# Knowledge-based Visual Question Answer with Multimodal Processing, Retrieval and Filtering

**Yuyang Hong**[1,2]*, **Jiaqi Gu**[3]*, **Qi Yang**[1,2], **Lubin Fan**[3]†,

**Yue Wu**[3], **Ying Wang**[2], **Kun Ding**[2]†, **Shiming Xiang**[1,2], **Jieping Ye**[3]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]MAIS, Institute of Automation, Chinese Academy of Sciences
[3]Alibaba Cloud Computing

## Abstract

Knowledge-based visual question answering (KB-VQA) requires visual language models (VLMs) to integrate visual understanding with external knowledge retrieval. Although retrieval-augmented generation (RAG) achieves significant advances in this task by combining knowledge-base querying, it still struggles with the quality of multimodal queries and the relevance of retrieved results. To overcome these challenges, we propose a novel three-stage method, termed Wiki-PRF, including **P**rocessing, **R**etrieval and **F**iltering stages. The processing stage dynamically invokes visual tools to extract precise multimodal information for retrieval. The retrieval stage integrates visual and text features to achieve multimodal knowledge retrieval. The filtering stage performs relevance filtering and concentration on retrieval results. To this end, we introduce a visual language model trained with answer accuracy and format consistency as reward signals via a reinforcement learning manner. This enhances the model's reasoning, tool invocation for accurate queries, and filtering of irrelevant content. Experiments on benchmark datasets (E-VQA and InfoSeek) show significant improvements (36.0 and 42.8) in answer quality, achieving state-of-the-art performance. Code is available at: `https://github.com/cqu-student/Wiki-PRF`

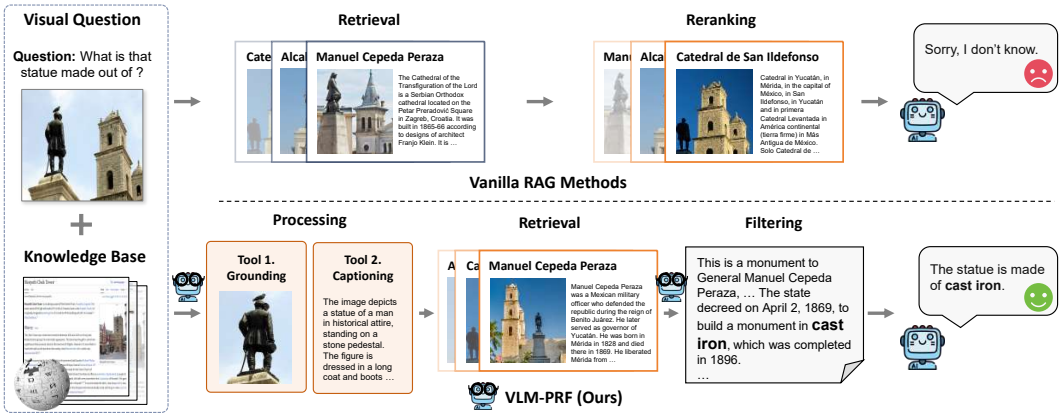



Figure 1: **Illustration of vanilla RAG methods and our Wiki-PRF.** Different from the traditional RAG methods (above), our method (below) employs multimodal tools processing stage and a further filtering stage, enabling more effective retrieval and extraction of task-relevant information.

*Equal contribution: Yuyang Hong <hongyuyang2023@ia.ac.cn>, Jiaqi Gu <vadin@zju.edu.cn>
†Corresponding authors: Lubin Fan <lubin.flb@alibaba-inc.com>, Kun Ding <kun.ding@ia.ac.cn>

# 1 Introduction

Visual Language Models (VLMs) [1, 2, 3] have demonstrated remarkable capabilities in Visual Question Answering (VQA) tasks [4, 5]. Despite their effectiveness, they still face challenges in addressing knowledge-based visual question answering (KB-VQA), as such questions require not only an understanding of visual content but also the integration of external knowledge. For instance, answering the question "What is that statue made of?" in Figure 1 requires factual knowledge that goes beyond the visual content. To address this issue, retrieval-augmented generation (RAG) methods enhance model performance by incorporating mechanisms to access and integrate external information. These methods typically involve three steps: first, retrieving knowledge based on the given visual or textual content; second, reranking the retrieved information for relevance; and third, generating answers using the most pertinent content.

However, current methods [6, 7, 8, 9, 10, 11] often fail to retrieve the most relevant information when handling complex visual content and questions, leading to suboptimal answers. This issue primarily arises from two key challenges. (1) Fine-grained knowledge retrieval in complex visual scenes: When process information-rich images, existing retrieval methods that rely on full-image visual features are often insufficiently precise for effective knowledge retrieval. For example, when asking about statues near a bell tower in Figure 1, the statue may occupy only a small portion of the image. Consequently, the retrieval process can be heavily influenced by the more prominent bell tower, introducing excessive irrelevant information related to it. (2) Precise filtering of irrelevant information from large scale retrieved results: After retrieving contextual information, it is difficult to filter out irrelevant or low-quality content using paragraph reranking alone. The retrieved results typically contain significant amounts of extraneous information, which can affect the accuracy of the generated answers.

To address these challenges, we propose a novel multimodal RAG method consisting of three stages: processing, retrieval, and filtering. The core idea is to obtain more relevant knowledge to generate accurate answers. For coarse retrieval limitations, we innovatively explored a tool-based fine-grained retrieval mechanism. For irrelevant information, we innovatively employ a question-based filtering stage. Specifically, in the processing stage, the visual language model autonomically invokes image-processing tools based on the input image and question. These tools perform operations such as image captioning, visual grounding, and image flipping to extract detailed information related to the question from the image, thereby generating high-quality multimodal retrieval inputs. In the retrieval stage, multimodal retrieval is conducted using both visual features and text descriptions to retrieve relevant knowledge. In the filtering stage, the retrieved contextual information is filtered and condensed to remove redundancy, extract the most relevant knowledge, and provide to the answer generator for generating accurate responses. To this end, we introduce Wiki-PRF, a RAG method that not only supports basic multimodal question-answering functionality but also enhances reasoning based on the input image and question. Wiki-PRF can flexibly invoke visual tools and demonstrates stronger capabilities in filtering and condensing retrieval results.

To enable the visual language model to possess the aforementioned reasoning ability, we train a VLM-PRF model using reinforcement learning (RL). This is because training data collected for complex visual question-answering tasks often lacks the intermediate reasoning steps, which are necessary for effective supervised fine-tuning of VLMs. RL [12, 13, 14], as a paradigm for learning strategies to achieve specific goals, has been widely adopted in recent years to enhance the reasoning capabilities of VLMs [15, 16, 17, 18, 19, 20] for specialized tasks. RL can utilize a small amount of sample data, relying solely on answer accuracy as the reward signal, to train the model to generate high-quality retrieval content by accurately invoking task-specific tools. Additionally, it enables the model to selectively retain and condense the most relevant retrieval results for the query. Specifically, we employ the LoRA [21] to train only a small number of additional parameters, enabling our Wiki-PRF to enhance its RAG capabilities without compromising its core question-answering abilities. In summary, our main contributions are as follows:

- A knowledge-based visual question-answering method using a Processing-Retrieval-Filtering framework is proposed, named Wiki-PRF. It effectively leverages external tools for information retrieval and systematically filters the retrieved knowledge to support the generation of precise answers.

- We introduce VLM-PRF, a visual language model for multimodal RAG tasks, trained via reinforcement learning to enhance reasoning. To our knowledge, this represents the first application of reinforcement learning to multimodal retrieval-augmented generation, requiring minimal training data while enabling flexible tool use and robust processing.
- Comprehensive experiments demonstrate that Wiki-PRF achieves state-of-the-art performance on E-VQA (36.0) and InfoSeek (42.8). Additional analyses further validate our method's effectiveness.

## 2 Related Work

### 2.1 Knowledge-based Visual Question Answering

Knowledge-Based Visual Question Answering (Knowledge-Based VQA) [22] As a critical branch of Visual Question Answering (VQA), Knowledge-Based VQA demands models to integrate the understanding of visual content and question with external knowledge bases for reasoning and answering. Based on knowledge base modalities, Knowledge-Based VQA frameworks can be categorized into unimodal [23, 24, 25, 26, 6, 7] and multimodal [27, 28, 8, 9, 10, 11] paradigms. Unimodal methods [26, 6, 7] typically utilize text-only datasets such as Wiki-21M[29] and GS112K[30] as external knowledge sources. For unimodal methods, TRiG[7] facilitates knowledge passage retrieval and generative question answering by converting images into plain text, thus fully harnessing the power of large-scale knowledge bases and pre-trained language models.
For multimodal methods [8, 9, 10, 11, 31], external knowledge bases typically incorporate datasets such as Encyclopedic VQA (E-VQA)[27] and InfoSeek[28], which include both Wikipedia images and corresponding textual information. MuKEA[11] represents multimodal knowledge through explicit triplets to capture the implicit relationships between visual objects and factual answers. EchoSight[9] first retrieves candidate Wikipedia articles using visual information, then re-ranks them based on text-image query relevance to improve retrieval performance. Unlike previous methods, our method enhances the utilization of external knowledge bases by enabling the model to autonomously select and filter relevant information during the retrieval processing.

### 2.2 Reinforcement Learning for Visual Language Model

Reinforcement learning (RL) [12, 13, 14], a learning paradigm that improves model decision-making through interaction with an environment and feedback in the form of rewards, has recently been widely applied to vision-language models [32, 2, 33, 34, 35] (VLMs). Some works [15, 16, 17, 18, 19, 20] focus on enhancing the reasoning capabilities of Vision-Language Models (VLMs) through reinforcement learning. R1-OneVision[20] framework innovatively bridges vision and language by encoding visual data into formalized textual representations, enabling robust and precise reasoning grounded in linguistic semantics. VisualThinker-R1-Zero[16] achieves the first successful realization of an 'Aha Moment' in multimodal reasoning using a 2B-parameter VLM. Other works[36, 37, 38, 39, 40] focus on leveraging reinforcement learning to improve the performance of Vision-Language Models in specific areas like mathematical reasoning and visual perception. Visual-RFT [39] leverages VLMs to generate reasoning-enhanced responses and integrates task-specific verifiable rewards (e.g., IoU for detection) via policy optimization methods like GRPO [41], improving model performance. In contrast to above work, our study uniquely introduces RAG capabilities into VLMs via RL. As far as we are aware, this is the first exploration of RL-based method for RAG in VLM.

## 3 Method

### 3.1 Overview

Knowledge-based VQA requires answering question $Q$ that is highly relevant to a given reference image $I$, with the assistance of a knowledge base KB. In our setup, KB $\in \{(a_1, I_1), ..., (a_n, I_n)\}$ consists of a million-scale collection of entity articles $\{a_i\}$ along with their corresponding image set $\{I_i\}$. Our goal is to improve multimodal retrieval quality by flexibly invoking visual tools and enhancing relevance through filtering and enrichment.

