
VRAG-RL: Empower Vision-Perception-Based RAG for Visually Rich Information Understanding via Iterative Reasoning with Reinforcement Learning

Qiuchen Wang¹, Ruixue Ding, Yu Zeng¹, Zehui Chen¹, Lin Chen¹,
Shihang Wang, Pengjun Xie, Fei Huang, Feng Zhao^{1*}

¹MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, USTC
qiuchenwang@mail.ustc.edu.cn, fzhao956@ustc.edu.cn

Abstract

Effectively retrieving, reasoning and understanding visually rich information remains a challenge for traditional Retrieval-Augmented Generation (RAG) methods. On the one hand, traditional text-based methods cannot handle visual-related information. On the other hand, current vision-based RAG approaches are often limited by fixed pipelines and frequently struggle to reason effectively due to the insufficient activation of the fundamental capabilities of models. As reinforcement learning (RL) has been proven to be beneficial for model reasoning, we introduce **VRAG-RL**, a novel RL framework tailored for complex reasoning across visually rich information. With this framework, VLMs interact with search engines, autonomously sampling single-turn or multi-turn reasoning trajectories with the help of visual perception tokens and undergoing continual optimization based on these samples. Our approach highlights key limitations of RL in RAG domains: (i) Prior Multi-modal RAG approaches tend to merely incorporate images into the context, leading to insufficient reasoning token allocation and neglecting visual-specific perception; and (ii) When models interact with search engines, their queries often fail to retrieve relevant information due to the inability to articulate requirements, thereby leading to suboptimal performance. To address these challenges, we define an action space tailored for visually rich inputs, with actions including cropping and scaling, allowing the model to gather information from a coarse-to-fine perspective. Furthermore, to bridge the gap between users' original inquiries and the retriever, we employ a simple yet effective reward that integrates query rewriting and retrieval performance with a model-based reward. Our VRAG-RL optimizes VLMs for RAG tasks using specially designed RL strategies, aligning the model with real-world applications. Extensive experiments on diverse and challenging benchmarks show that our VRAG-RL outperforms existing methods by 20% (Qwen2.5-VL-7B) and 30% (Qwen2.5-VL-3B), demonstrating the effectiveness of our approach. The code is available at <https://github.com/Alibaba-NLP/VRAG>.

1 Introduction

Retrieval-Augmented Generation (RAG) [12, 18, 4] enables Language Models (LMs) to leverage external information to tackle problems. Due to the limitations of traditional textual RAG methods in handling visually rich information, efforts have been made to introduce RAG into the visual domain by integrating Vision-Language Models (VLMs) [1, 7, 33, 15, 34] with search engines. However, current visual RAG methods still fall short in effectively reasoning with search engines and

*Corresponding author

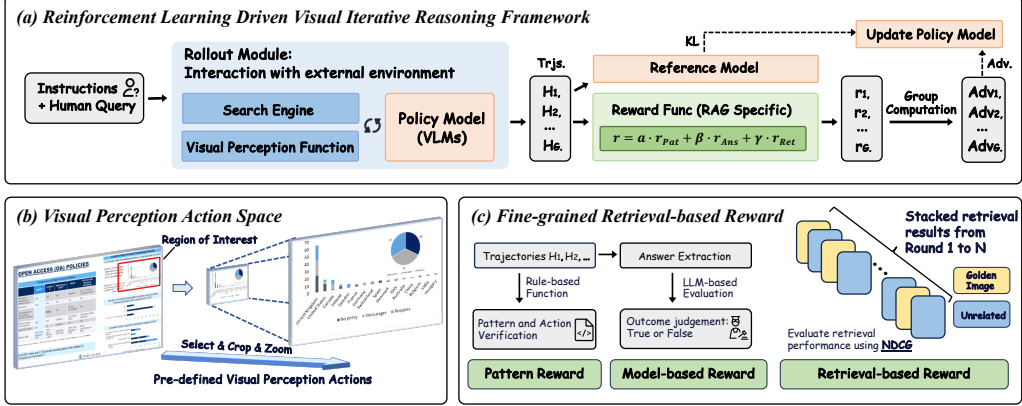


Figure 1: **Overall Framework of our Reinforcement Learning Framework.** (a) demonstrates the interaction process between the model and the external environment, as well as the implementation of the GRPO algorithm. (b) shows the proposed visual perception action space which allows the model to extract information from a coarse-to-fine perspective. (c) is the specially designed reward for RAG, which combines outcome and retrieval performance across the entire sampling process.

understanding complex visual information. Reinforcement Learning (RL) has been recognized as an effective approach for optimizing VLMs in complex reasoning tasks [39, 20, 14, 30, 49]. Therefore, RL offers a promising approach to address the challenges faced by visual RAG methods.

