

---

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

## *Supplementary Materials*

---

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

### 5 A Workflow of Wiki-PRF

6 We detail the complete workflow of Wiki-PRF below.

#### 7 • Processing Stage

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

## 39 B Prompts Details in Wiki-PRF

### 40 B.1 Processing Stage

#### 41 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}.

#### 43 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."

#### 45 Prompt for Grounding:

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

### 47 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."

### 49 B.3 Prompt for Answer

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

## 51 C Additional Experiments

### 52 C.1 Training Loss

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

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

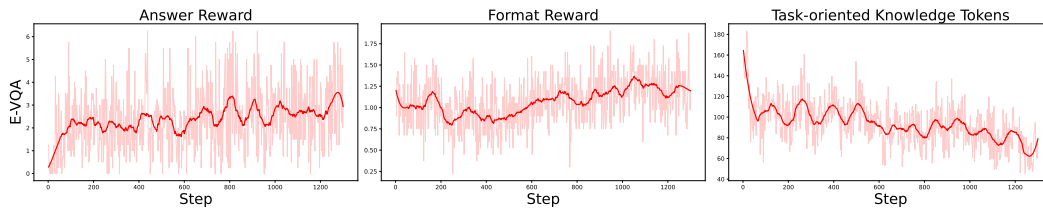


Figure A: The training curve of VLM-PRF-7B in E-VQA.

## 62 C.2 Tool Inference Time

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

Table A: **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

66

## 67 C.3 Weights of Rewards

68 As shown in Table B, we investigate the influence of varying the weight of answer reward (i.e.,  $\alpha$ )  
 69 and the weight of format reward (i.e.,  $\beta + \gamma$ ) in the overall objective function on the InfoSeek dataset.  
 70 We fix the values of  $\beta$  and  $\gamma$  to 1.0, and continuously adjust the ratio of  $\alpha$  and  $(\beta + \gamma)$ . By gradually  
 71 decreasing  $\alpha : (\beta + \gamma)$  from 3 : 1 to 1 : 3, we observe that the optimal performance is achieved when  
 72 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 B: **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

73

## 74 C.4 The Number of Selected Sections

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

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

Table C: **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 D: **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%

## 83 D Qualitative Results

### 84 D.1 Comparison of Wiki-PRF

85 We conduct a comparison between our method and two baselines: Vanilla RAG and Wiki-PRF  
 86 without the reinforcement learning fine-tuning (Wiki-PRF w/o RL). As shown in Figure B, we  
 87 present a comparison across various scenes, including plants, buildings, and animals. Examples 2  
 88 and 3 in Figure B and example 1 in Figure C demonstrate the accuracy of our method in answering

89 number-related questions. Examples 3 and 4 in Figure B show that our method can still accurately  
90 answer questions when the target subject is far away. The comparison results fully illustrate the  
91 effectiveness of our method.

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

93 In this section, we mainly show the case examples of Wiki-PRF. Figure D and Figure G show  
94 examples of visualizations of different tools working individually. Figure E and Figure F show  
95 scenarios where the two tools work together, but are called in different orders. Figure H demonstrates  
96 that through tool calls, Wiki-PRF extends information retrieval to retrieve information that contributes  
97 to the answer. Figure I shows an example of calling the captioning combination, proving that Wiki-  
98 PRF will make specific tool combinations according to the question. The filtered results across all  
99 samples are shorter than the original messages while retaining the correct answer, illustrating the  
100 effectiveness of the filtering stage.