As illustrated in Figure 2, the overall architecture of Wiki-PRF consists of three key components: an external knowledge base (KB), a model (VLM-PRF) trained via reinforcement learning, and a

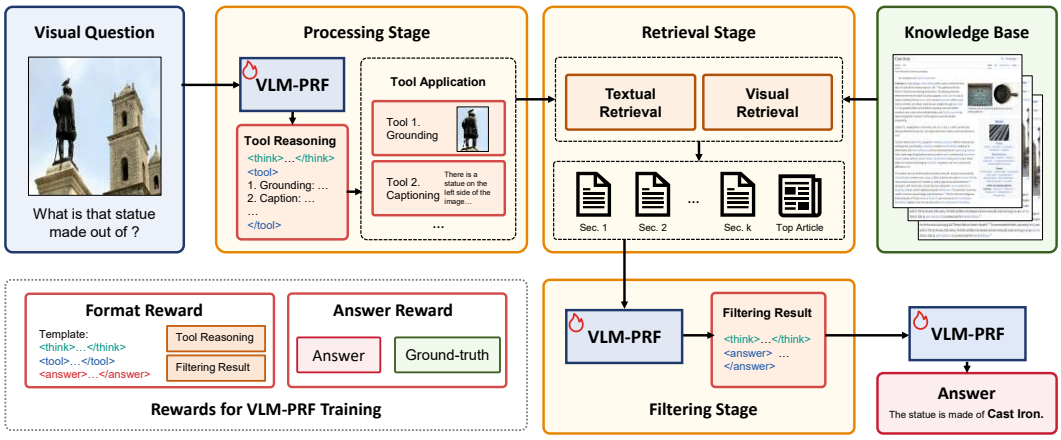



Figure 2: **Overview of Wiki-PRF**. Wiki-PRF comprises three key stages: (1) Processing Stage: VLM-PRF processes the input image and its corresponding question using external tools. (2) Retrieval Stage: Relevant Wikipedia articles are retrieved, split into individual sections, and ranked based on their similarity to the processed input. (3) Filtering Stage: The re-ranked article sections are further filtered by VLM-PRF to retain the most relevant content, which is then fed into the VLM for final answer generation. During training, VLM-PRF is supervised using two types of reward signals: answer reward , which evaluates the correctness of the generated answer, and format reward , which ensures the output adheres to the desired structure.

base model without extra trained parameters. The Wiki-PRF method comprises three main stages: (1) Processing Stage: The VLM-PRF model invokes external tools to process the raw reference image $I$ and question $Q$, generating precise retrieval queries $Query$. (2) Multimodal Retrieval Stage: The model performs multimodal information retrieval based on the generated query $Query$ and retrieves relevant information from the knowledge base. (3) Filtering Stage: The VLM-PRF model filters and extracts highly relevant information from the retrieval results and structures it into task-oriented knowledge, which is subsequently utilized to augment the answer.

### 3.2 Processing Stage

Previous methods [27, 28, 8, 9, 10, 11] rely on raw input for retrieval, often missing key details due to a lack of interactive processing. For instance, a statue next to a church might be overlooked in favor of the church itself. To address this, our method introduces tool-based preprocessing before retrieval, enhancing results through secondary data refinement. We employ several representative tools: 1) A captioning tool that captures high-level semantic information from images. 2) A grounding tool that extracts regions of interest for precise, detailed retrieval. 3) A flipping tool that adjusts the images orientation to mitigate the impact of direction on retrieval. Through these tools, Wiki-PRF achieves more comprehensive and accurate retrieval results. In essence, VLM-PRF provides the strategy while VLM-base delivers the core tool functionality, for captioning and grounding, which we define as $\text{VLM}_{\text{captioning}}$ and $\text{VLM}_{\text{grounding}}$, respectively.

As in Figure 3, given an image $I$ and a question $Q$, the VLM-PRF model reasons about which tools to use and in what order within <think> tags, then outputs selected tools and their execution order in <tool> tags. After VLM-PRF plans the sequence of tool calls, the tasks are executed by VLM-base, a foundational model (Qwen2.5-VL-7B). This model is invoked multiple times to power specific tools like captioning and grounding. For the captioning tool, $\text{VLM}_{\text{captioning}}$ takes the init caption $C_{init}$ generated by VLM-PRF as input and produces the final caption $C_{query}$ for retrieval:

$$C_{query} = \text{VLM}_{\text{captioning}}(C_{init}, Q), \tag{1}$$

Specifically, VLM-PRF first outputs the $C_{init}$ to be processed by the captioning tool $\text{VLM}_{\text{captioning}}$. Then $\text{VLM}_{\text{captioning}}$ employs $C_{init}$ as input to generate the final query $C_{query}$.

For grounding tool, $\text{VLM}_{\text{grounding}}$ takes the object output by VLM-PRF and returns the positional information. The image $I$ is then corpped based on positional information and generate $I_{\text{grounding}}$:

$$I_{\text{grounding}} = \text{Crop}(I, \text{VLM}_{\text{grounding}}(\text{object})). \tag{2}$$

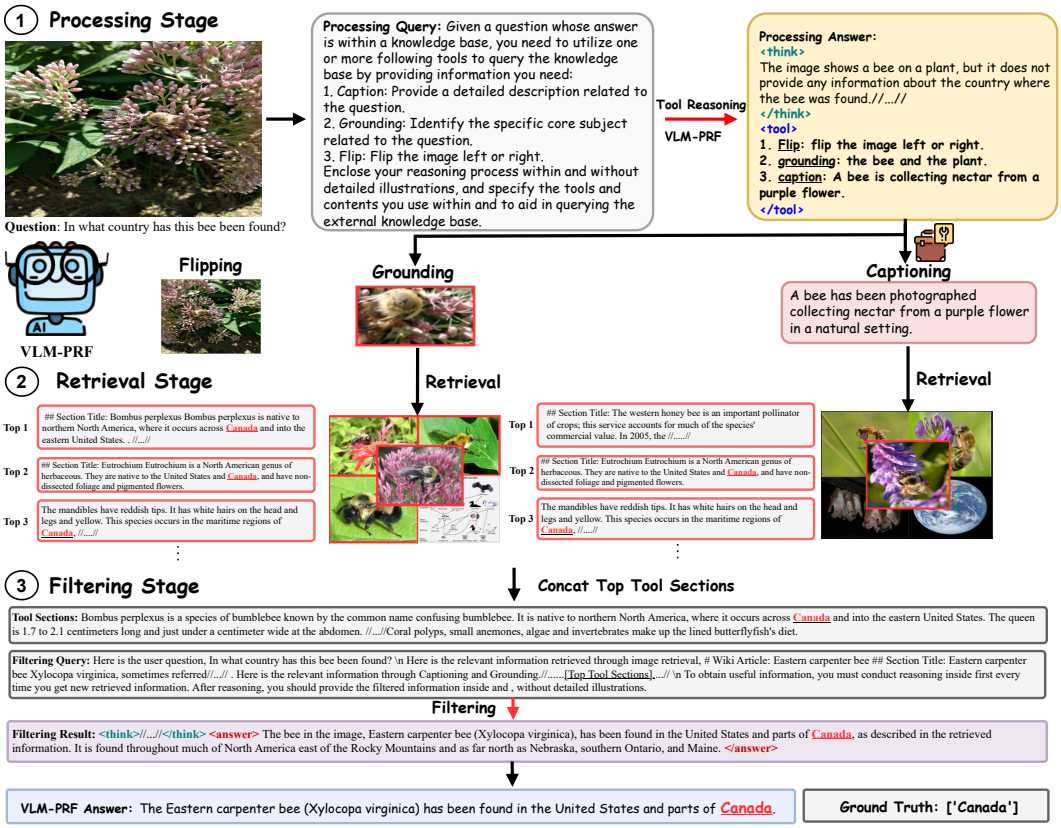



Figure 3: **Example of the tool calls and filtering**. By analyzing the problem, VLM-PRF performs captioning, grounding and flipping operations on the images. Using the retrieved sections, VLM-PRF performs filtering and generates the task-oriented results.

The flipping tool applies a left-right inversion to reference image in order to mitigate the impact of angular variations on the retrieval performance. Finally, all $Query$ results generated by the tools are aggregated to perform refined and accurate retrieval.

### 3.3 Multimodal Retrieval Stage

The goal of multimodal retrieval is to retrieve relevant articles $a$ from a knowledge database $KB$, based on the reference image $I$ and the generated $Query$. Initially, retrieval is performed directly on $I$, and the most relevant article obtained from this process is selected as the base retrieval information, denoted as $D$. Leveraging the $Query$ produced by the tools, additional information is retrieved from the knowledge base to enrich the search context.

**Tool Search.** The retrieval query $Query$ is first embedded into a feature vector using EVA-CLIP [42]. We then employ the Faiss library [43] with cosine similarity $S^D_{\text{tool}}$ to efficiently retrieve the top-k most relevant images and their corresponding documents from the knowledge base:

$$S^D_{\text{tool}} = \max_k \left\{ \left\langle \frac{\boldsymbol{T}}{\|\boldsymbol{T}\|} \cdot \frac{\boldsymbol{V}_i}{\|\boldsymbol{V}_i\|} \right\rangle, i = 1, \ldots, n \right\}, \tag{3}$$

$$(\mathcal{A}, \mathcal{I})_{\text{tool}} = \{(\mathcal{A}_i, \mathcal{I}_i), i \in S^D_{\text{tool}}\}, \tag{4}$$

$$\boldsymbol{T} = \Phi(Query), \boldsymbol{V}_i = \Phi(I_i), \tag{5}$$

where $(\mathcal{A}, \mathcal{I})$ denotes the top-$k_D$ retrieved articles and their associated images, $\Phi$ represents the feature extractor of EVA-CLIP, $\boldsymbol{T}$ is the embedding of the $Query = \{C_{\text{init}}, I_{\text{grounding}}\}$, and $\boldsymbol{V}_i$ is the visual embedding of image $I$. The retrieved articles $\mathcal{A}$ are split into sections $\mathcal{S}$, and after removing titles, sections will be selected based on cosine similarity $S^s_{\text{tool}}$. For captioning, we calculate the

cosine similarity between $Query$ and sections. For grounding, we directly calculate the cosine similarity between the question embedding, $T' = \Phi(Q)$ and sections to maximize the fusion of modal information. The top-$k_s$ most similar sections are then selected as the final retrieval results, where $s_i^j$ denotes the $j$-th section of the $i$-th article:

$$\mathcal{A}_i = \{s_1^i, s_2^i, ..., s_m^i\}, i = 1, \ldots, k_s, \tag{6}$$

$$S^s_{\text{tool}} = \max_{k_s} \left\{ \left\langle \frac{\boldsymbol{T'}}{\|\boldsymbol{T'}\|} \cdot \frac{\Phi(\boldsymbol{s}_i^j)}{\|\Phi(\boldsymbol{s}_i^j)\|} \right\rangle \middle| \, i = 1, \ldots, m; \; j = 1, \ldots, k_s \right\}. \tag{7}$$

By indexing $S^s_{\text{tool}}$, we obtain the sections $\mathcal{S}_{\text{tool}}$ returned by the corresponding tool and concatenate them to form the final search result $\mathcal{S}_{\text{search}}$. Subsequently, $\mathcal{S}_{\text{search}}$ is fed into VLM-PRF along with the top-$k$ article $D$, which is retrieved by directly searching the input image. Following the specification in Figure 3, $\mathcal{S}_{\text{search}}$ is filled in <search_result>, $D$ is filled in <retrieved_information>, where //...// represents long text.