Inspired by these advancements, we introduce **VRAG-RL**, a novel multimodal RL framework specifically designed for iterative reasoning in visually rich information RAG. Our approach is based on three critical observations: **(i) Insufficient activation of reasoning capabilities with visual information.** Existing methods underutilize the reasoning potential of VLMs when incorporating visual information. For instance, prior approaches tend to merely embed images into the context without adequately addressing visual-specific perception processes, resulting in insufficient reasoning token allocation and limiting the models’ ability to fully leverage visual data for complex reasoning tasks. **(ii) Inefficient and disjointed Retrieval.** In previous work, limited by the inability to articulate complex requirements, models struggled to retrieve relevant information efficiently, which may lead to repetitive and meaningless interactions, restricting the overall effectiveness. **(iii) Inconsistent multi-turn reasoning and unstable training with VLMs.** Current RL frameworks for LMs often struggle with maintaining stability and consistency during multi-turn reasoning. Handling complex, multi-step reasoning tasks can be particularly challenging, as models may encounter difficulties in maintaining effective reasoning across interactions with external environments, leading to inconsistent performance and suboptimal results. This challenge is further exacerbated for VLMs, which are limited by their instruction-following and reasoning capabilities.

Building upon these insights, VRAG-RL introduces improvements in various modules: (i) We propose a visual perception action space that includes selecting regions of interest and zooming into these areas. VLMs with visual perception tokens in the action space are capable of acquiring information from coarse-to-fine perspective. As shown in Figure 1 (b), when dealing with images or charts within documents, VLMs can give higher attention to information-dense areas through the proposed perception tokens. This allows the model to more effectively activate reasoning abilities within a limited context length, preventing the overlooking of details. (ii) Furthermore, rather than relying solely on a simple outcome-based reward, we factor in the effectiveness of the retrieval process as part of the reward structure. In particular, during the interaction between the model and the search engine, retrieving pertinent images promptly enhances the model’s ability to address questions effectively, whereas persistently retrieving irrelevant documents adds noise and hampers the reasoning process. As illustrated in Figure 1 (c), by integrating retrieval performance into reward, we establish comprehensive guidance for retrieval-augmented generation frameworks. (iii) Inspired by the current think-then-answer approach and the ReAct paradigm, we model the interaction between the VLMs and the search engine, along with the visual perception action space, as a process of iterative reasoning and tool invocation. Figure 1 (a) illustrates our training pipeline, which supports automatic sampling and integrates the GRPO algorithm. To ensure stability in multi-turn sampling

and training, we have carefully designed the sampling strategy including post-processing for each interaction, and model-based reward together with the retrieval reward mentioned above guides the model training. Additionally, we have re-annotated existing datasets of visually rich documents and developed a data construction pipeline to efficiently scale data for RL and SFT.

Our major contributions are as follows:

- We propose VRAG-RL, a novel reinforcement learning framework tailored for training VLMs to effectively reason, retrieve, and understand visually rich information.
- We define a visual perception action space that includes selecting, cropping, and scaling regions of interest, allowing VLMs to gather information progressively from coarse-grained to fine-grained levels. This action space enhances the models’ ability to focus on information-dense areas and activates their vision-specific reasoning capabilities more effectively.
- We introduce a comprehensive reward structure that integrates retrieval performance and model-based outcome reward. This reward mechanism aligns the model more closely with real-world applications, bridging the gap between users’ original intentions and the retriever.
- Extensive experiments demonstrate the effectiveness of our method. VRAG-RL significantly outperforms strong baselines, achieving over 20% improvement on various benchmarks.

2 VRAG-RL

In this section, drawing on insights and foundational ideas, we present a comprehensive description of our **VRAG-RL** framework. We start with the formulation of the problem (§2.1), then introduce the action space designed for visual perception (§2.2) and the fine-grained reward specifically defined for the RAG task (§2.3). Finally, we illustrate the model interaction process in the rollout module and the reinforcement learning training implementation of our framework (§2.4).

2.1 Problem Formulation

Given a query denoted as q , we have a huge collection of images $\mathcal{C} = \{\mathbf{I}_1, \mathbf{I}_2, \dots, \mathbf{I}_N\}$, consisting of N images. Each image contains a variety of visually rich elements, such as flowcharts, charts, tables, and diverse layouts, derived from real-world documents across multiple domains, including slides and reports. Our goal is to efficiently reason, accurately retrieve the most relevant images, extract valuable information from the complex visual data, and generate the final answer a to the query q .

2.2 Visual Perception Action Integration for Understanding Information-Dense Regions

Previous works merely involved migrating textual RAG to the multi-modal domain, which simply meant inserting images into the context and then reasoning and responding. However, these efforts overlooked the characteristics of image data, where the efficiency of visual perception is closely related to image resolution, visual element layouts, information density, and other visually related factors. Motivated by these findings, we introduce a dynamic novel visual perception paradigm into VLMs that involves region selection and re-encoding at the token level, as illustrated in Figure 2.