## 101 **E Broader Impacts of Wiki-PRF**

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



## E-VQA



**Question:** Which animals eats the fruits / leaves of this plant?

**Vanilla RAG :** The fruits and leaves of *Monarda fistulosa* //...// many insects and small birds may utilize the nectar and pollen from the flowers for food. Additionally, birds may occasionally consume the seeds. ❌

**Wiki-PRF (w/o RL) :** Based on the information provided in the context, the article does not specify which animals eat the fruits or leaves of the *Monarda bradburiana*. //...// Additionally, herbivorous mammals are said to avoid the foliage due to its odor. ❌

**Wiki-PRF :** The passage does not provide //...// The known pollinators for this plant are bees, butterflies, and hummingbirds, as well as DOUFIORAE MONARDAE, which itself is a pollinator. ✔️

**Ground Truth:** ['bees, hummingbirds, and lepidoptera']



**Question:** How many meters high can this plant grow?

**Vanilla RAG :** The image shows *Fuchsia magellanica*, which typically grows to heights of 1 to 2 meters. ❌

**Wiki-PRF (w/o RL) :** This plant can grow up to 10 meters in height in frost-free climates. ❌

**Wiki-PRF :** The plant *Fuchsia magellanica* can grow up to 3.0 meters high in frost-free climates and up to 1.5 meters high in colder conditions. ✔️

**Ground Truth:** ['3.0']



**Question:** When was this palace built?

**Vanilla RAG :** The Ruzhany Palace was built between the 16th and 19th centuries. ❌

**Wiki-PRF (w/o RL) :** The palace was built from 1806 to 1810. ❌

**Wiki-PRF :** The palace was built between the late 16th century and 1602. ✔️

**Ground Truth:** ['1602']



**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. //...// ❌

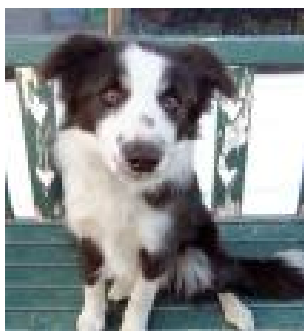
**Wiki-PRF (w/o RL) :** The palace in the image //...// by Juan Gómez de Mora or Juan de Herrera. Therefore, it is not possible to determine the designer of the palace from the given information or the image. ❌

**Wiki-PRF :** Ricardo Velázquez Bosco designed the Palace of Agriculture. ✔️

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

Figure B: Comparison on E-VQA

## Infoseek



**Question:** What is the height of this animal in centimetre?

**Vanilla RAG :** The image provided is of a dog, but it does not specify the height in centimeters. //...// we would need a measurement of the dog's height, which is not available in the image or the context provided. ❌

**Wiki-PRF (w/o RL) :** The image provided is of a dog, but it is not a border collie. Therefore, it is not possible to determine the exact height of this dog in centimeters based on the image alone. ❌

**Wiki-PRF :** The height of the animal in the picture is approximately 48 cm to 56 cm tall, as it is a male Border Collie. ✓

**Ground Truth:** '[{"wikidata": "53.0", "range": "[47.7, 58.300000000000004]"}]'



**Question:** What is the conservation status of this animal?

**Vanilla RAG :** The Greater Kudu is classified as Vulnerable (VU) by the International Union for Conservation of Nature ❌