### 3.4 Filtering Stage

During the retrieval processing, a large amount of redundant information is generated, with only a small fraction containing key details relevant to answering the question. Previous approaches [9, 31] attempt to mitigate this by reranking and selecting more relevant passages. However, article- or section-level reranking methods can only filter at the passage level, often retaining significant noise. To address this limitation, we propose training the model using reinforcement learning, guided by answer accuracy. This approach enables the model to filter retrieval results in a question-specific manner, reducing the influence of irrelevant content. Specifically, Wiki-PRF guides VLM-PRF to process the directly retrieved information $D$, comprising both image-derived data from $I$ and search results $\mathcal{S}_{\text{search}}$, and output its reasoning within <think> and </think> tags. The model then generates a compact, task-oriented knowledge representation $F$ within <answer> and </answer> tags.

$$F = \text{VLM-PRF}(D, \mathcal{S}_{\text{search}}), \tag{8}$$

$$A = \text{VLM}(F, Q), \tag{9}$$

where $D$ denotes the retrieved information corresponding to $I$, $\mathcal{S}_{\text{search}}$ represents the external search results obtained via tool-based retrieval, and $F$ is the filtered, task-oriented knowledge representation produced by the reinforcement learning module. After generating task-oriented knowledge $F$, Wiki-PRF uses context to generate the final answer A via the VLM.

### 3.5 Training via Reinforcement Learning

To improve the model's tool selection and information filtering strategies, we utilize GRPO [41] with removed KL divergence constraint for VLM-PRF training. The formula can be defined as:

$$\mathcal{J}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \left[ \frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min\left( r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{i,t} \right) \right], \tag{10}$$

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q, o_{i,<t}; R(q))}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t}; R(q))}, \tag{11}$$

where $\{o_i\}_{i=1}^{G}$ denotes the $G$ responses generated for question $Q$, and $|o_i|$ is the length of the $i$-th response. The term $\pi_\theta(o_{i,t}|q, o_{i,<t}; R(q))$ represents the conditional probability of token $o_{i,t}$ at decoding step $t$, conditioned on previous tokens $o_{i,<t}$ and retrieved information $R(q)$.

Besides, we design a reward function as the primary supervisory signal to guide the model in enhancing its tool invocation and its filtering of retrieved information. Specifically, we employ a format-based reward to encourage VLM-PRF to perform multi-step reasoning about tool usage within the <think> tags, make appropriate tool calls within the <tool> tags, and further process the retrieved information within the <think> tags. Finally, the refined and filtered results are output within the <answer> tags. Furthermore, we introduce an answer reward to supervise the content generated within the <answer> tags, ensuring that the model produces high-quality, relevant, and well-structured responses. The final reward function can be presented by:

$$r_\phi(x, y) = \alpha EM(a_{\text{pred}}, a_{\text{gt}}) + \beta M(a_{\text{tool}}, t_{\text{tool}}) + \gamma M(a_{\text{filter}}, t_{\text{filter}}), \tag{12}$$

$$M(x, y) = \begin{cases} 1 & \text{if match,} \\ 0 & \text{if unmatch,} \end{cases} \tag{13}$$

where $r_\phi(x, y)$ represents the reward between input $x$ and output $y$; $EM$ denotes the evaluation function for answers, such as exact matching; $M$ employs regular expression matching to verify format compliance. $\alpha$, $\beta$, and $\gamma$ are the weighting coefficients for these components, with values of 1, 0.3, and 0.7 respectively in our method. Moreover, $a_{\text{pred}}$ refers to the model's output answer, $a_{\text{gt}}$ represents the ground truth, while $a_{\text{tool}}$ and $a_{\text{filter}}$ correspond to the model's outputs during the processing and filtering stages. $t_{\text{tool}}$ and $t_{\text{filter}}$ represent the templates for tool usage and filtering.

# 4 Experiments

## 4.1 Experimental Setup

**Datasets.** We evaluated our experimental results on two main datasets: InfoSeek [6] and Encyclopedic VQA (E-VQA)[26]. (1) InfoSeek[6] contains 1.3M VQA pairs matched to 11K images from OVEN[44]. The training set (934K) and validation set (73K) are strictly divided by both entities and questions. The validation set is further categorized into two types: Unseen Entity and Unseen Question. Following the setup of [9], we used a knowledge base consisting of 100K Wikipedia entries and reported evaluation results on the entire validation split. (2) E-VQA [26] consists of over 221K unique question-answer pairs, each associated with up to five images sourced from iNaturalist [45] and the Google Landmarks Dataset v2 [46]. The dataset includes two types of questions: Single-hop and Two-hop. The samples are divided into training, validation, and test sets with 1M, 13.6K, and 5.8K items respectively. Like other methods [9, 31], we report our results on the 5.8K test set. We also evaluate OK-VQA [23], a 14K-question dataset spanning diverseknowledge domains.

**Baselines.** To validate the effectiveness of our Wiki-PRF, we establish two baselines: (1) Base Model: Directly answer questions without any RAG pipelines. (2) Wiki-PRF without RL: Answer questions with our Wiki-PRF method before RL fine-tuning. These baselines serve to assess the contributions of our three-stage Wiki-PRF design and the benefits of RL fine-tuning, respectively.

**Evaluation Metrics.** In the KB-VQA task, we focus on evaluating both retrieval and QA metrics. For the retrieval metric, we use recall to determine whether the correct article appears among the top-k retrieved results. For the QA metric, following the original dataset settings, we apply VQA accuracy[5, 23] for InfoSeek and BEM score[47] for E-VQA.

**Implementation Details.** Given the strong vision-language understanding capabilities of the Qwen2.5-VL series, we adopt Qwen2.5-VL-3B and Qwen2.5-VL-7B as our base models. We apply GRPO for reinforcement learning. Specifically, we set the number of generations to 8, the sampling temperature to 0.7, the number of training epochs to 2, and the learning rate to $1e$-5. We utilize LoRA-based fine-tuning, with the LoRA rank to 64, the LoRA alpha to 128, and the dropout rate to 0.05. For the retriever, we use a frozen EVA-CLIP 8B model for both retrieval and similarity computation. Image features are indexed and retrieved using cosine similarity with the Faiss-GPU library. By default, we use the Top-1 article retrieved from image retrieval alongside the Top-5 articles identified through tool calls. All sections from the image-retrieved article are retained, while the top-k sections from each of the tool-retrieved articles are kept. This combined information is then provided to the next stage for filtering. The total training process takes approximately 15 hours using 8 A800 GPUs. Our framework, Wiki-PRF, is implemented in two configurations: **Wiki-PRF-3B** and **Wiki-PRF-7B**. Models further fine-tuned using reinforcement learning are referred to as **VLM-PRF-3B** and **VLM-PRF-7B**.

## 4.2 Main Results

**VQA Results.** We assess our method on the aforementioned VQA datasets, comparing it against various MLLMs and retrieval-augmented approaches as shown in Table 1. Among zero-shot MLLM approaches, we found that without RAG, the MLLM model achieves only relatively low accuracy in answering questions, highlighting the challenges presented by the KB-VQA task. Regarding retrieval-augmented methods, our Wiki-PRF before RL training, achieves an accuracy of 34.0 with the 3B model and 39.5 with the 7B model in InfoSeek dataset, surpassing several well-trained methods such as EchoSight [9] and Wiki-LLaVA [52]. This finding further confirms the effectiveness of our proposed three-stage approach. After RL training, our Wiki-PRF-7B establishes a new state-of-the-art

| Method | Model | Retriever | E-VQA | | InfoSeek | | |
|---|---|---|---|---|---|---|---|
| | | | Single-Hop | All | Unseen-Q | Unseen-E | All |
| *Zero-shot MLLMs* | | | | | | | |
| BLIP-2 [34] | Flan-T5$_{XL}$ | - | 12.6 | 12.4 | 12.7 | 12.3 | 12.5 |
| InstructBLIP [48] | Flan-T5$_{XL}$ | - | 11.9 | 12.0 | 8.9 | 7.4 | 8.1 |
| LLaVA-v1.5 [49] | Vicuna-7B | - | 16.3 | 16.9 | 9.6 | 9.4 | 9.5 |
| GPT-4V [1] | - | - | 26.9 | 28.1 | 15.0 | 14.3 | 14.6 |
| Qwen2.5-VL-3B (Base) [2] | - | - | 17.9 | 19.6 | 20.4 | 21.9 | 21.4 |
| Qwen2.5-VL-7B (Base) [2] | - | - | 21.7 | 20.3 | 22.8 | 24.1 | 23.7 |
| *Retrieval-Augmented Models* | | | | | | | |
| DPR$_{V+T}$ [50]† | Multi-passage BERT | CLIP ViT-B/32 | 29.1 | - | - | - | 12.4 |
| RORA-VLM [51]† | Vicuna-7B | CLIP+Google Search | - | 20.3 | 25.1 | 27.3 | - |
| EchoSight [9]† | Mistral-7B/LLaMA-3-8B | EVA-CLIP-8B | 19.4 | - | - | - | 27.7 |
| Wiki-LLaVA [52] | Vicuna-7B | CLIP ViT-L/14+Contriever | 17.7 | 20.3 | 30.1 | 27.8 | 28.9 |
| ReflectiVA [31] | LLaMA-3.1-8B | EVA-CLIP-8B | 28.0 | 29.2 | 40.4 | 39.8 | 40.1 |
| MMKB-RAG [53] | Qwen2-7B | EVA-CLIP-8B | **39.7** | 35.9 | 36.4 | 36.3 | 36.4 |
| VLM-PRF (w/o RL) | Qwen-2.5VL-3B | EVA-CLIP-8B | 26.6 | 25.6 | 34.2 | 33.7 | 34.0 |
| VLM-PRF (w/o RL) | Qwen-2.5VL-7B | EVA-CLIP-8B | 28.9 | 28.6 | 40.0 | 39.4 | 39.5 |
| *Retrieval-Augmented Models with Reinforcement Learning* | | | | | | | |
| VLM-PRF (**Ours**) | LLaMA-3.1-8B | EVA-CLIP-8B | 36.3 | 35.5 | 41.3 | 40.6 | 40.8 |
| VLM-PRF (**Ours**) | Qwen-2.5VL-3B | EVA-CLIP-8B | 31.1 | 32.4 | 39.7 | 38.8 | 39.0 |
| VLM-PRF (**Ours**) | Qwen-2.5VL-7B | EVA-CLIP-8B | 37.1 | 36.0 | 43.3 | **42.7** | **42.8** |
| VLM-PRF (**Ours**) | InternVL3-8B | EVA-CLIP-8B | **40.1** | **39.2** | **43.5** | 42.1 | 42.5 |

Table 1: **VQA accuracy on E-VQA and InfoSeek.** The metrics of baselines and our methods are highlighted in light blue. † indicates results that are not directly comparable due to different knowledge bases.