Definition of Visual Perception Actions. We define the visual perception action space for VLMs by taking into account the specific characteristics of visual information. This enables the model to select regions with high information density or regions relevant to the query for a detailed view, acquiring information from a coarse to fine perspective. We integrate search queries, answer summaries, and visually specific actions into a unified action space to align with the model’s pre-training domain.

The policy model π_θ interacts with the environment in the Thought-Action-Observation ($\mathcal{T}, \mathcal{A}, \mathcal{O}$) paradigm. In each interaction, the model generates the next action $\mathcal{A}_t \sim \pi_\theta(\cdot \mid \mathcal{H}_{t-1})$ based on the trajectory \mathcal{H}_{t-1} from step $t-1$ and earlier. A role-based function is used to extract visual perception tokens $\langle \text{region} \rangle$ and $\langle / \text{region} \rangle$, whose main purpose is to select, crop, and zoom in on the region of interest of the image that has already been retrieved in the context:

$$\mathcal{A}_t \times \mathcal{O}_k \rightarrow \mathcal{O}_t, k \in \{1, 2, \dots, t-1\}, \quad (1)$$

Given a $w \times h$ image as an observation \mathcal{O}_k , a bounding box $[x_{min}, y_{min}, x_{max}, y_{max}]$ within perception tokens can precisely delineate the position of region \mathcal{R} , where (x_{min}, y_{min}) and (x_{max}, y_{max})

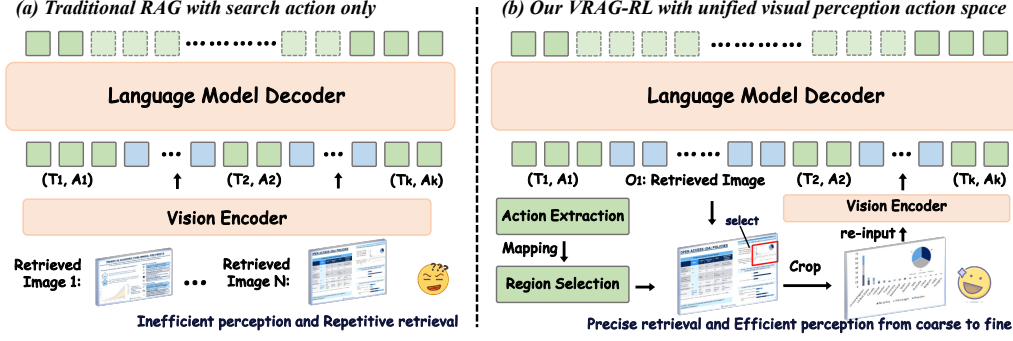


Figure 2: **Comparison between our VRAG-RL and the traditional RAG in terms of perception methods.** (a) Traditional methods lack effective perception, which easily leads to repetitive and ineffective retrieval calls and suboptimal outcomes. (b) Our VRAG-RL is efficient and accurate, enabling the model to perceive information-dense regions from a coarse-to-fine perspective.

represent the coordinates of the top-left and bottom-right pixels of region \mathcal{R} . Some current models' pre-training domains for grounding tasks normalize the coordinates to $[0, \delta]$, resulting in actual coordinates of $(x \times \frac{w}{\delta}, y \times \frac{h}{\delta})$, while other models, such as Qwen2.5VL, directly use the original coordinates without normalization. Then we will map the selected region \mathcal{R} from the image tokens in context to the $w_{raw} \times h_{raw}$ raw image, and crop this raw image to obtain $\hat{\mathcal{R}}$:

$$\hat{\mathcal{R}} = Crop(\mathbf{I}_{raw}, [x_{min} \times \frac{w_{raw}}{w_{encoder}}, y_{min} \times \frac{h_{raw}}{h_{encoder}}, x_{max} \times \frac{w_{raw}}{w_{encoder}}, y_{max} \times \frac{h_{raw}}{h_{encoder}}]). \quad (2)$$

where (w_{raw}, h_{raw}) are the shape of the original image \mathbf{I}_{raw} , $(w_{encoder}, h_{encoder})$ are determined by the vision encoder such that $w_{encoder} \times h_{encoder} = Pixels_{max}$. Finally, $\hat{\mathcal{R}}$ is integrated into the context as an observation: $\hat{\mathcal{R}} \rightarrow \mathcal{O}_t$. Actually, the image token embedded in the context does not represent the original size of the image. The maximum pixel size $Pixels_{max}$ for the vision encoder is often considerably smaller than the pixel of visually rich documents found in real-world applications. This is the reason why the region cropped from the original image and scaled within the vision encoder has a higher density of vision tokens. This simple yet effective "crop and re-input" strategy enhances visual perception performance by directly increasing perceptual resolution [50, 26, 37].