**Wiki-PRF (w/o RL) :** The conservation status of the greater kudu is Vulnerable. ❌

**Wiki-PRF :** The conservation status of the greater kudu is "Least Concern." ✓

**Ground Truth:** ['LR/lc', 'Least Concern', 'LC']



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

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

**Wiki-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. ❌

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

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



**Question:** Which christian liturgical rite is this building associated with?

**Vanilla RAG :** Catholic Church ❌

**Wiki-PRF (w/o RL) :** Catholic ❌

**Wiki-PRF :** Roman Rite ✓

**Ground Truth:** ['Roman Rite']

Figure C: Comparison on InfoSeek

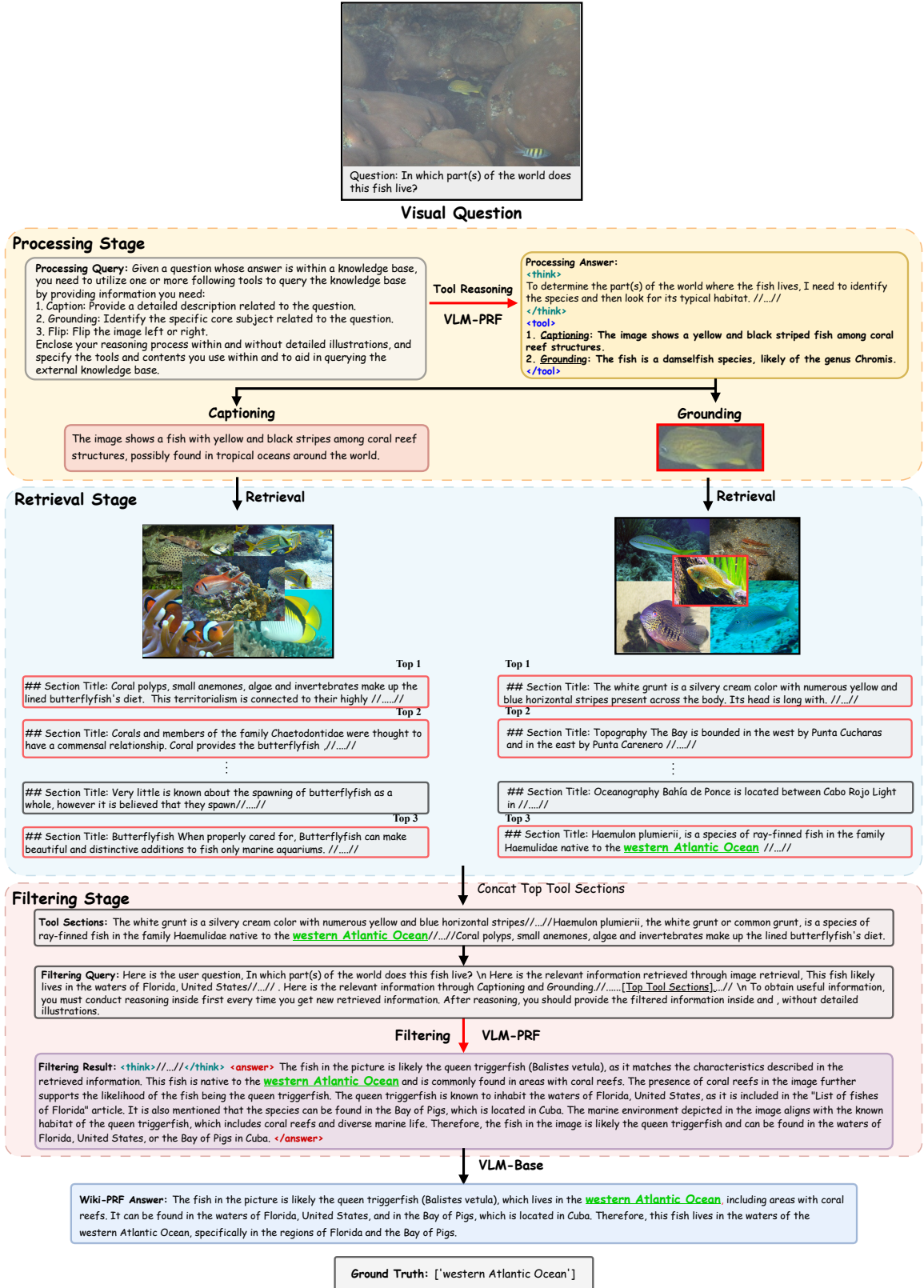


Figure D: Illustration of Wiki-PRF on Question E-VQA\_114 from E-VQA.