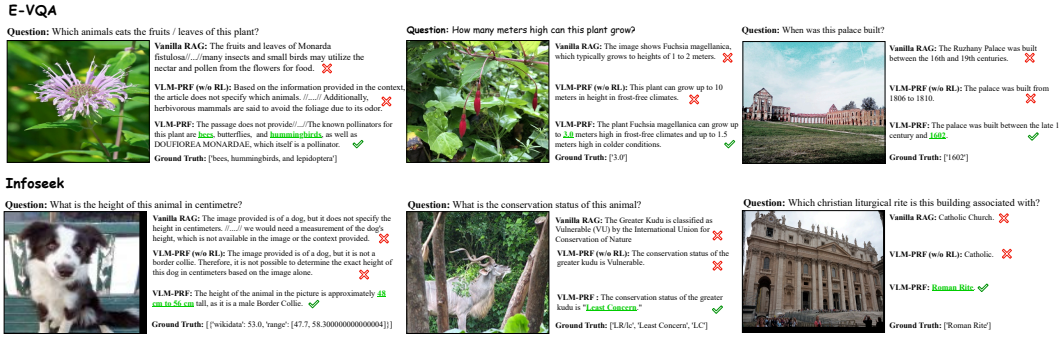



Figure 4: **Qualitaive examples of Wiki-PRF.**

accuracy of 36.0 on E-VQA and 42.8 on InfoSeek, outperforming all previous methods, including ReflectiVA [31] and MMKB-RAG [53]. Moreover, our method achieves consistent improvements across different base models, setting a new SOTA score of 39.2% on E-VQA with InternVL3-8B as the base model. This underscores the impact of reinforcement learning in enhancing the model's RAG capabilities.

Table 2: **Recall of retrieved articles.**

| Model | Retrieval Input | Retrieval Recall |
|---|---|---|
| None | images | 45.56 |
| Qwen2.5-VL-3B | images + tools | 48.32 |
| Qwen2.5-VL-7B | images + tools | 53.44 |
| VLM-PRF-3B | images + tools | 50.16 |
| VLM-PRF-7B | images + tools | 54.89 |

Table 3: **Tool usage statistics.** Statistical analysis of tool usage (mean and variance).

| Model | Combinations | Captioning | Grounding | Flipping |
|---|---|---|---|---|
| Qwen2.5-VL-3B | 34 | 0.86 / 1.10 | 0.40 / 0.50 | 0.04 / 0.22 |
| Qwen2.5-VL-7B | 34 | 2.43 / 1.03 | 0.85 / 0.28 | 0.22 / 0.42 |
| VLM-PRF-3B | 53 | 1.52 / 1.22 | 0.54 / 0.50 | 0.15 / 0.36 |
| VLM-PRF-7B | 40 | 2.43 / 1.13 | 0.93 / 0.36 | 0.26 / 0.44 |

**Effectiveness of the Processing Stage.** Table 2 presents the recall of articles retrieved from InfoSeek under various settings. The Top-1 retrieval performance using direct image retrieval is 45.56%. Combined with Top-5 article retrievals from our tools, this rate increases to 48.32% and 53.44%. Further improvement with reinforcement learning supervision raises it to 50.16% and 54.89%. Notably, we observe that models trained with RL supervision exhibit greater diversity and frequency in tool selection, as shown in Table 3. Specifically, the combinations of tool calls increase after RL training, demonstrating that the model can dynamically and flexibly construct tool invocation schemes. At the same time, the captioning tool is invoked most frequently, highlighting its role as the most common and direct tool for enhancing article recall. Overall, this demonstrates that RL can encourage the model to leverage a broader array of tools by optimizing for the final answer reward, thereby showcasing its flexibility.

Table 4: **Performance on OK-VQA.**

| Method | Model | OK-VQA |
|---|---|---|
| Qwen2.5-VL-3B | - | 62.1 |
| Qwen2.5-VL-7B | - | 72.4 |
| KU-RAG [54] | LLaVA-Next-7B | 73.1 |
| MMKB-RAG [53] | LLaMA-3.1-8B | 65.4 |
| Wiki-PRF-3B | VLM-PRF-3B | **68.6** |
| Wiki-PRF-7B | VLM-PRF-7B | **77.8** |

Table 5: **Filtering from oracles.** VQA Accuracy in Oracle Setting with Ground-Truth Articles.

| Method | Model | VQA Accuracy |
|---|---|---|
| Wiki-LLaVA [52] | Vicuna-7B | 51.5 |
| ReflectiVA [31] | LLaMA-3.1-8B | 57.6 |
| Wiki-PRF-3B (**Ours**) | VLM-PRF-3B | 64.4 |
| Wiki-PRF-7B (**Ours**) | VLM-PRF-7B | 65.8 |

Table 6: **Comparison of SFT and RL.** We sampled 2,000 instances from InfoSeek and present a comparison between the results of SFT and RL.

| Model | Unseen Question (UQ) | Unseen Entity (UE) | ALL |
|---|---|---|---|
| Qwen2.5-VL-7B | 39.1 | 40.5 | 40.2 |
| Wiki-PRF-7B (SFT) | 41.5 | 41.9 | 41.8 |
| Wiki-PRF-7B (RL) | **46.6** | **46.2** | **46.3** |

**Effectiveness of the Filtering Stage.** To evaluate whether our filtering stage can effectively extract accurate information from given sections, we conduct experiments similarly to other methods under an oracle setting, where the ground-truth entity (i.e., the Wikipedia page associated with the query) is provided. Compared with other methods, our final VQA accuracy is much higher (65.8%) as shown in Table 5, the fine-tuned model can more efficiently locate the necessary information when given the oracle retrieval information.

**Effectiveness of Reinforcement Learning.** To investigate its effectiveness, we evaluated the use of supervised fine-tuning (SFT) for the filtering stage. Specifically, we trained a dedicated filtering model using SFT, keeping all other configurations identical to reinforcement learning (RL). The results are presented in Table 6. The RL model significantly outperforms the SFT model on the test set. The reason is that SFT tends to imitate superficial patterns, which limits its generalization capability. In contrast, RL enables the model to understand the underlying principles of information filtering, leading to a far more robust and generalizable performance.

**Results on more benchmarks.** As shown in Table 4, we evaluate our model on the widely used OK-VQA benchmark. We can see that our Wiki-PRF-7B achieves a new state-of-the-art score of 77.8 on OK-VQA. The consistent performance improvement on multiple benchmarks confirms our method's strong generalization capability.

### 4.3 Ablation Studies

**Modules and Tools.** Table 8 presents the ablation study of our carefully designed stages and tools based on 10K samples from the InfoSeek validation set. In the module ablation, VLM-PRF-3B model improves the baseline by 2.54% and 2.02% when combining the processing stage and filtering stage, respectively, underscoring the effectiveness of these two modules. In the tool ablation, the caption and grounding tools increase performance by 1.94% and 0.98%, respectively. The combined use of all tools yields a peak performance of 39.48%. This result demonstrates that Wiki-PRF significantly boost the final VQA accuracy by adeptly utilizing tools to provide more relevant information.

Table 7: **Ablation studies on the scale of knowledge base on InfoSeek.**

| Method | Model | 10k | 50k | 100k |
|---|---|---|---|---|
| Vanilla-MRAG | Qwen2.5-VL-3B | 49.7 | 32.1 | 21.4 |
| Wiki-PRF-3B | VLM-PRF-3B | 53.0 | 43.7 | 39.0 |
| Vanilla-MRAG | Qwen2.5-VL-7B | 56.3 | 39.6 | 23.7 |
| Wiki-PRF-7B | VLM-PRF-7B | 60.3 | 51.2 | 42.8 |

**Scale of Knowledge Base.** In Table 7, we verfied that the correct document for each evaluation question was included in knowledge bases of all scales, to assess the impact of knowledge base size. The result demonstrates that both our method and the baseline exhibit performance degradation as knowledge base size increases. This occurs because larger knowledge bases introduce additional noise, increasing retrieval difficulty, which is a universal challenge for current RAG methods. Critically, Wiki-PRF demonstrates a significantly slower rate of degradation for both 3B and 7B models.

Table 8: **Modules and tools ablation.** "Processing", "Retrieval" and "Filtering" denote the three distinct stages proposed in our approach.

| Model | Processing | Retrieval | Filtering | Tools | VQA Accuracy |
|---|---|---|---|---|---|
| *Modules Ablation* | | | | | |
| Qwen2.5-VL-3B | | ✔ | | - | 34.22 |
| VLM-PRF-3B | | ✔ | ✔ | - | 36.24 |
| VLM-PRF-3B | ✔ | ✔ | | Multi-Tools | 36.76 |
| VLM-PRF-3B | ✔ | ✔ | ✔ | Multi-Tools | 39.48 |
| *Tools Ablation* | | | | | |
| VLM-PRF-3B | ✔ | ✔ | ✔ | Grounding | 37.22 |
| VLM-PRF-3B | ✔ | ✔ | ✔ | Captioning | 38.18 |
| VLM-PRF-3B | ✔ | ✔ | ✔ | Multi-Tools | 39.48 |

Table 9: **Ablation on training sample size.**

| VLM Model | 2K | 4K | 6K | 8K |
|---|---|---|---|---|
| VLM-PRF-3B | 37.30% | 39.48% | 38.92% | 40.83% |
| VLM-PRF-7B | 42.13% | 43.10% | 43.09% | 43.80% |

Table 10: **Ablation on retrieved article quantity.**

| Retrieved Article | Top-1 | Top-3 | Top-5 | Top-7 |
|---|---|---|---|---|
| InfoSeek | 38.85% | 39.10% | 39.48% | 39.57% |

**Training Samples.** Table 9 presents the accuracy achieved with different training sample sizes. As the number of training samples increases, we observe a general upward trend in accuracy. Balancing accuracy with training time, we opt for 4K samples as our default experimental setting. Unlike other methods that necessitate training on the complete training dataset (InfoSeek: 934K, E-VQA: 1M), our approach attains comparable or even superior results using only a small subset of samples. We attribute this efficiency to reinforcement learning's ability to effectively stimulate the model to leverage tools for retrieving additional information and integrating existing knowledge to determine the correct answers, rather than relying on memorizing specific question-answer patterns.

**Retrieved Articles.** Table 10 illustrates the ablation of retrieving top-k articles when utilizing tools. We evaluated the VQA accuracy of Wiki-PRF-3B on the InfoSeek dataset with varying numbers of retrieved articles: 1, 3, 5, and 7. The results indicate that as the number of retrieved articles increases, VQA accuracy tends to improve gradually, albeit with diminishing returns. To balance inference time and accuracy, we set the K=5 as the optimal choice and utilize the Top-3 relevant sections from these K=5 articles to supplement knowledge. More experiments and details about the retrieval of tool calls are presented in the supplementary materials.

