on the source dataset and task type, and is computed as follows:

$$R_{\text{task}}(\tilde{o}_i, o_i^*, \tau_i) = \begin{cases} \mathbb{1}[\Psi_{\text{num}}(\tilde{o}_i, o_i^*)], & \text{if } \tau_i = \text{numerical}, \\ \mathbb{1}[\text{IoU}(\tilde{o}_i, o_i^*) \geq 0.5], & \text{if } \tau_i = \text{multi}, \\ \mathbb{1}[\tilde{o}_i = o_i^*], & \text{otherwise.} \end{cases}$$

where \tilde{o}_i , o_i^* and τ_i denote respectively the extracted prediction, the ground-truth answer and the task type of the i -th sample, and Ψ_{num} evaluates success or failure in numerical match. When multiple alternative ground-truths are provided for a sample, we compute the reward with respect to each and take the maximum. For samples from InfoSeek, we use

⁶In particular, we use the `en_core_web_sm` model available at <https://spacy.io/models/en>.

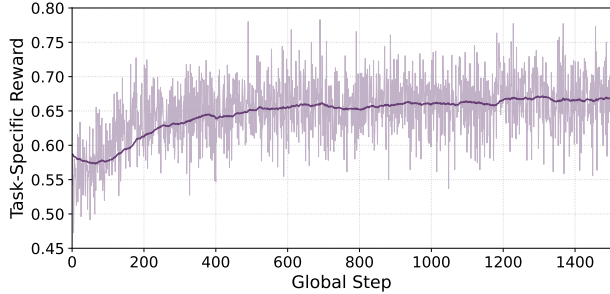


Figure 5. Task-specific accuracy reward progression across training iterations of the ReAG 7B generator.

exact string matching for *entity* and *time* questions, while *numerical* questions are evaluated with Ψ_{num} :

$$\Psi_{\text{num}}(\tilde{o}, o^*) = \begin{cases} |\tilde{o} - o^*| \leq 0.1, & \text{if } \text{is_scalar}(\tilde{o}) \\ & \wedge \text{is_scalar}(o^*), \\ \tilde{o} \in o^*, & \text{if } \text{is_scalar}(\tilde{o}) \\ & \wedge \text{is_interval}(o^*), \\ \text{IoU}(\tilde{o}, o^*) \geq 0.5, & \text{if } \text{is_interval}(\tilde{o}) \\ & \wedge \text{is_interval}(o^*). \end{cases}$$

For samples from Encyclopedic-VQA dataset, we adopt exact match scoring for *single-answer* questions, while for *multi-answer* questions the prediction is rewarded as correct only if intersection-over-union between predicted and ground-truth items reaches or surpasses 0.5.

The evolution of task-specific accuracy reward during training is reported in Fig. 5.

C. Additional Experimental Results

In this section, we provide additional experiments and analyses that complement the results reported in the main paper.

C.1. Results with Google Lens Retriever

We further extend the analysis in the main paper by evaluating all methods under an alternative retrieval setup. Specifically, in Table 5, we employ the Wikipedia pages retrieved by Google Lens⁷ when provided with the query image for each question of Encyclopedic-VQA, which have been officially released along with the dataset. Even though ReAG does not use the fine-grained retriever in this setting, it consistently outperforms ReflectiVA [13] across different generator scales. Notably, ReAG at the 3B scale performs comparable to ReflectiVA at the 8B scale and HAMMR [9] at 55B. In addition, the stronger reasoning capabilities of ReAG allow it to benefit substantially from improved retrieval quality, improving single-hop accuracy from 48.0 (3B) to 55.5 (7B), showing a notable gain of +7.5 points.

⁷A visual recognition service by Google, available at <https://lens.google.com/>.

Table 5. VQA accuracy scores on the Encyclopedic-VQA test set with Google Lens employed as retriever.