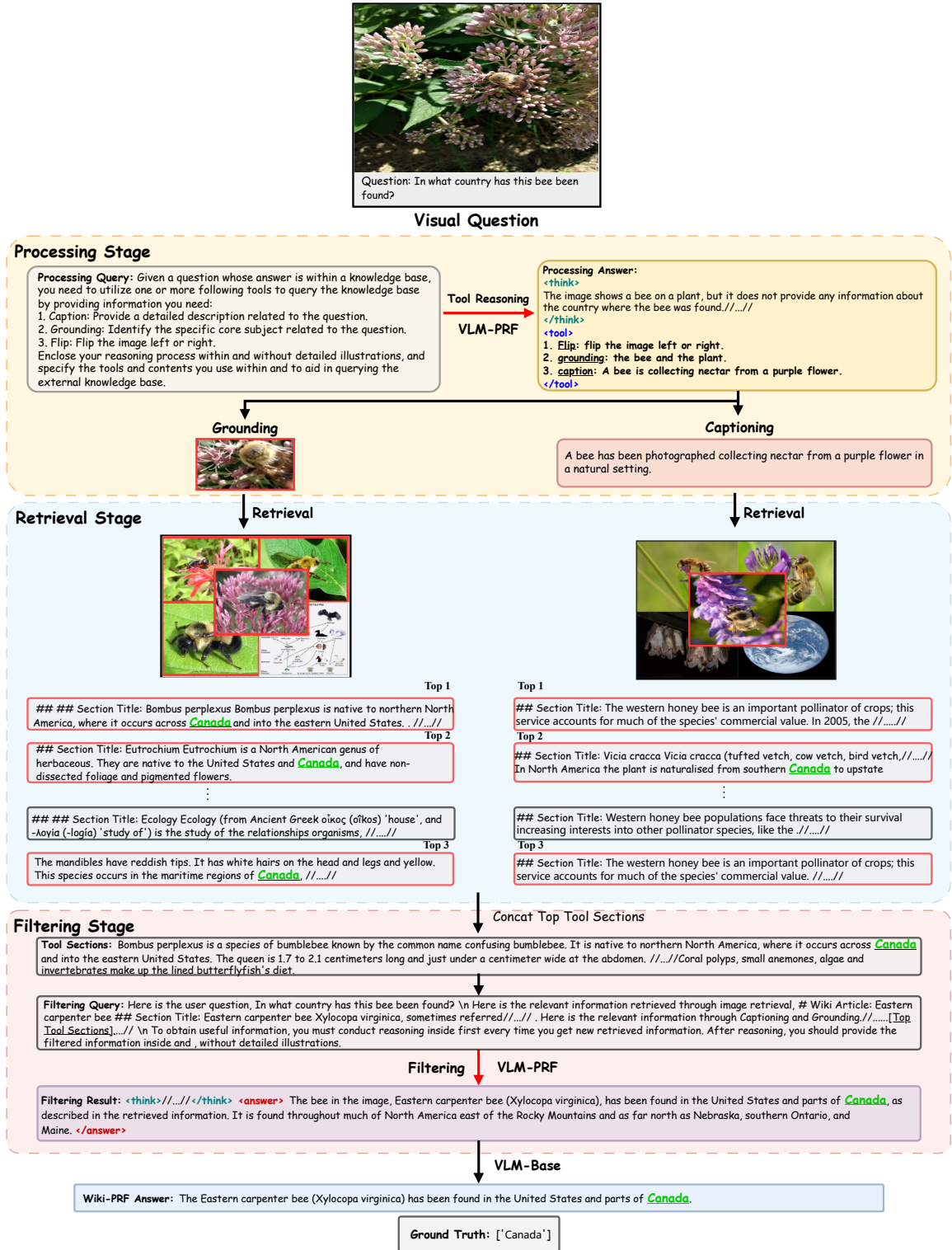


Figure E: Illustration of Wiki-PRF on Question E-VQA\_1182 from E-VQA.