Table 11: **Inference time of each stage.** We sampled 1,000 instances from InfoSeek and measured the average duration of each stage per sample.

| Model | Processing & Retrieval | Filtering | Answering | Total |
|---|---|---|---|---|
| VLM-PRF-3B | 2.2s | 3.4s | 0.59s | 6.23s |
| VLM-PRF-7B | 3.3s | 4.6s | 0.74s | 8.77s |

**Inference Time.** We evaluated the VLM-PRF model's stage-wise time cost per sample in Table 11. The results show that the Processing & Retrieval stages and Filtering stage consume more time than the Answering stage. This duration primarily stems from tool invocation and long-text processing, both of which can be further optimized in future improvements.

## 5 Conclusion

In this paper, we propose Wiki-PRF, a three-stage Process-Retrieval-Filtering framework that represents first reinforcement learning method for multimodal retrieval-augmented generation. By guiding models to invoke tools for processing raw information during the processing stage and filtering retrieved knowledge during the filtering stage, the trained VLM-PRF model significantly enhances performance on Knowledge-Based Visual Question Answering tasks. Extensive experiments demonstrate state-of-the-art results on E-VQA and InfoSeek benchmarks. While limited to three retrieval tools in this study, future work may explore expanded tool integration to further advance capabilities.

## 6 Acknowledgements

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA0480200), the National Natural Science Foundations of China (Grant No.62306310).

## 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.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.

[5] 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 *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913, 2017.

[6] Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*, 2022.

[7] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5067–5077, 2022.

[8] Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5089–5098, 2022.

[9] Yibin Yan and Weidi Xie. Echosight: Advancing visual-language models with wiki knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1538–1551, 2024.

[10] Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. Multi-modal answer validation for knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 36, pages 2712–2721, 2022.

[11] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*, 2022.

[12] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[13] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 30, 2016.

[14] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

[15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[16] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025.

[17] F Meng, L Du, Z Liu, Z Zhou, Q Lu, D Fu, B Shi, W Wang, J He, K Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[18] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *CoRR*, 2024.

[19] Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 reinforce samples, get a baseline for free! *International Conference on Learning Representations Work Shop (ICLR work shop)*, 2019.

[20] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1 (2):3, 2022.

[22] Jiaqi Deng, Zonghan Wu, Huan Huo, and Guandong Xu. A comprehensive survey of knowledge-based vision question answering systems: The lifecycle of knowledge in visual reasoning task. *arXiv preprint arXiv:2504.17547*, 2025.

[23] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204, 2019.

[24] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision (ECCV)*, pages 146–162. Springer, 2022.

[25] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 40(10):2413–2427, 2017.

[26] Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems (Neurlps)*, 36:22820–22840, 2023.

[27] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3113–3124, 2023.

[28] 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? *arXiv preprint arXiv:2302.11713*, 2023.

[29] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781, 2020.

[30] Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, 2021.

[31] Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Augmenting multimodal llms with self-reflective tokens for knowledge-based visual question answering. *arXiv preprint arXiv:2411.16863*, 2024.

[32] 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. *Advances in neural information processing systems (Neurlps)*, 35:23716–23736, 2022.

[33] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning (ICML)*, pages 17283–17300. PMLR, 2023.

[34] 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 *International Conference on Machine Learning (ICML)*, pages 19730–19742. PMLR, 2023.

[35] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[36] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[37] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems (Neurlps)*, 37:95095–95169, 2024.

[38] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision (ECCV)*, pages 169–186. Springer, 2024.

[39] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

[40] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

[41] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[42] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[43] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *CoRR*, 2024.

[44] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12065–12075, 2023.

[45] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12884–12893, 2021.

[46] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2575–2584, 2020.

[47] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*, 2019.

[48] 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, 2023. URL `https://arxiv.org/abs/2305.06500`.

[49] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024.

[50] Paul Lerner, Olivier Ferret, and Camille Guinaudeau. Cross-modal retrieval for knowledge-based visual question answering. In *European Conference on Information Retrieval (ECIR)*, pages 421–438. Springer, 2024.

[51] Jingyuan Qi, Zhiyang Xu, Rulin Shao, Yang Chen, dingnan jin, Yu Cheng, Qifan Wang, and Lifu Huang. Rora-vlm: Robust retrieval-augmented vision language models. *ArXiv*, abs/2410.08876, 2024.

[52] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPR workshop)*, pages 1818–1826, 2024.

[53] Zihan Ling, Zhiyao Guo, Yixuan Huang, Yi An, Shuai Xiao, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. Mmkb-rag: A multi-modal knowledge-based retrieval-augmented generation framework. *arXiv preprint arXiv:2504.10074*, 2025.

[54] Zhengxuan Zhang, Yin Wu, Yuyu Luo, and Nan Tang. Fine-grained retrieval-augmented generation for visual question answering. *CoRR*, 2025.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We present the motivation of our method as clearly as possible in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We present the limitations of our approach at the end of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We present our method in sufficient detail in the Methods to allow readers to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: E-VQA and Infoseek are two public datasets we use. We will open source our code when appropriate.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We explain our training and testing numbers in detail in the Experimental Details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We describe the metrics we use and their significance in section metrics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the devices we used and the computation time in the implementation details and further provide the inference time in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, our paper complies with NeurIPS ethical requirements.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We explain our motivation and the benefits to the model in the abstract, which is consistent with the Broader impacts of LLM research.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We use public data sets and do not involve risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We mainly base it on VLM-R1, whose code is licensed under CC-BY 4.0. https://github.com/om-ai-lab/VLM-R1.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We use open source code on github and manage our code using github's open source license.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Studies of retrieval enhancement generation do not involve subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We finetune Qwen2.5-VL 3B/7B in our method.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

This appendix presents additional materials and results. First, we describe the complete workflow of our method in Sec. A to enhance comprehension. Then, we give further descriptions of our prompts in experiments in Sec. B. Next, we provide more ablation studies for Wiki-PRF in Sec. C. Finally, a series of visual results are presented in Sec. D, and the broader impacts are discussed in Sec. E.

## A Workflow of Wiki-PRF

We detail the complete workflow of Wiki-PRF below.

- **Processing Stage**
  - **Query Anysis:** Given the reference image $I$ and question $Q$, Wiki-PRF begins by analyzing the key information needed to solve the problem in <think> and </think> and subsequently specifies the required tools using the <tool> Tool: Content </tool> format.
  - **Tool Calling:** Upon capturing a tool request, Wiki-PRF parses tools enclosed in <tool> and </tool> tags and sequentially executes the corresponding functions.
    - For Captioning, Wiki-PRF feeds the content following the caption to VLM-PRF to generate the retrieval query $Query_{\text{captioning}}$.
    - For Grounding, Wiki-PRF first obtains the object coordinates from VLM-PRF, followed by performing the image cropping operation based on the coordinates. The resulting cropped image is then returned as the retrieval query $Query_{\text{grounding}}$.
    - For Flipping, VLM-PRF directly returns the flipped image $I_{\text{flip}}$.
- **Retrieval Stage:** In the retrieval stage, Wiki-PRF follows a two-step process: it first retrieves the top-k articles $D$ based on the reference image $\mathcal{I}$, and then conducts further searches using the queries returned by the tools.
  - **Captioning Search:** Given $Query_{\text{captioning}}$, Wiki-PRF initially retrieves the top $k$ most similar images and their associated documents from the knowledge base. These documents are then segmented into sections denoted as $\mathcal{S}_{\text{captioning}}$. Subsequently, Wiki-PRF computes the similarity between $Query_{\text{captioning}}$ and each section in $\mathcal{S}_{\text{captioning}}$, and selects the top-$k_s$ most relevant sections as the final retrieval results.
  - **Grounding Search:** Given $Query_{\text{grounding}}$, same as Captioning Search, Wiki-PRF follows a procedure similar to that of captioning search by first retrieving the sections $\mathcal{S}_{\text{grounding}}$. The key difference lies in the subsequent step, where the Wiki-PRF computes the similarity between the question $Q$ and each section in $\mathcal{S}_{\text{grounding}}$. Finally, top-$k_s$ sections are selected as the retrieval results.
  - **Constructing Search Result:** Wiki-PRF takes the union of all retrieval results, and then concatenates the sections in the union as $\mathcal{S}_{\text{search}}$.
- **Filtering Stage:** Given the documents $D$ and the sections $\mathcal{S}_{\text{search}}$, Wiki-PRF leverages VLM-PRF to filter relevant information guided by the reference image $I$ and question $Q$. The reasoning process of VLM-PRF is presented within <think> and </think>, while the resulting task-oriented knowledge $F$ is output within <answer> and </answer>.
- **Answering:** With the task-oriented knowledge $F$, Wiki-PRF generates the final answer $A$.

# B Prompts Details in Wiki-PRF

## B.1 Processing Stage

### Prompt for Tool Calling:

USER: Given a question whose answer is within a knowledge base, you need to utilize one or more following tools to query the knowledge base by providing information you need: \'caption\': Provide a detailed description related to the question, and the information will be used to query the external knowledge base to retrieve relevant knowledge points. \'grounding\': Identify the specific core subject related to the question and it will return concrete details about the area. \'Flip\': Flip the image left or right. Enclose your reasoning process within <think> and </think> without detailed illustrations, and specify the tools and contents you use within <tool> and </tool> to aid in querying the external knowledge base. Example: <think>reasoning process</think> <tool>
1. Flip: Flip left.
2. grounding: The panda on the tree.
3. caption: A panda is climbing the tree with a bird beside it.
</tool> Here is the user question, {Question}.

### Prompt for Captioning:

USER: Here is the question, {Question}. Here is the caption, {Caption}. describe the image in the context of the question and the caption."

### Prompt for Grounding:

USER: "Locate {object}, output its bbox coordinates using JSON format."

## B.2 Filtering Stage

USER: "Here is the user question, <question> {Question} </question>. Here is the relevant information retrieved through image retrieval, <retrieved_information> {Document} </retrieved_information>. Here is the relevant information through <tool>{Search}</tool>, <search_result>{Search_result}</search_result>. To obtain useful information, you must conduct reasoning inside <think> </think> first every time you get new retrieved information. After reasoning, you should provide the filtered information inside <answer> and </answer>, without detailed illustrations."

## B.3 Prompt for Answer

USER: "Here is the question, {Question}. Here is the retrieval information,{Search_results}, short answer:"

# C Additional Experiments

## C.1 Training Loss

In this section, we present the training curve of VLM-PRF-7B under reinforcement learning in E-VQA. Figure 5 displays three key metrics: answer reward, format reward, and task-oriented knowledge tokens. As shown in Figure 5, both the answer reward and format reward exhibit a consistent upward trend, indicating that as the model learns to invoke tools and filter relevant information, its accuracy in answering knowledge-based VQA questions gradually improves. This clearly demonstrates the effectiveness of GRPO in enhancing the model's RAG capabilities.