Trajectory Data Scaling-Up Based on Multi-Expert Sampling. To effectively train the model, especially smaller-scale models, to learn the utilization of Visual Perception Tokens while retaining their foundational capabilities, we need to train them with high-quality data through Supervised Fine-Tuning before applying RL. We propose a multi-expert sampling strategy to scale up the trajectory data, aiming to sample diverse interactions within the same reasoning trajectory for each data.

The core idea is to utilize large-scale models π_{LM} to effectively guide the reasoning process and tool selections within a trajectory, while smaller expert models π_{EM} annotate coordinate under the guidance of large-scale models. At the t_{th} interaction between the model and the environment:

$$\mathcal{H}_t = \{\mathcal{T}_1, \mathcal{A}_1, \mathcal{O}_1, \dots, \mathcal{O}_{t-1}, \mathcal{T}_t, \mathcal{A}_t, \mathcal{O}_t\}, \quad (3)$$

where \mathcal{H}_t is the trajectory, representing the sequence of past observations and actions leading up to the current step. The π_{LM} equipped with extensive capacities for understanding and processing complex multi-modal interactions, act as pioneers in determining the overarching reasoning pathway:

$$\{\mathcal{T}_t, \mathcal{A}_t\} = \pi_{LM}(\cdot \mid \mathcal{H}_{t-1}), \quad (4)$$

We use a rule-based function to extract action and thought. If the action is search, the engine returns the original image as \mathcal{O}_t . Otherwise, each time a visual perception token is output, we employ grounding-specific expert models to re-locate the coordinates of regions of interest:

$$\hat{\mathcal{A}}_t = \pi_{EM}(\cdot \mid \mathcal{H}_{t-1}; \mathcal{T}_t), \quad (5)$$

where the expert models π_{EM} benefit from the guidance provided by the large model's thought \mathcal{T}_t , leveraging these insights to enhance their precision in region localization. The newly generated

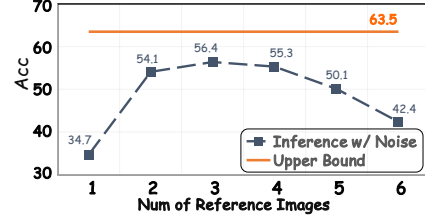
coordinates of the region of interest $\hat{\mathcal{A}}_t$ will replace the old visual perception tokens \mathcal{A}_t generated by π_{LM} , and the re-encoded image serves as observation $\hat{\mathcal{O}}_t$:

$$\hat{\mathcal{O}}_t = \mathcal{P}_V(\mathcal{O}_{t-1}, \hat{\mathcal{A}}_t). \quad (6)$$

where \mathcal{P}_V represents the visual processing function, the selected region will undergo cropping, zooming in, and re-encoding before being inserted into the context.

2.3 Fine-Grained Reward Function Tailored for Enhancing RAG Framework

Unlike traditional RL methods that focus only on output results, VRAG-RL emphasizes optimizing retrieval in RAG, as retrieval quality directly affects overall performance. We designed a reward function with three components: pattern reward, retrieval efficiency reward, and model-based outcome reward, guiding the model to efficiently retrieve information and generate high-quality answers.



Retrieval Efficiency Reward. As shown in Figure 3, when the information is sufficient, an excessively long context can interfere with the model. Therefore, the earlier and more comprehensive the retrieval of relevant information, the better the model can construct a coherent and informative context for generating high-quality answers. Inspired by Normalized Discounted Cumulative Gain, and using our predefined relevance of the recalled images, we define:

$$\text{DCG}(\mathcal{D}_{trj}) = \sum_{i=1}^{|\mathcal{D}_{trj}|} \frac{2^{s_i} - 1}{\log_2(i + 1)}, \quad s_i = \begin{cases} 1, & \text{if } d_i \in \mathcal{D}_{rel} \\ 0, & \text{if } d_i \notin \mathcal{D}_{rel} \end{cases}, \quad (7)$$

where $d_i \in \mathcal{D}_{trj}$ represents stacked retrieved images within the trajectory, \mathcal{D}_{rel} is the collection of relevant golden images, s_i is the predefined relevance score. We believe that the performance is optimal when all relevant documents are retrieved first, the Ideal-DCG is defined as:

$$\text{IDCG}(\mathcal{D}_{rel}) = \sum_{i=1}^{|\mathcal{D}_{rel}|} \frac{2^{s_{rel}} - 1}{\log_2(i + 1)} + \sum_{i=|\mathcal{D}_{rel}|+1}^n \frac{2^{s_{unrel}} - 1}{\log_2(i + 1)} = \sum_{i=1}^{|\mathcal{D}_{rel}|} \frac{1}{\log_2(i + 1)}, \quad (8)$$

where $s_{rel} = 1$ and $s_{unrel} = 0$ respectively represent the relevance scores of ideally relevant and irrelevant documents. Our Retrieval Efficiency Reward is defined as:

$$r_{Ret} = \frac{\text{DCG}(\mathcal{D}_{trj}, \mathcal{D}_{rel})}{\text{IDCG}(\mathcal{D}_{rel})}. \quad (9)$$

where r_{Ret} , the modified NDCG, is directly used as the reward to reflect retrieval performance.