Model	Generator	E-VQA	
		Single-Hop	All
Qwen2.5-VL-3B [4]	-	35.2	-
Qwen2.5-VL-7B [4]	-	44.4	-
HAMMR [9]	PaLI-X-55B	47.8	-
mR ² AG [59]	LLaVA-v1.5-7B	55.9	-
ReflectiVA [13]	LLaVA-MORE-8B	48.2	46.1
ReflectiVA [13]	Qwen2.5-VL-3B	45.8	43.2
ReAG (Ours)	Qwen2.5-VL-3B	48.0	48.0
ReflectiVA [13]	Qwen2.5-VL-7B	48.3	46.8
ReAG (Ours)	Qwen2.5-VL-7B	55.5	56.9

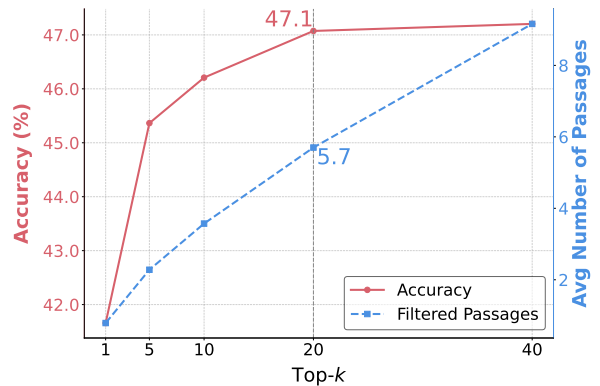


Figure 6. Performance of ReAG 7B (red) and average number of filtered passages (blue) when varying the number k of retrieved documents. Accuracy and number of passages are computed as the average of E-VQA and InfoSeek scores.

C.2. Varying the Number of Retrieved Documents

In Fig. 6, we analyze the effect of varying the number of retrieved documents k on the overall performance and on the average number of filtered passages fed to the generator. As shown, the model achieves the best results around $k = 20$, which represents the optimal trade-off between coverage and noise, and is therefore adopted as the default retrieval depth in our pipeline. Retrieving too few documents results in insufficient contextual evidence, causing a drop in recall and limiting the ability of the model to access the necessary information. Conversely, increasing k beyond this point does not yield meaningful performance gains while substantially inflating the computational cost of the filtering stage.

C.3. Effectiveness of the Critic Model

Critic Analysis. In Fig. 7, we provide a detailed analysis on the effectiveness of the proposed critic model, employed in ReAG to filter relevant passages. Specifically, the plot reports the average number of passages retained after the filtering performed by the critic model when varying the number k of retrieved documents. Across all retrieval sizes, the critic model substantially reduces the number of retained

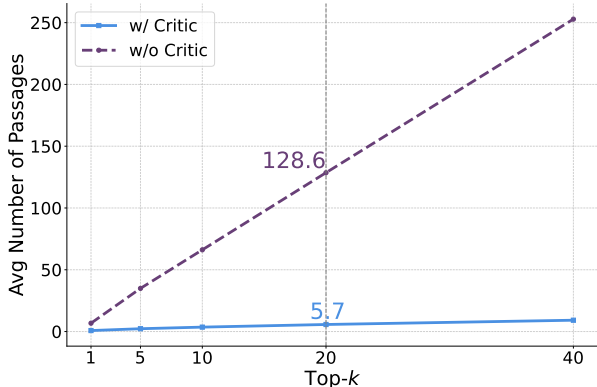


Figure 7. Comparison of the average number of passages fed to the generator with and without the critic filtering.

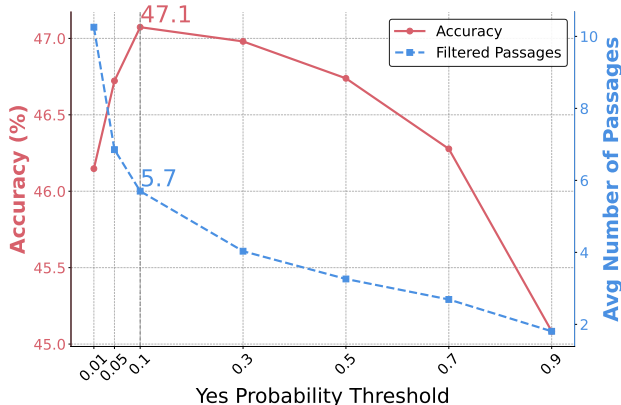


Figure 8. Performance of ReAG 7B (red) and average number of filtered passages (blue) when varying the *yes*-probability threshold in our critic model. Accuracy and number of passages are computed as the average of E-VQA and InfoSeek scores.