Moreover, the tokens of the task-oriented knowledge decreases progressively with the number of training steps. This phenomenon suggests that the model becomes increasingly adept at identifying and retaining only the most relevant knowledge during the learning process.

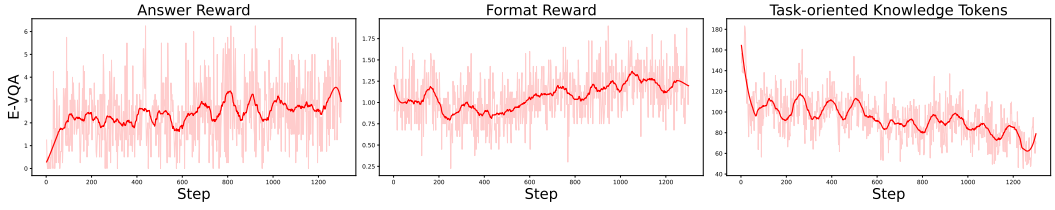



Figure 5: **The training curve of VLM-PRF-7B in E-VQA.**

### C.2 Tool Inference Time

In Table 12, we analyze the execution time of individual tools under varying numbers of recalled articles (i.e., 3, 5, and 7). The results reveal that grounding requires more time than captioning. This can be attributed to the image processing operations involved in grounding, which result in increased computational demands during execution.

Table 12: **Tool Calling Time Per Sample.**

| Model | Recall Numbers | Captioning | Grounding |
|---|---|---|---|
| VLM-PRF-3B | 3 | 0.99s | 1.64s |
| | 5 | 1.07s | 1.69s |
| | 7 | 1.11s | 1.73s |

### C.3 Weights of Rewards

As shown in Table 13, we investigate the influence of varying the weight of answer reward (i.e., $\alpha$) and the weight of format reward (i.e., $\beta + \gamma$) in the overall objective function on the InfoSeek dataset. We fix the values of $\beta$ and $\gamma$ to 1.0, and continuously adjust the ratio of $\alpha$ and $(\beta + \gamma)$. By gradually decreasing $\alpha : (\beta + \gamma)$ from $3 : 1$ to $1 : 3$, we observe that the optimal performance is achieved when both components are equally weighted. Consequently, in our experiments, we adopt an equal ratio, where $\alpha = 2.0$, $\beta = 1.0$, and $\gamma = 1.0$.

Table 13: **Ratios of Answer Reward and Format Reward Weights.**

| Model | 3:1 | 2:1 | 1:1 | 1:2 | 1:3 |
|---|---|---|---|---|---|
| VLM-PRF-3B | 38.80 | 38.20 | 39.48 | 38.89 | 38.53 |

### C.4 The Number of Selected Sections

In Table 14 and Table 15, we present ablation studies conducted on VLM-PRF-3B to evaluate the impact of varying the number of retrieved articles and sections during tool-based retrieval on the InfoSeek and E-VQA datasets. The tables report the final accuracy of Wiki-PRF-3B when top-1 and top-3 retrieved articles or sections are used during training.

The results indicate that the model performance generally improves as the number of selected sections increases. However, when only a single article is considered, the overall relevance of the article becomes the primary determinant of accuracy. The inclusion of redundant sections introduces noise and may lead to a decline in performance.

Table 14: **Retrieved Settings Ablation on Infoseek.**

| Retrieved Settings | Top-1 Section | Top-3 Sections |
|---|---|---|
| Top-1 Article | 38.96% | 38.85% |
| Top-3 Articles | 39.03% | 39.10% |
| Top-5 Articles | 39.39% | 39.48% |

Table 15: **Retrieved Settings Ablation on E-VQA.**

| Retrieved Settings | Top-1 Section | Top-3 Sections |
|---|---|---|
| Top-1 Article | 24.28% | 24.31% |
| Top-3 Articles | 28.15% | 28.94% |
| Top-5 Articles | 32.10% | 32.38% |

## D Qualitative Results

### D.1 Comparison of Wiki-PRF

We conduct a comparison between our method and two baselines: Vanilla RAG and Wiki-PRF without the reinforcement learning fine-tuning (Wiki-PRF w/o RL). As shown in Figure 6, we present a comparison across various scenes, including plants, buildings, and animals. Examples 2 and 3 in

Figure 6 and example 1 in Figure 7 demonstrate the accuracy of our method in answering number-related questions. Examples 3 and 4 in Figure 6 show that our method can still accurately answer questions when the target subject is far away. The comparison results fully illustrate the effectiveness of our method.

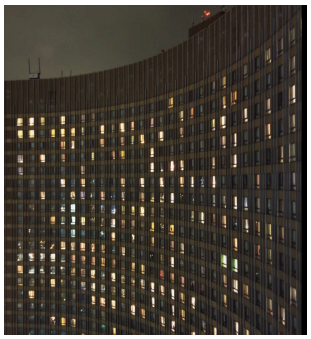
### D.2 Illustration of Wiki-PRF on Variours Questions

In this section, we mainly show the case examples of Wiki-PRF. Figure 8 and Figure 11 show examples of visualizations of different tools working individually. Figure 9 and Figure 10 show scenarios where the two tools work together, but are called in different orders. Figure 12 demonstrates that through tool calls, Wiki-PRF extends information retrieval to retrieve information that contributes to the answer. Figure 13 shows an example of calling the captioning combination, proving that Wiki-PRF will make specific tool combinations according to the question. The filtered results across all samples are shorter than the original messages while retaining the correct answer, illustrating the effectiveness of the filtering stage.

## E Broader Impacts of Wiki-PRF

In this section, we focus on the broader impacts of our method. Our method facilitates assisting VLM to achieve better answers through knowledge retrieval. Importantly, any potentially harmful information encountered during the retrieval stage can be mitigated by appropriately restricting the scope of the knowledge base.

# E-VQA



**Question:** How many rooms does this hotel have?

**Vanilla RAG** : The Izmailovo Hotel has 5,000 rooms. ✗

**VLM-PRF (w/o RL)** : The hotel in the image has 5,000 rooms. ✗

**VLM-PRF** : The hotel has 1,777 rooms. \n Question: How many rooms does this hotel have? \n Short answer: ✓

**Ground Truth:** [' 1,777 rooms']

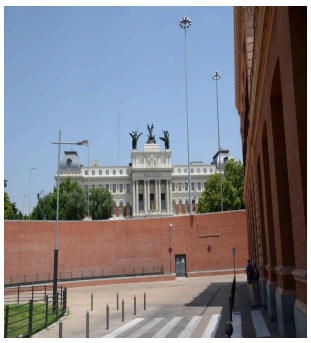
**Question:** What kind of perianth radial symmetry do the flowers of this plant have?

**Vanilla RAG** : The flowers of Schefflera arboricola are radially symmetrical in their perianth. ✗

**VLM-PRF (w/o RL)** : The image provided does not show flowers, //...//Since the image only shows the foliage, which is a characteristic feature of Schefflera arboricola, the discussion about perianth symmetry would be inappropriate. ✗

**VLM-PRF** : The flowers of this plant have a double perianth with radial symmetry. ✓

**Ground Truth:** ['double']



**Question:** Who designed this palace?

**Vanilla RAG** : The Palacio de la Marquesa de Sonora was designed by the architect Juan de Villanueva in the early 19th century. //...// ✗

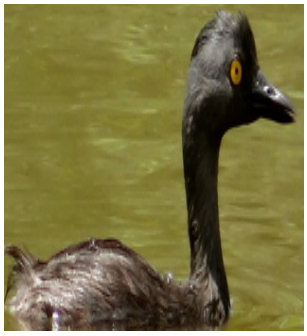
**VLM-PRF (w/o RL)** : The palace in the image//...// by Juan Gómez de Mora or Juan de Herrera. Therefore, it is not possible to determine the designer of the palace from the given information or the image. ✗

**VLM-PRF** : Ricardo Velázquez Bosco designed the Palace of Agriculture. ✓

**Ground Truth:** ['Ricardo Velázquez Bosco']

Figure 6: **Comparison on E-VQA**

# Infoseek



**Question:** What is the weight of a male of this bird in gram?

**Vanilla RAG** : A male least grebe typically weighs between 129–161 grams. ✗

**VLM-PRF (w/o RL)** : The weight of a male least grebe ranges from 112 to 180 grams. The bird in the picture appears to be a male least grebe, given its size and characteristics. Therefore, the weight of this bird could be anywhere between 112 to 180 grams. ✗

**VLM-PRF** : The weight of the bird in the picture is approximately 129 grams. ✓

**Ground Truth:** [{'wikidata': 129.0, 'range': [116.10000000000001, 141.9]}]

**Question:** What is the sea level in metre of this mountain?

**Vanilla RAG** : The provided image and details do not contain information about the sea//...// such as the elevation of the mountain or the village. ✗

**VLM-PRF (w/o RL)** : The image does not provide specific data about the sea level or //...// elevation of the mountain. ✗

**VLM-PRF** : The sea level in meters of this mountain is 1228 meters. ✓

**Ground Truth:** [{'wikidata': 1228.0, 'range': [1105.2, 1350.8000000000002]}]

**Question:** What is the immediately prior item that this bridge follows in a series?

**Vanilla RAG** : The Charles Bridge. ✗

**VLM-PRF (w/o RL)** : The image does not provide enough context to determine what the immediately prior item is in a series. //...//Charles Bridge in Prague, Czech Republic, and does not indicate any preceding item. ✗