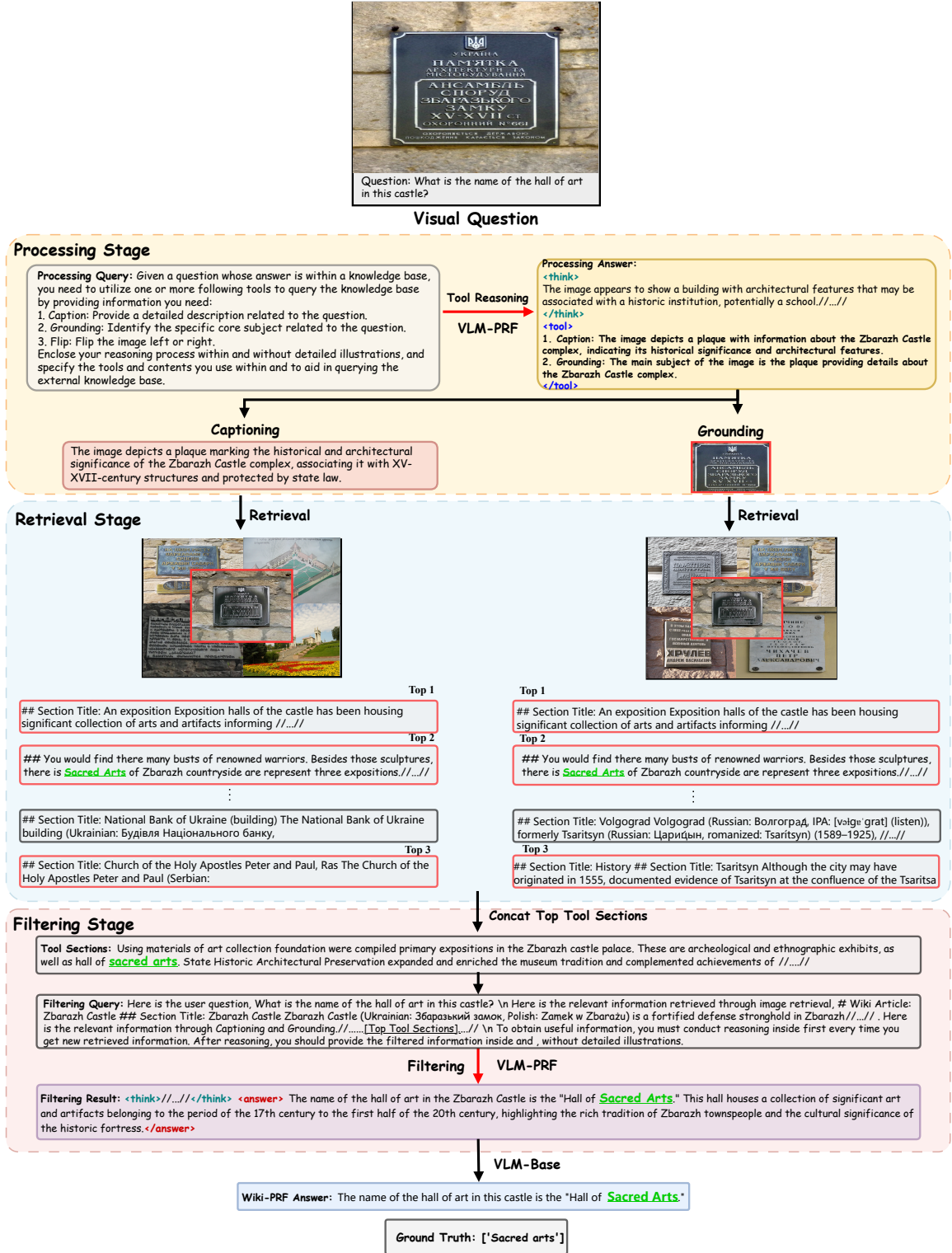


Figure F: Illustration of Wiki-PRF on Question E-VQA\_1747 from E-VQA.