Pattern Consistency and Model-Based Outcome Rewards. The rule-based pattern reward is designed to encourage the model to follow the reasoning patterns during the interaction process:

$$r_{Pat} \sim \text{Parse}(\mathcal{H}), \quad (10)$$

where \mathcal{H} is the generated trajectory. $\text{Parse}(\cdot)$ employ action tokens `<search>` and `</search>` to extract predefined actions in the action space. This is crucial for a reasoning agent with a predefined action space, as it helps effectively extract actions and thoughts. Regarding outcome reward, unlike rule-based methods that are prone to falling into local optima, we adopt a model-based reward:

$$r_{Ans} \sim \pi_{RM}(\cdot | \mathcal{Q}, \mathcal{A}_{golden}, \mathcal{A}_{pred}), \quad (11)$$

where \mathcal{Q} represents the input query, \mathcal{A}_{golden} is the reference golden answer, and \mathcal{A}_{pred} is the answer generated by the VLMs. Based on these inputs, the evaluation model π_{RM} assesses the correctness of the final answer. Please refer to Appendix A for the detailed prompt used in the model-based reward.

Integrated Reward Function. The final reward function is a weighted combination of the three components described above, with weights used to balance the contributions of each component:

$$r_\phi = \alpha \cdot r_{Ret} + \beta \cdot r_{Ans} + \gamma \cdot r_{Pat}. \quad (12)$$

where $\alpha + \beta + \gamma = 1$. In practice, we usually set $\gamma = 0$ as the model can effectively learn the pattern after SFT. We set $\gamma = 0.1$ when performing RL with cold start to help the model learn the predefined pattern. By integrating these three components into the reward function, our VRAG-RL provides a comprehensive and fine-grained evaluation mechanism that guides the model in optimizing its reasoning and retrieval capabilities in a way that aligns closely with real-world applications.

2.4 Reinforcement Learning Framework with Iterative Reasoning

We apply RL to multimodal RAG agent tasks to enhance the capability of VLMs in retrieving and reasoning. Our RL framework is primarily divided into two parts for discussion: the rollout process for multimodal agent and the reinforcement learning training strategy for multi-turn interactions.

Multi-Round Generation with Search Engine and Visual Perception Actions. As shown in Algorithm 1, the model interacts with the external environment in multiple turns, where the observation, which is the image, is inserted into the trajectory in the role of the user. This is necessary to align with the model’s pre-training domain, where only the user token can insert image tokens.

Algorithm 1 Interaction of VLM with the External Environment through Iterative Reasoning

Input: Input query x , Policy model π_θ , External environment \mathcal{V} , Maximum iterations T .

Output: Final trajectory y .

```

1: Initialize rollout sequence  $y \leftarrow \emptyset$  and action count  $t \leftarrow 0$ 
2: while  $t < T$  do
3:   Generate VLM response sequence  $y_t \sim \pi_\theta(\cdot \mid x, y)$ 
4:   Concatenate  $y_t$  to the  $y$  sequence with the role of assistant:  $y \leftarrow y + y_t$ 
5:   if <search> </search> detected in  $y_t$  then
6:     Extract search query  $q \leftarrow \text{Parse}(y_t)$  and Retrieve related image  $I_t = \text{Ret}(q)$ 
7:   else if <region> </region> detected in  $y_t$  then
8:     Extract visual perception tokens  $loc \leftarrow \text{Parse}(y_t)$  and Processing image  $I_t = P_V(loc, y)$ 
9:   else if <answer> </answer> detected in  $y_t$  then
10:    return final generated trajectory  $y$ 
11:   end if
12:   Concatenate vision tokens  $I_t$  to the sequence  $y$  with the role of user:  $y \leftarrow y + I_t$ 
13:   Increment action count  $t \leftarrow t + 1$ 
14: end while
15: return final generated trajectory  $y$ 

```

Training Strategy for Reinforcement Learning in Multi-Step Interactions. We propose a RL framework that enables VLM to learn how to interact with search engines and gather visually rich information from a coarse-to-fine perspective. The optimization objective is formulated as:

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot \mid x; \mathcal{V})} [r_\phi(x, y)] - \beta \mathbb{D}_{\text{KL}} [\pi_\theta(y \mid x; \mathcal{V}) \parallel \pi_{\text{ref}}(y \mid x; \mathcal{V})], \quad (13)$$

where the π_θ is the policy model, π_{ref} is the reference model, \mathbb{D}_{KL} is KL-divergence, and $y \sim \pi_\theta(\cdot \mid x; \mathcal{V}) = \pi_\theta(\cdot \mid x) \otimes \mathcal{V}$ is the rollout process. Our approach implements Group Relative Policy Optimization (GRPO) [13], which optimizes the model’s retrieval-augmented reasoning capability with group-sampled role-play trajectories. Please refer to Appendix C for more details.