passages (*e.g.*, from an average of 128.6 to 5.7 at $k = 20$), while preserving answer-relevant information. This highlights the strong ability of the critic model to discard noisy or off-topic passages, leading to a more compact and semantically aligned evidence set for multimodal reasoning.

Critic Threshold. In Fig. 8, we report how the performance varies as a function of the *yes*-probability threshold used in our critic model (cf. Eq. 4). The results show that instead of simply letting the fine-tuned MLLM decide if the passage is relevant or not (*i.e.*, $\text{thresh} = 0.5$), leveraging the confidence of the model in predicting the “Yes” token allows us to gain more control over the filtering phase. A threshold $\text{thresh} = 0.1$ yields the best trade-off between precision and recall in retrieving relevant passages. This setting enables the critic model to reliably filter out passages for which it is most confident of their non-relevance to the query.

Filtering Analysis. To evaluate the quality of our filtering stage, we report passage-level statistics measuring evidence retention (Recall) and irrelevant-passage recognition (Specificity) in Table 6. Our Critic retains relevant evidence with higher recall while filtering irrelevant passages with higher

Table 6. Recall and specificity for passage filtering on E-VQA.

Model		Recall	Specificity
ReflectiVA [13]	LLaVA-MORE-8B	89.6	93.8
ReflectiVA [13]	Qwen2.5-VL-3B	91.3	93.8
ReflectiVA [13]	Qwen2.5-VL-7B	92.3	95.1
ReAG (Ours)	Qwen2.5-VL-3B	94.6	95.7

specificity compared to ReflectiVA. As a result, the generator receives a cleaner context with fewer passages on average, reducing noise while preserving useful evidence.

C.4. Sources of Performance Gains

Impact of Evidence Quality. As shown in the ablation study (Table 4), equipping the generator with our retrieval pipeline – composed of the Critic and fine-grained retrieval – already yields substantial gains in a Zero-Shot setting, reaching 40.2 on E-VQA and 27.1 on InfoSeek. This indicates that a significant portion of the improvements stems from higher-quality evidence selection rather than parametric memorization.

Training for Visual Robustness. The SFT+RL procedure further improves performance by teaching the generator to leverage retrieved evidence when available and rely on visual reasoning otherwise. To encourage this behavior, the training data includes both relevant passages and distractors (30% soft negatives and 70% hard negatives), forcing the model to distinguish useful evidence from noise.

Finally, the relatively small differences in the *w/o evidence* setting (Fig. 3, left) are partly due to a design choice in our retrieval pipeline. Unlike [13], which inserts a random passage when all evidence is filtered out, our method provides no fallback context, avoiding the risk of conditioning the generator on misleading information.

D. Critic and Generator Prompts

Critic Prompt Design. The prompt illustrates the instruction used to query the critic model. For each image-question pair (I_q, q) and a candidate passage p , the critic is explicitly asked to determine whether p contains any information that could help answer the question. The formulation uses a minimal, binary response space (*i.e.*, “Yes”/“No”), which simplifies supervision and ensures consistent outputs across diverse samples. This concise design encourages the model to focus on relevance estimation rather than generative reasoning, enabling more stable fine-tuning and robust filtering of noisy passages during retrieval.

Critic System Prompt

You are a multimodal reasoning assistant specialized in Knowledge-Based Visual Question Answering (KB-VQA). Your task is to evaluate whether a given text passage provides useful and relevant

information for answering a question about an image.

You will be given:

- **Image:** a visual scene containing entities, actions, and context.