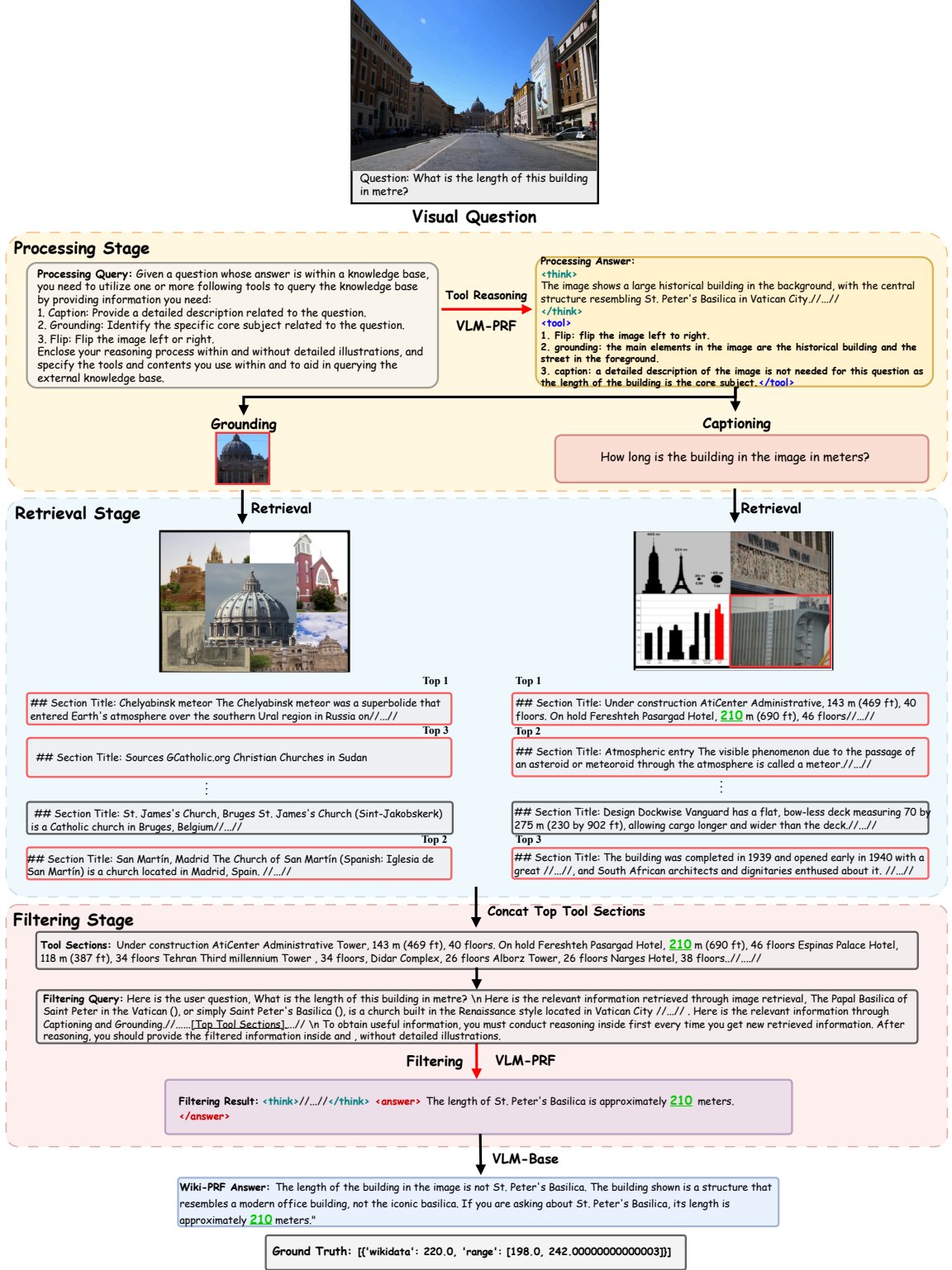