3 Experiments

3.1 Experimental Settings

Datasets, Metric and Baselines. To evaluate the effectiveness of VRAG-RL, we compare our method with the text-based and vision-based baselines: (1) **Vanilla RAG** [11] uses the original question as a query for the search engine, then VLMs perform direct inference. (2) **ReAct** [48]: The model performs rewriting, retrieving, and reasoning in the think-then-act paradigm. (3) **Search-R1(-VL)** is the baseline adapted from Search-R1 [19], and the settings are aligned across all experiments

Table 1: **Main Results.** The best performance are marked in bold. SlideVQA and ViDoSeek mainly focus on reasoning type, while MMLongBench focuses on the visual type of reference content. OCR-based (🔍) RAG and purely visual (👁️) RAG are evaluated with the same prompt and setting.

METHOD	SLIDEVQA		ViDoSEEK		MMLONGBENCH					OVERALL
	Single-hop	Multi-hop	Extraction	Logic	Text	Table	Chart	Figure	Layout	
Qwen2.5-VL-3B-Instruct										
🔍 Vanilla RAG	15.1	12.1	8.8	14.3	3.9	5.1	1.7	3.1	2.5	11.2
🔍 ReAct [48]	11.8	9.9	5.3	7.4	6.5	3.7	3.9	5.2	2.5	8.4
🔍 Search-R1 [19]	17.5	13.8	13.3	20.7	3.4	3.2	4.5	4.1	6.8	14.1
👁️ Vanilla RAG	19.4	12.2	10.1	17.3	2.2	4.1	5.2	4.7	4.3	13.2
👁️ ReAct [48]	15.7	10.9	6.7	14.2	2.7	3.6	3.4	3.1	5.1	10.9
👁️ Search-R1-VL [19]	26.3	20.1	20.1	29.8	8.5	7.8	7.9	9.3	7.6	21.3
👁️ VRAG-RL (Ours)	65.3	38.6	63.1	73.8	22.7	16.1	21.9	21.4	19.5	53.5
Qwen2.5-VL-7B-Instruct										
🔍 Vanilla RAG	26.1	10.6	24.7	30.9	8.5	5.4	11.7	4.4	3.3	20.9
🔍 ReAct [48]	21.2	13.3	14.3	21.3	5.9	5.1	7.3	5.5	1.7	15.8
🔍 Search-R1 [19]	28.4	19.7	20.8	30.6	9.9	6.0	7.9	10.1	5.9	22.2
👁️ Vanilla RAG	29.1	17.4	26.4	41.3	13.1	14.7	15.9	4.3	7.6	24.2
👁️ ReAct [48]	34.8	20.4	27.5	42.1	10.1	12.4	10.2	6.2	7.1	26.9
👁️ Search-R1-VL [19]	48.3	42.3	40.5	50.3	19.9	13.4	12.9	11.4	10.2	37.4
👁️ VRAG-RL (Ours)	69.3	43.1	60.6	74.8	26.1	26.3	24.8	25.9	21.2	57.1

to ensure fairness. We evaluate our method on three challenging, visually rich benchmarks: **ViDoSeek** [41], **SlideVQA** [40] and **MMLongBench** [29]. The model-based evaluation metric is binary 0 or 1, indicating the accuracy of the model’s responses. Please refer to Appendix E and F for more details.

Training and Inference Setups. We conducted SFT and RL on llama-factory [53] and verl [38] respectively. We use full parameter fine-tuning and cosine learning scheduler with a warmup ratio of 0.1 during SFT. When training with the GRPO algorithm, we set the group size to 5 and the coefficient for the KL loss is typically set to 0.01, but if we perform cold start, we set it to 0 to disable the KL loss constraint on the model. During training and inference, we built a search engine from a database of approximately $\sim 70k$ visual documents. All the experiments are conducted on 8 NVIDIA A100 80G GPUs. Please refer to Appendix G for detailed hyperparameters used in our paper.