- **Question:** a natural-language question that refers to the image.

- **Text Passage:** an external knowledge snippet retrieved from a database.

You must analyze the semantic alignment between the text, the image, and the question. Follow these steps carefully before giving your final decision:

1. Understand the visual scene: Identify the key objects, people, actions, and context visible in the image.

2. Interpret the question: Determine what information the question seeks (e.g., factual, reasoning, counting, attribute-based).

3. Analyze the text passage: Extract the main claims, facts, and entities mentioned in the text.

Compare for relevance: Assess whether the information in the text:

- Contains at least one sentence that supports answering the question about the image, OR

- Provides background knowledge needed to interpret or reason about the image-question pair.

Important:

- If even a single sentence in the passage is relevant or useful, consider the entire passage as relevant and answer "Yes".

- If no part of the passage contributes meaningfully to answering the question, answer "No".

Output only one word:

"Yes" -> if the text provides relevant or useful information for answering the question.

"No" -> if the text is irrelevant or unhelpful.

Critic User Prompt

Here is the question on the image above:

{Question}

Here is the text passage to analyze:

{Passage}

Does the text passage contain at least one sentence that may have some information useful to answer the user question?

"Yes"/"No" answer:

Generator Prompt Design. This prompt defines the instruction for the generator model, which receives the image, question, and textual context. During training, the model is provided with only the single passage associated with the current example from the ReflectiVA dataset [13], whereas at inference it takes the subset of passages selected by the critic model. The generator is prompted to synthesize a fi-

nal answer grounded in both visual and textual evidence. At inference, providing critic-filtered passages as input encourages concise, evidence-based reasoning, reduces the impact of irrelevant or noisy information, and improves factual grounding in multimodal responses. The generator system prompt is adapted from Dr. GRPO [33]. Notably, when the critic model filters all retrieved passages (*i.e.*, when $j = 0$) the user prompt is changed and only the question with the image is fed to the generator.

Generator System Prompt

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, *i.e.*, <think>reasoning process here</think><answer>short answer here</answer>.

Generator User Prompt

{Question}

The following paragraphs may contain useful information to help answer the question correctly:

<paragraph>{Passage₁}</paragraph>

...

<paragraph>{Passage_j}</paragraph>.

Reasoning-Trace Prompt Design. To extract reasoning traces used during the SFT training stage (cf. Sec. 3.3 of the main paper), we employ a structured prompting strategy that elicits explicit, step-by-step inference from a teacher MLLM. The system prompt instructs the model to analyze the image, the question, and one retrieved passage, then produce a hidden reasoning trace (<think>...</think>) that (i) grounds its steps in visual evidence (*e.g.*, objects, attributes, spatial relations), (ii) evaluates the content of the passage and explicitly states whether it is relevant or irrelevant, and (iii) connects visual and textual cues via a logical chain. The user prompt supplies the question, the retrieved passage with its relevance tag, and the correct answer; the model must output the trace plus the final answer in a strict schema. Collected reasoning traces are used to initialize the generator with explicit reasoning trajectories that link the image, retrieved evidence, and the question, thereby strengthening its reasoning capabilities before the RL stage.

Reasoning Traces Generation System Prompt

You are a multimodal reasoning assistant. Your goal is to analyze the image, the question, and the retrieved passage, and

then produce a hidden reasoning trace that logically leads to the given answer. The reasoning must be step-by-step, plausible, and based on both the visual evidence and the retrieved text passage. You MUST explicitly state, within your reasoning trace, whether the passage is relevant or not according to the information provided (i.e., if it is labeled as "irrelevant", your reasoning must clearly and logically explain why it is not relevant, and if it is labeled as "relevant", your reasoning must logically support its relevance).

Do not mention, restate, or hint at the correct answer in the reasoning trace. Your reasoning trace should include:

- Description of relevant visual evidence (objects, spatial relations, attributes).
- Analysis of the retrieved passage (what it states, whether it supports or contradicts the image/question, and its relevance).
- Logical deduction that connects the visual and textual evidence to reach a conclusion.