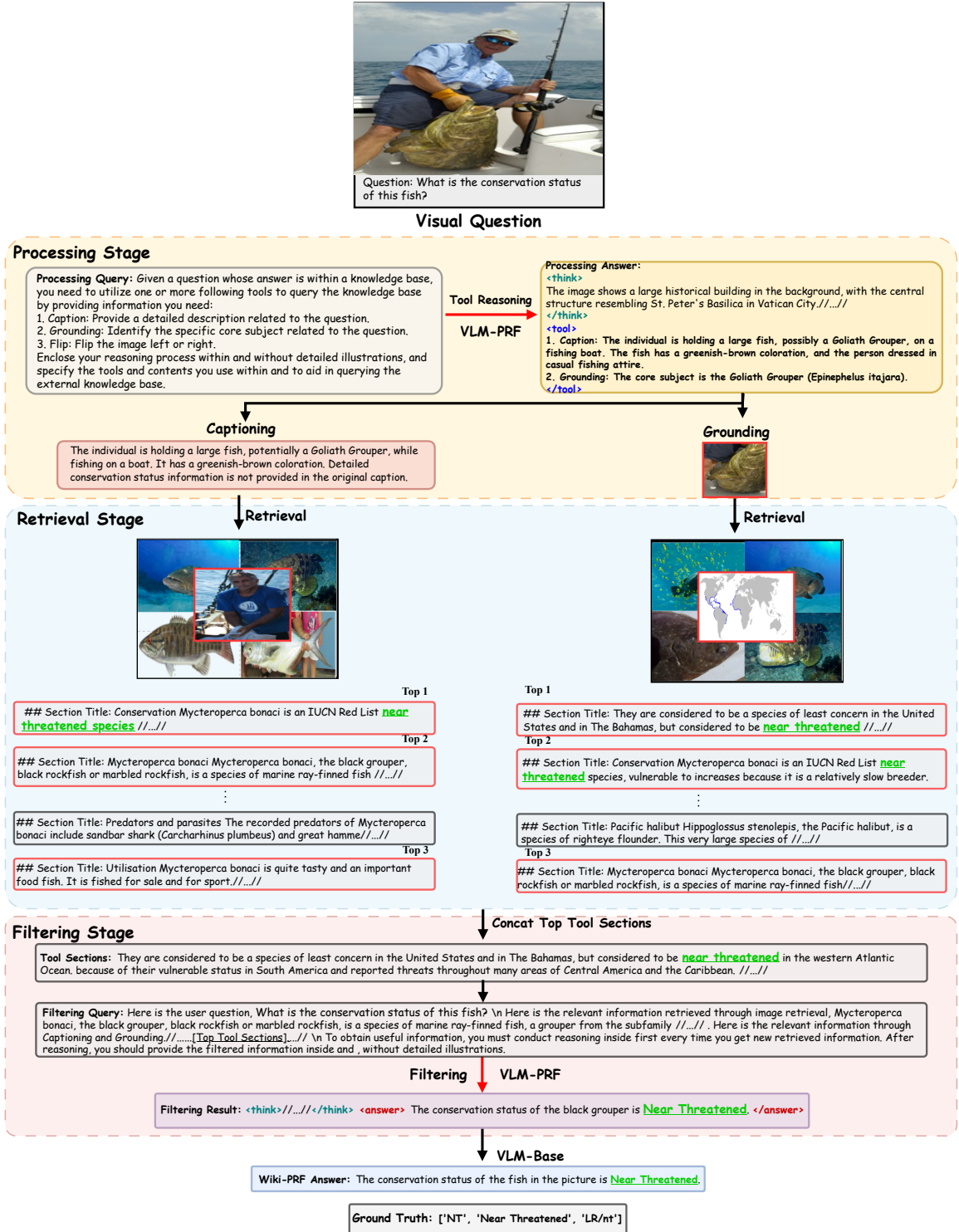