3.2 Results

Main Results. As shown in Table 1, compared to purely visual methods, OCR-based methods exhibit significant limitations on visually intensive benchmarks. On the one hand, visual information inherently contains elements that cannot be represented by text, such as element positions, layout, and color, etc. On the other hand, the perceptual capabilities of OCR models are considerably inferior to those of the current advanced VLMs, which restricts the overall performance ceiling of the framework. Visual-based methods have proven to be a more elegant solution compared to OCR-based methods, especially in tasks related to visual understanding. For prompt-based baselines of vision domain, Vanilla RAG and ReAct exhibit poor performance, far behind RL-based baselines and our method on various benchmarks. The 7B model, compared to the 3B model, possesses superior perception and understanding capabilities, exhibiting strong performance across various datasets. For RL-based baselines, our method also performs better than search-R1-VL on both Qwen2.5-VL-7B-Instruct ($34.7 \rightarrow 57.1$) and Qwen2.5-VL-3B-Instruct ($21.3 \rightarrow 53.5$). The evaluation results on SlideVQA and ViDoSeek demonstrate our model’s significant improvement in reasoning capabilities across various reasoning tasks. Furthermore, as MMLongBench includes multiple visual elements, which indicates the model’s improvement in visual perception capabilities, this phenomenon is related to our proposed visual perception action space. The results across various benchmarks prove the effectiveness and generalization of our method in the retrieval and reasoning of visually rich information.

Approach Ablations. As shown in Table 2, taking Qwen2.5-VL-7B-Instruct as an example, we decompose the key components of VRAG-RL to examine the impact of different rewards and action space on performance separately. In a macro view, removing each module results in a clear drop in the accuracy, which validates the power of our RAG-specific reward and Visual-perception action space.

The action space module we defined shows a certain degree of improvement on different bases, which **proves the effectiveness of the visual perception-based strategy**. Consistent with the findings demonstrated in MMLongBench in Figure 5, the visual perception action space we introduced has generally enhanced the framework’s performance, particularly in improving high-density visual information. Furthermore, ablation experiments on the reward model further demonstrate that retrieving relevant information is a prerequisite for high-quality generation, highlighting the role of high-quality retrieval in RAG, which **proves the importance of our RAG-specific reward**. Comparisons and analyses of experiments across different settings collectively demonstrate the effectiveness and generalization of our modules, and their combination comprehensively enhances end-to-end performance from various perspectives.

3.3 Analysis

Better retrieval facilitates high-quality generation.

Our VRAG-RL framework significantly enhances the retrieval efficiency, which is crucial for constructing a coherent and informative context for high-quality generation. As demonstrated in Figure 3, the context length has a substantial impact on model performance. When the context is too long, it can introduce noise and interfere with the model’s ability to generate accurate answers. In contrast, when relevant information is retrieved early and comprehensively, the model can build a more focused and informative context. As shown in Figure 4, our model is more effective at retrieving relevant information compared to traditional prompt-based rewrite methods. Our approach provides the vision model with a better context for generating high-quality answers.

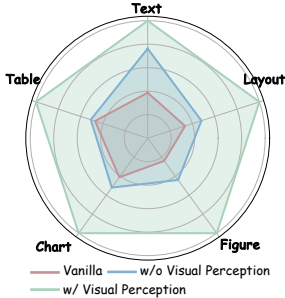


Figure 5: Relative performance on MMLongBench.

Visual perception action space provides a fine-grained perspective.

The visual perception action space introduced in our framework further enhances understanding by allowing the model to focus on information-dense regions of images. Figure 5 illustrates the relative performance comparison between our approach with visual perception action space and various baselines, from which we can observe that VRAG-RL not only performs well in textual tasks but also shows noticeable improvements in tasks requiring visual perception abilities, particularly in Layout, Chart, and Figure. This is particularly important given the current limitations in computational resources, especially considering that VLMs are highly memory-intensive. Using this dynamic resolution strategy, the model can achieve more detailed perception within the constraints of limited computational resources, rather than simply maximizing the resolution of the original image. Our method achieves an improvement in perceptual abilities while optimizing resource utilization. Perhaps this human-like way of thinking and acting is the key to AGI.

Reinforcement learning helps the model to perform multi-step reasoning effectively.

One major challenge of the prompt-based method is that as the number of interactions increases, the model’s capability to follow instructions weakens. However, pre-training with SFT helps the model reason in a pre-defined pattern compared to cold start, but it also impacts the model’s inherent foundational capabilities to some extent. To further explore the activation of multi-turn reasoning abilities in models by RL, we compared the iterative reasoning performance of models with and without RL, as shown in Table 3. For our method with action space, effective actions are crucial for interacting with the external environment. The Invalid Action Rate indicates incorrect action responses, which include not only pattern errors but also hallucinations caused by wrong cropping, answering before retrieval, and so on. Inefficient reasoning often includes repeated

Table 2: Ablation study on three benchmarks.

Vanilla	REWARD		ACTION SPACE		Accuracy
	Vanilla	RAG-Specific	Search	Visual-Perception	
✓	✓		✓		47.2
			✓	✓	49.3
		✓	✓		54.9
		✓	✓	✓	57.1

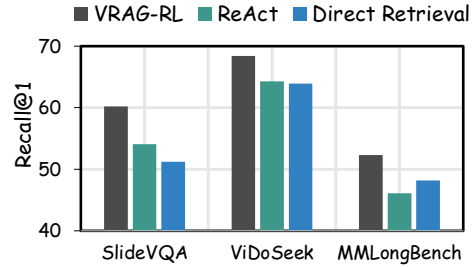


Figure 4: The retrieval performance of traditional prompt-based RAG and our approach.