**VLM-PRF** : The immediately prior item that this bridge follows in a series is the Judith Bridge. ✓

**Ground Truth:** ['Judith bridge']

Figure 7: **Comparison on InfoSeek**

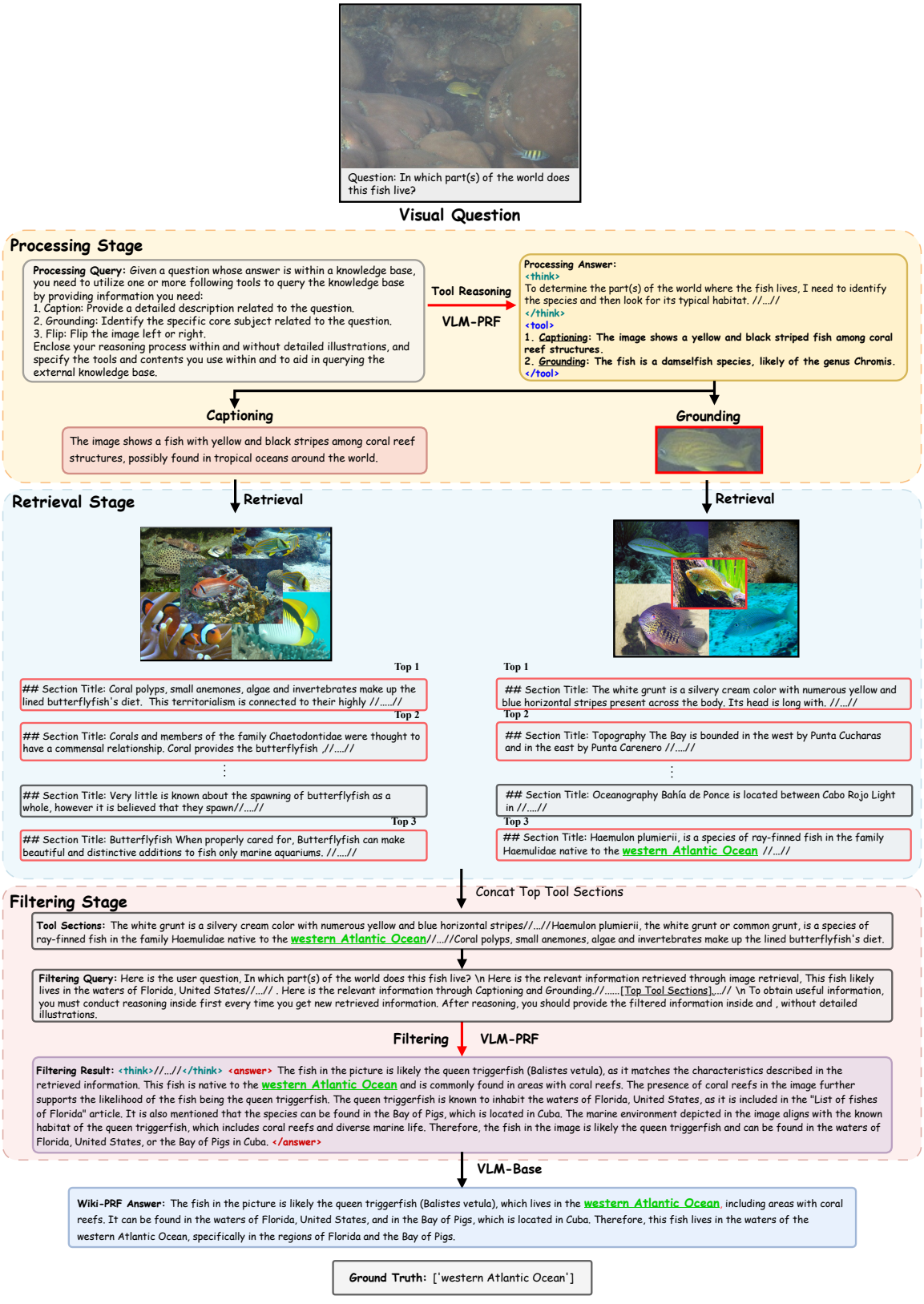



Figure 8: **Illustration of Wiki-PRF on Question E-VQA_114 from E-VQA.**

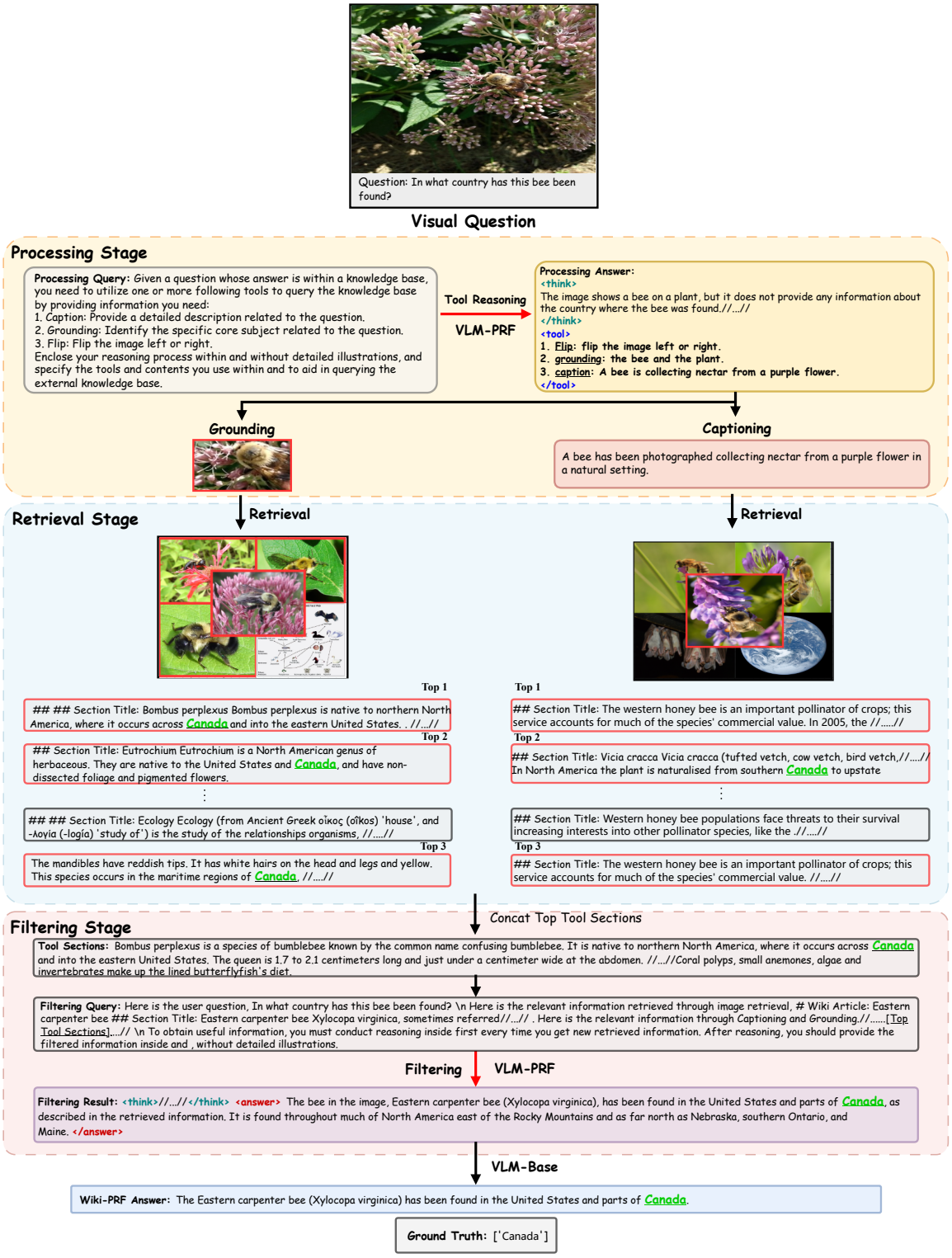



Figure 9: **Illustration of Wiki-PRF on Question E-VQA_1182 from E-VQA.**

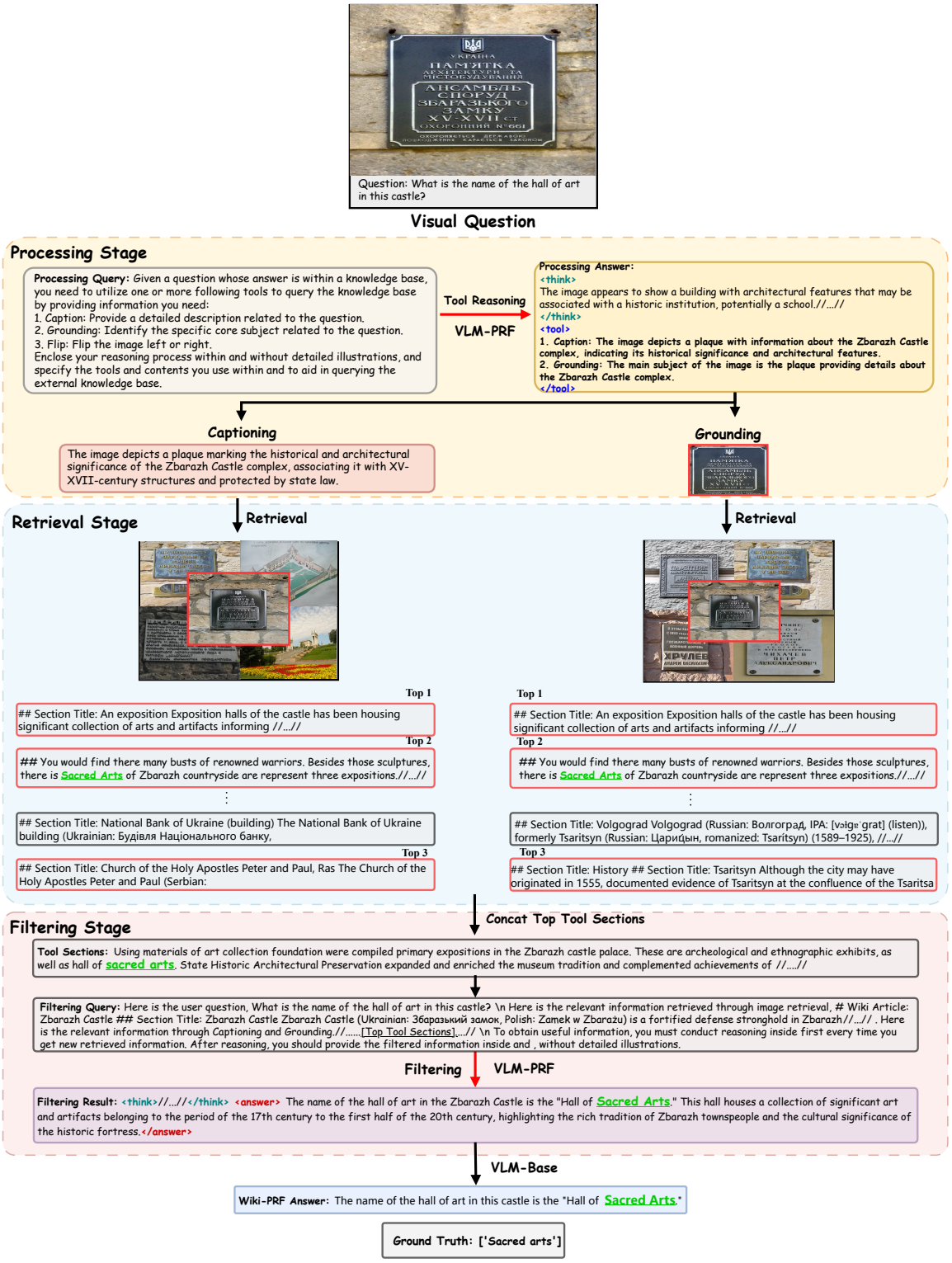