Figure G: Illustration of Wiki-PRF on Question Infoseek\_00012299 from Infoseek.



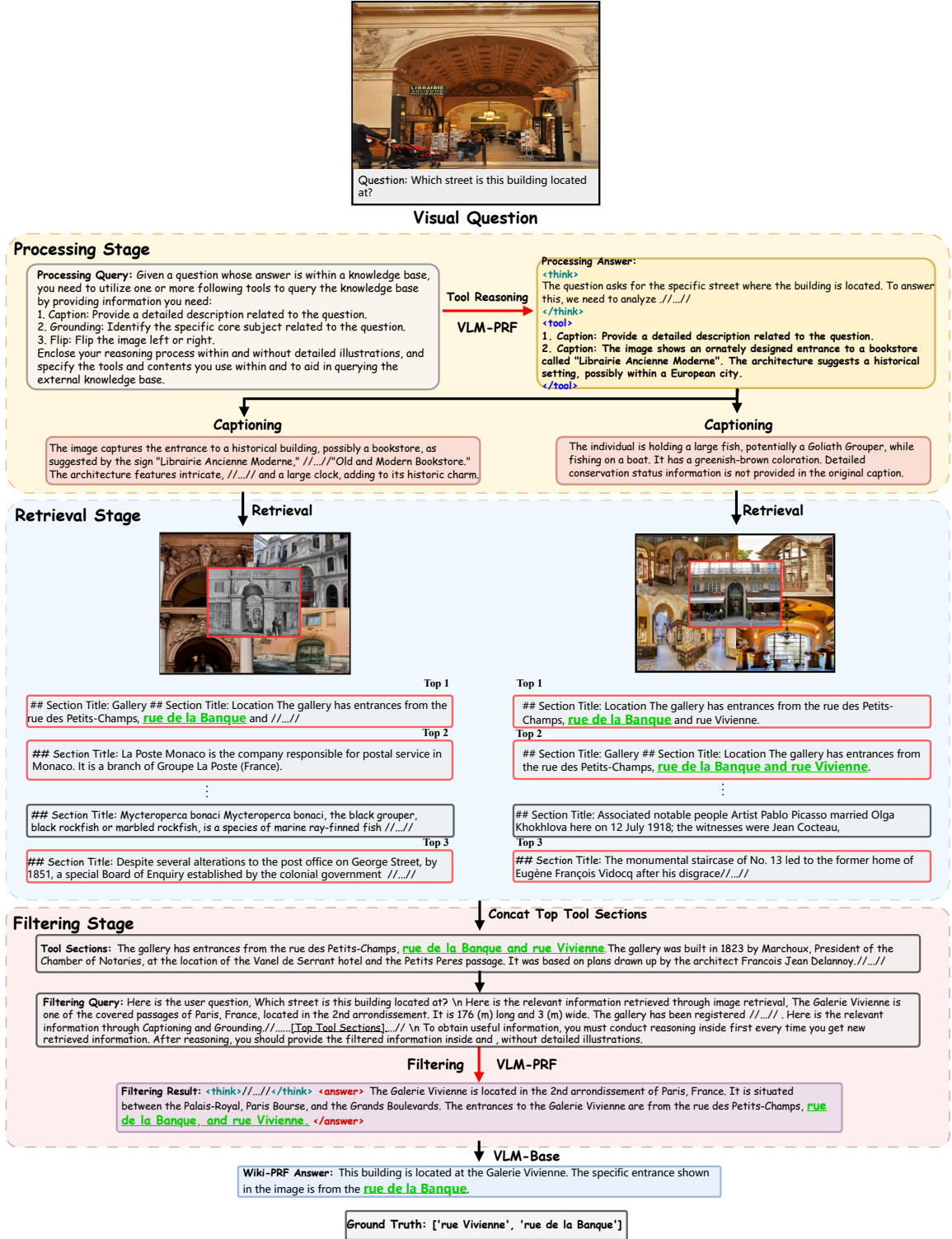


Figure I: Illustration of Wiki-PRF on Question Infoseek\_00005094 from Infoseek.