Here you have two good examples of reasoning traces:

EXAMPLE 1 with Relevant passage:
 Question: {Example 1 Question}
 Retrieved **Relevant** Passage:
 {Example 1 Relevant Passage}
 Correct answer: {Example 1 Answer}
 Output:
 <think> {Example 1 Reasoning Trace} </think>
 <answer> {Example 1 Answer} </answer>

EXAMPLE 2 with Irrelevant passage:
 Question: {Example 2 Question}
 Retrieved **Irrelevant** Passage:
 {Example 2 Irrelevant Passage}
 Correct answer: {Example 2 Answer}
 Output:
 <think> {Example 2 Reasoning Trace} </think>
 <answer> {Example 2 Answer} </answer>

Output your reasoning and the correct answer using the exact format below:
 <think> [your reasoning trace here] </think>
 <answer> [the provided answer] </answer>

Reasoning Traces Generation User Prompt

Question: {Question}
 Retrieved **{Relevant}** Passage: {Passage}
 Correct Answer: {Answer}
 Please produce a reasoning trace that could logically lead to this answer, based on both the image and the retrieved passage if relevant.
 Do not mention or hint at the answer explicitly in your reasoning.
 Concentrate on providing a coherent explanation that supports the indicated relevance or irrelevance of the passage

in the reasoning trace, integrating both textual and visual evidence.
 Make sure to insert the correct answer between the answer tags.

E. Additional Qualitative Results

Reasoning Traces. To further interpret the behavior of our model, we visualize qualitative examples of the reasoning traces generated by ReAG in Fig. 9. The zero-shot baseline produces partial reasoning but lacks a consistent structure and does not adhere to the output format defined by the evaluation datasets. In contrast, ReflectiVA follows the correct answer format but fails to generate explicit reasoning traces, limiting interpretability. In contrast, ReAG generates coherent, well-structured traces that reveal the step-by-step logic behind its predictions. These examples highlight the ability of the proposed solution to integrate visual and textual cues, assess the relevance of retrieved passages, and maintain consistent reasoning even under noisy or irrelevant evidence, where baselines often over-rely on passages or hallucinate unsupported details.

KB-VQA Qualitative Results. Fig. 10 presents additional qualitative examples from the InfoSeek and Encyclopedic-VQA benchmarks, comparing the responses of ReAG, ReflectiVA [13]. As shown, ReAG produces answers that remain aligned with both the visual content and the retrieved evidence, benefiting from its critic-guided filtering and structured reasoning. These examples further highlight the robustness of ReAG in handling complex, knowledge-driven VQA scenarios.

F. Limitations and Impact

While ReAG demonstrates strong performance across standard benchmarks, it still faces some limitations. First, the generator produces a detailed reasoning trace, which improves the explainability of the final answer but may also increase latency, as more tokens must be generated before producing the answer. Second, the quality of ReAG depends on the reliability of the retrieved evidence. Although the critic effectively filters irrelevant passages, retrieval failures or missing knowledge can still lead to incomplete or incorrect reasoning. Moreover, the model may occasionally over-structure its explanations, producing reasoning that is correct in format but not perfectly aligned with human logic.

Despite these limitations, the explicit separation of evidence filtering from reasoning and answer generation enables ReAG to achieve strong performance while promoting greater transparency and explainability, potentially inspiring future research on modular and trustworthy multimodal reasoning frameworks.



Q: Controversy surrounds the origin of what in the city from which one would drive to reach this nature reserve?

Qwen2.5-VL-7B (ZS) [4]:

Based on the information provided, the city from which one would drive to reach the nature reserve is Port Augusta West. The text mentions that Port Augusta West is a suburb of Port Augusta, South Australia, and lists it as the starting point for driving to the nature reserve ✗

ReflectiVA [13]:

The Town ✗

ReAG (Ours):

<think> The image shows a nature reserve with a tree silhouette against a sunset sky, which matches the setting described in the retrieved passages. The question specifically asks for the origin of something in a city from which one would drive to reach this particular nature reserve. None of the retrieved passages provide any relevant information whatsoever regarding the question asked. Therefore, the final logical is that these passages must be disregarded insofar as the question is concerned. </think> <answer> Willi Stoph ✓ </answer>



Q: What genus are this plant and schefflera actinophylla in?