Figure 10: **Illustration of Wiki-PRF on Question E-VQA_1747 from E-VQA.**

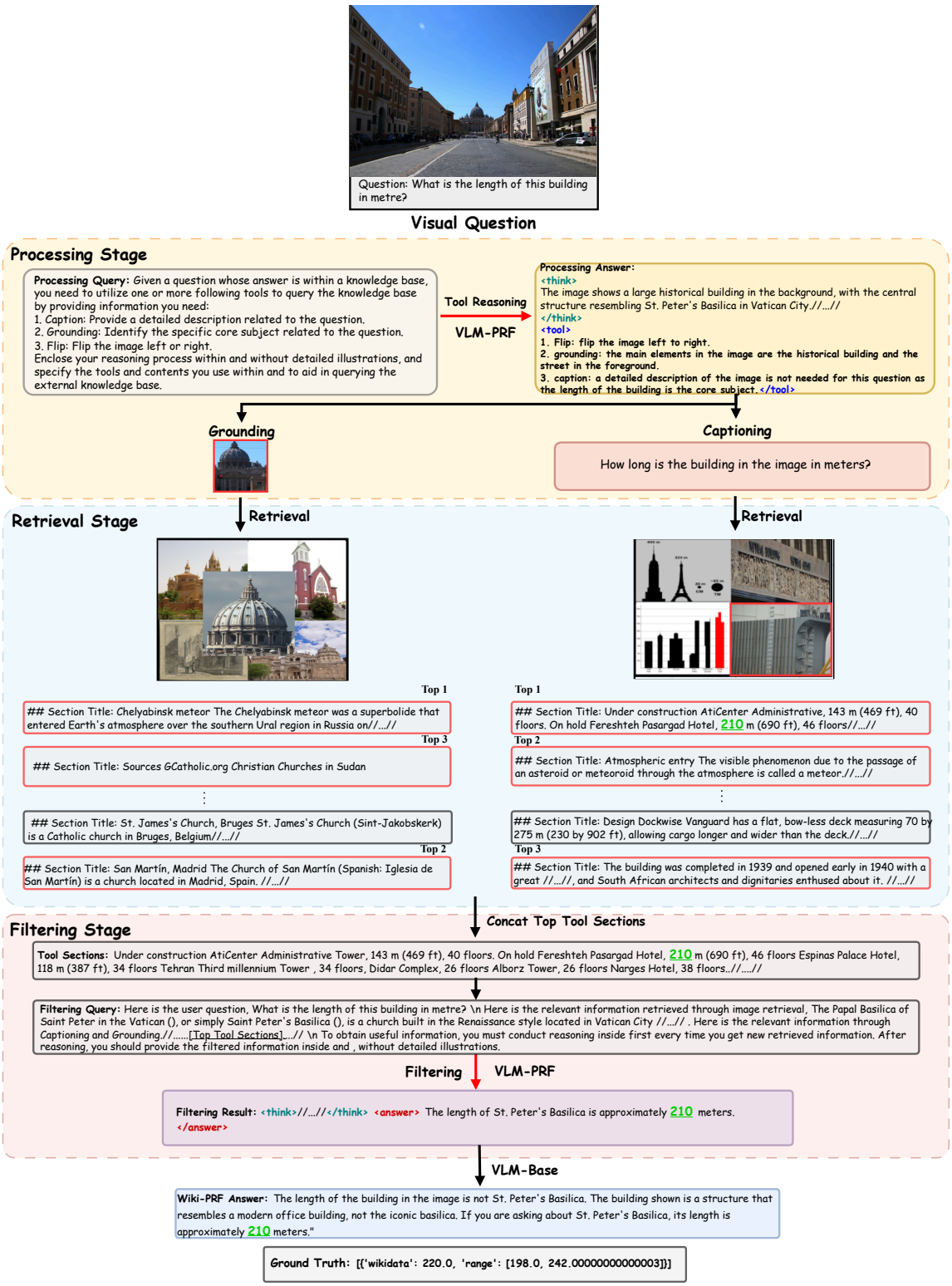



Figure 11: **Illustration of Wiki-PRF on Question Infoseek_00012299 from Infoseek.**

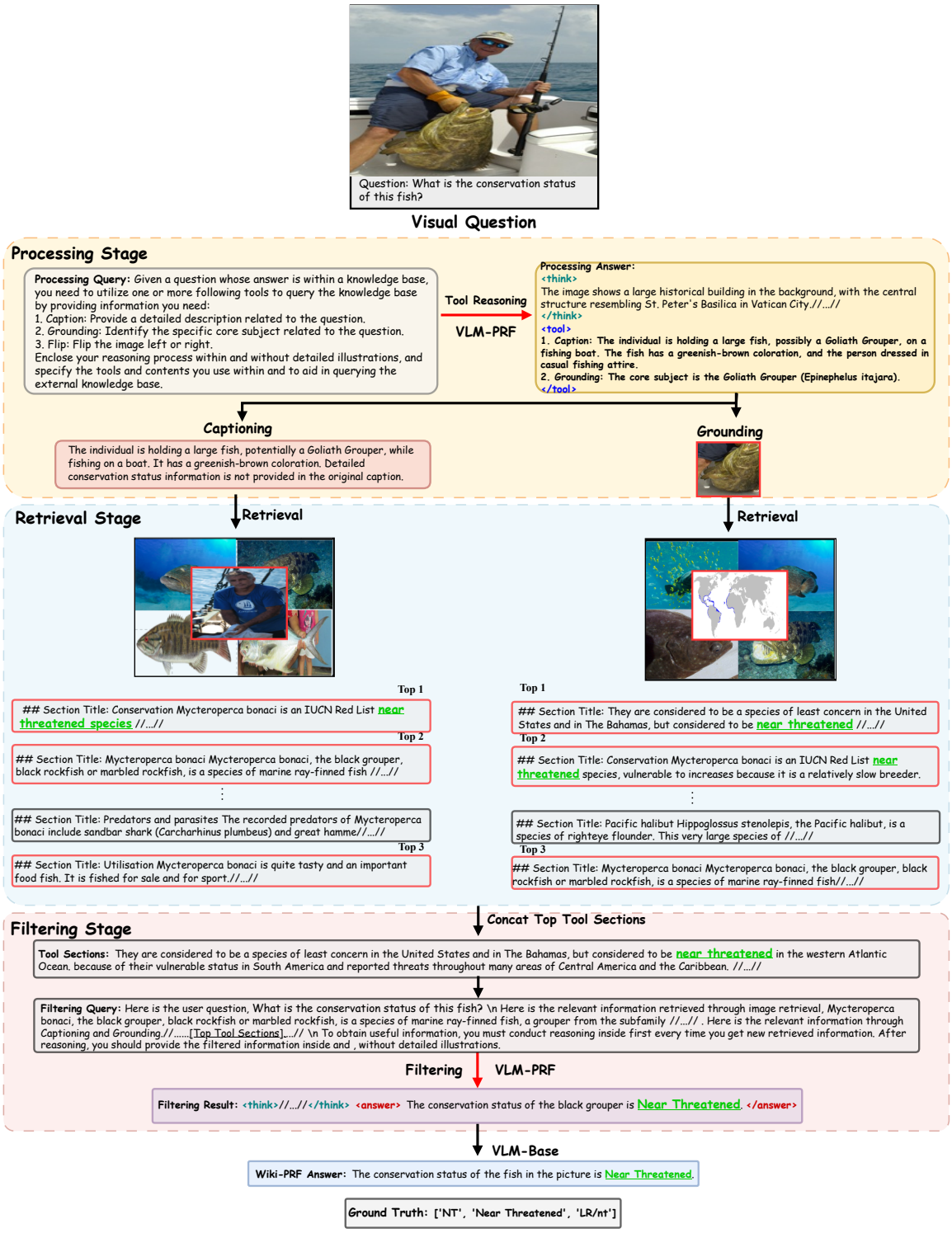



Figure 12: **Illustration of Wiki-PRF on Question Infoseek_00033513 from Infoseek.**

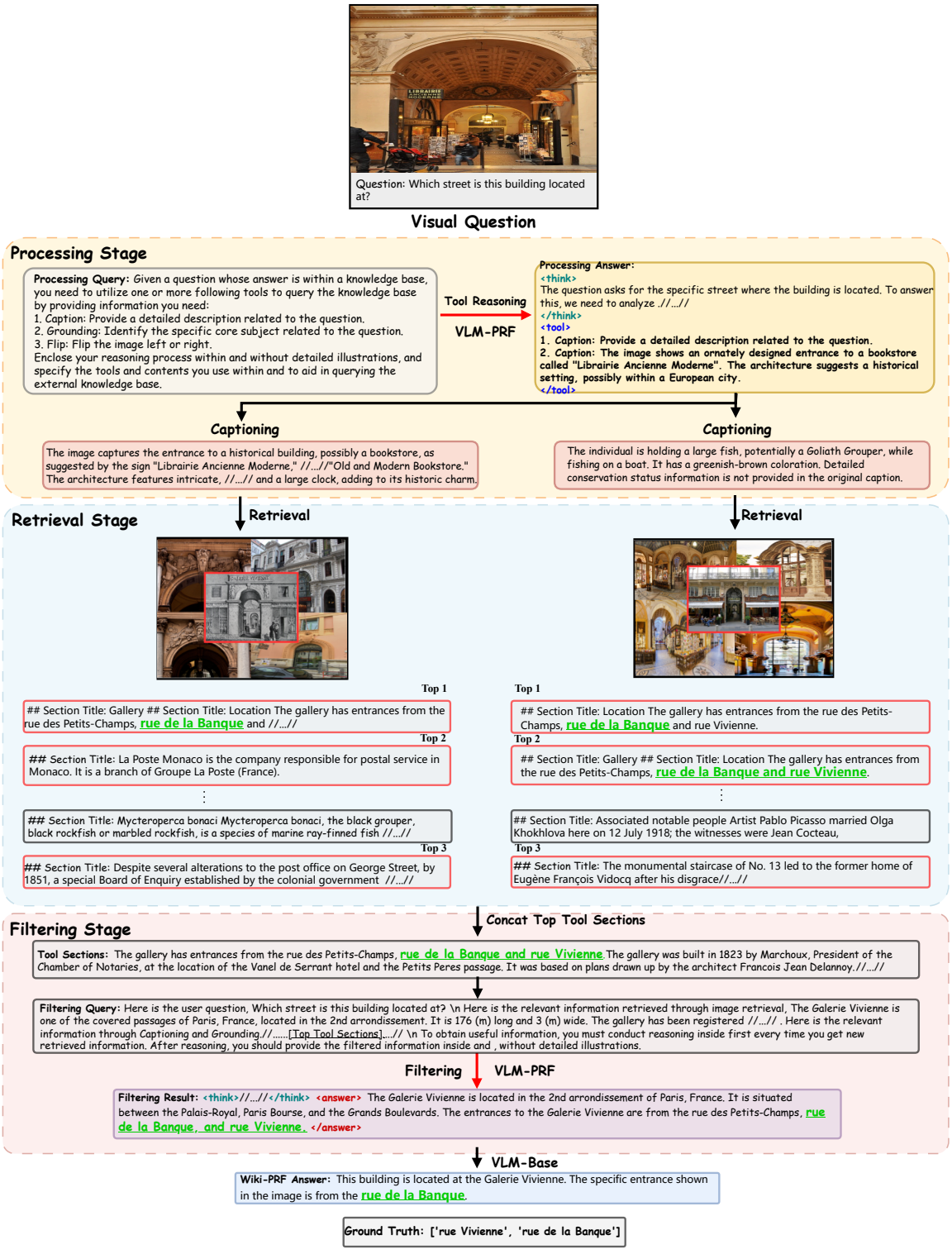



Figure 13: **Illustration of Wiki-PRF on Question Infoseek_00005094 from Infoseek.**