Qwen2.5-VL-7B (ZS) [4]:

Not enough information is provided to determine the genus in question ✗

ReflectiVA [13]:

Schefflera ✗

ReAG (Ours):

<think> The image shows a plant that matches the description of Schefflera actinophylla in the retrieved passage. The question specifically asks for the genus of this particular plant in the image and Schefflera actinophylla. Both plants are identified as belonging to the same genus in the retrieved passage. The question directly corresponds to the information presented in the retrieved passage without any need to refer to the other retrieved passages. Therefore, the final verdict is that the final question must ignore these retrieved passages entirely. </think> <answer> Heptapleurum ✓ </answer>



Q: Who designed this dock?

Qwen2.5-VL-7B (ZS) [4]:

Not enough information is provided to determine who designed the dock ✗

ReflectiVA [13]:

Jesse Hartley ✗

ReAG (Ours):

<think> The image shows the Albert Dock in Liverpool, which matches the question asked. The question specifically asks for the designer of this particular dock. The retrieved passage directly states that the dock in question, the Albert Dock, was designed by Jesse Hartley. Since the question specifically matches the dock in the image with the one mentioned in the passage, the retrieved passage directly answers the question without any need to refer to the question itself or the retrieved passage. </think> <answer> Royal Albert Dock ✓ </answer>

Figure 9. Qualitative examples from Encyclopedic-VQA where we show the ability of ReAG to reason about the question-image pair and the retrieved passages. For reference, we also report the generation from ReflectiVA [13] and the corresponding zero-shot model.

Q: What did the zuni people use the bark of this plant to dye black?



Qwen2.5-VL-7B (ZS) [4]:
The plant you are referring to is the Desert Paintbrush [...] ✗
ReflectiVA [13]:
Dyes ✗
ReAG (Ours):
Deerskin ✓

Q: What part of a tree is this plant often found under?



Qwen2.5-VL-7B (ZS) [4]:
Polystichum is often found under trees in disturbed habitats such as road cuts, talus slopes, and stream banks ✗
ReflectiVA [13]:
Trunk ✗
ReAG (Ours):
Canopy ✓

Q: In what german city is this landmark the most famous?



Qwen2.5-VL-7B (ZS) [4]:
The landmark in the picture is the Brunswick Lion ✗
ReflectiVA [13]:
Braunschweig ✗
ReAG (Ours):
Brunswick ✓

Q: What type of bird is the Contopus cooperi?



Qwen2.5-VL-7B (ZS) [4]:
Contopus cooperi (Olive-sided Flycatcher) ✗
ReflectiVA [13]:
Passerine ✗
ReAG (Ours):
Flycatcher ✓

Q: Who dies without a male heir in the netherlands the year this plant was introduced to canada?



Qwen2.5-VL-7B (ZS) [4]:
Queen Beatrix ✗
ReflectiVA [13]:
The dutch royal family ✗
ReAG (Ours):
King William III ✓

Q: In what month does it not snow in this lake?



Qwen2.5-VL-7B (ZS) [4]:
The lake is not affected by snowfall, [...] ✗
ReflectiVA [13]:
July ✗
ReAG (Ours):
July and August ✓

Q: Who designed this palace?



Qwen2.5-VL-7B (ZS) [4]:
Not enough information is provided to determine [...] ✗
ReflectiVA [13]:
Peter Frederick Robinson ✗
ReAG (Ours):
John Nash ✓

Q: What type of hawksbeard is this plant commonly known as?



Qwen2.5-VL-7B (ZS) [4]:
The plant in the image is likely rough hawksbeard ✗
ReflectiVA [13]:
Mouse-ear ✗
ReAG (Ours):
Narrowleaf ✓

Q: What monument was placed under state control in 1882?



Qwen2.5-VL-7B (ZS) [4]:
Newgrange was placed under state control in 1882 ✗
ReflectiVA [13]:
Newgrange ✗
ReAG (Ours):
Newgrange and Knowth and Dowth ✓

Q: What is the habitat of this animal?



Qwen2.5-VL-7B (ZS) [4]:
In the upper intertidal zone on open coast [...] ✗
ReflectiVA [13]:
Salt marshes and quiet bays ✗
ReAG (Ours):
Intertidal zone on the open coast ✓

Q: What month of meteorological winter in the northern hemisphere did hungary purchase this building?



Qwen2.5-VL-7B (ZS) [4]:
According to the information provided, the U.S. Civil Service Building was built [...] in 1911 ✗
ReflectiVA [13]:
December ✗
ReAG (Ours):
Third ✓

Q: Who was the duke of milan when Edward Cowper began leading the monks at this abbey?



Qwen2.5-VL-7B (ZS) [4]:
The Duke of Brittany, William de Lis, was the owner of the patronage of Roche Abbey ✗
ReflectiVA [13]:
Charles I ✗
ReAG (Ours):
Maximilian Sforza ✓

Q: What is the religious order of this monastery?



Qwen2.5-VL-7B (ZS) [4]:
Bremian Knighthood ✗
ReflectiVA [13]:
Premonstratensian ✗
ReAG (Ours):
Order of the Holy Cross ✓

Q: In which country or region does this animal live?



Qwen2.5-VL-7B (ZS) [4]:
It is native to southern Canada, the contiguous USA, and Mexico ✗
ReflectiVA [13]:
Eurasia ✗
ReAG (Ours):
North America ✓

Q: What did Alexander Wilson do to this bird in 1811?



Qwen2.5-VL-7B (ZS) [4]:
Alexander Wilson did not harm the bird in any way [...] ✗
ReflectiVA [13]:
I don't see any information about a bird species ✗
ReAG (Ours):
Discovered ✓

Q: What type of habitat does this plant prefer?



Qwen2.5-VL-7B (ZS) [4]:
It cannot be determined from the given information alone, but it is likely to [...] ✗
ReflectiVA [13]:
Dry ✗
ReAG (Ours):
Chaparral ✓

Q: Which teams played against each other in the first match ever held in this stadium?



Qwen2.5-VL-7B (ZS) [4]:
The first match ever held in FirstEnergy Stadium, [...] ✗
ReflectiVA [13]:
Minnesota vikings ✗
ReAG (Ours):
Browns and Minnesota Vikings ✓

Q: What wars ended the year the fichtelberg cable car was built on this mountain?



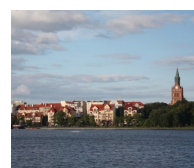
Qwen2.5-VL-7B (ZS) [4]:
Not enough information is provided to determine [...] ✗
ReflectiVA [13]:
World war I ✗
ReAG (Ours):
The American Indian Wars ✓

Q: For what football cup final is this stadium the venue?



Qwen2.5-VL-7B (ZS) [4]:
The football cup final for which this stadium is the venue is the Taça de Portugal ✗
ReflectiVA [13]:
Uefa women's championship ✗
ReAG (Ours):
Portuguese ✓

Q: What is the surface area of this lake?



Qwen2.5-VL-7B (ZS) [4]:
Ełk Lake is a fresh water lake in the Masurian Lake District of Poland's Warmia-Mazury Province ✗
ReflectiVA [13]:
3.82 ✗
ReAG (Ours):
3.82 square kilometres ✓

Q: How many eggs does this bird typically lay?



Qwen2.5-VL-7B (ZS) [4]:
Three eggs ✗
ReflectiVA [13]:
Three ✗
ReAG (Ours):
Three to four ✓

Figure 10. Qualitative results on InfoSeek and Encyclopedic-VQA image-question pairs comparing ReAG, ReflectiVA [13], and the corresponding zero-shot model.