Table 3: Average Finish Rate (%) and Average Invalid Action Rate (%).

Method	Invalid Action Rate ↓	Finish Rate ↑
SFT	9.4	84.2
+ RL	5.1	97.1

meaningless searches, leading to a decrease in the finish rate. Our method with RL effectively reduces the invalid rate and increases the finish rate. It guides the model to make optimal decisions at each step of the reasoning process, enabling it to flexibly adjust strategies when faced with different types of out-of-domain visual information, thereby better completing complex reasoning tasks.

Model-based reward offers more stable training compared to rule-based reward. Previous works often use EM as the reward, which is too strict. Unlike short answers for data-related questions, it is difficult for the model’s responses to exactly match the golden answer, resulting in inefficient training. However, using recall as a reward may lead to misjudgments and cause models to hack the function, resulting in repetitive responses that destabilize training. In contrast, a model-based reward leverages an evaluation model to assess the quality and relevance of generated responses in a more flexible manner. This approach not only aligns better with real-world applications but also provides a more stable and effective training signal, as demonstrated in Appendix A. The model-based reward thus enables VRAG-RL to achieve more robust performance across visual reasoning tasks.

Time efficiency. As shown in Figure 6, our method’s multi-turn interaction with external environments can lead to increased latency. The latency of vanilla RAG remains consistent, as it only performs a single search and provides an answer. ReAct RAG, a prompt-based method, also demonstrates multi-turn interaction capabilities due to the fundamental reasoning abilities of the model. However, it is limited to only two defined actions: answer and search. Due to the lack of sufficient perception capabilities, it often falls into repetitive search loops. Our approach equips the model with a visual perception space that can effectively understand visually rich images. The model can quickly extract answers after retrieval, thus avoiding ineffective searches. Despite the increase in latency, the overall performance improves due to the higher quality of generated answers, making the trade-off between latency and accuracy highly beneficial for visually rich retrieval and understanding tasks.

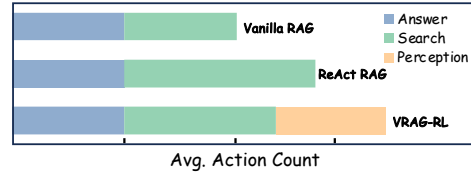


Figure 6: Latency Analysis on Generation.

Case Study. In Figure 7 and 8 (Appendix H), we list the trajectories of our VRAG-RL to illustrate how our model reasons and interacts with the environment. These cases highlight two challenges in visually rich information RAG: (1) accurately retrieving relevant images, and (2) the reference information often requires higher-resolution perception. In Figure 7, we can observe that the model demonstrated reflective capability, and eventually identified subtle clues in the relevant images. Moreover, as shown in Figure 8, the model engages in visual perception actions only when required, showcasing human-like reasoning instead of simply replicating patterns from its training data.

4 Related Work

Vision-based Retrieval-augmented Generation. RAG demonstrates significant advantages in addressing knowledge-intensive problems [22, 12, 2]. Traditional text-based RAG methods typically involve designing different agents to interact with search engines [45, 5, 6, 44, 25, 32, 21, 10]. However, with the widespread adoption of electronic documents, knowledge is no longer confined to text. Recently, there has been an increasing amount of research on OCR-free retrieval methods that directly align textual queries with images [51, 11]. Furthermore, more and more work is focusing on multimodal RAG agents [41, 8, 16, 24, 46], enabling more accurate retrieval and extraction of visual information. Our work builds upon these developments by incorporating visual perception actions into visual-based RAG, effectively activating the reasoning and understanding capabilities of VLMs.

Reinforcement Learning with Large Models. Reasoning capabilities are crucial for models to effectively address complex problems, and RL has been proven to be a powerful approach to enhance these capabilities [13, 15]. Previous work applied RL in the training of LLMs [31, 43, 35, 36, 13, 27]. Additionally, more and more works aim to use RL to enhance the reasoning capabilities of VLMs [3, 30, 28, 52]. Recent advancements have seen RL being widely applied to the training of large model-driven agents [42]. These agents, especially RAG agents, require robust multi-step reasoning capabilities to interact effectively with external environments [17, 23]. However, there is still a scarcity of RL frameworks specifically tailored for multimodal iterative reasoning, which is essential

for handling visually rich information. Our work aims to fill this gap by introducing a novel RL framework that enables VLMs to perform iterative reasoning with visual perception actions, thereby enhancing their reasoning capabilities in complex, multi-modal retrieval-augmented reasoning tasks.

5 Conclusion and Future Work

In this paper, we introduce VRAG-RL, a novel reinforcement learning framework tailored for complex reasoning across visually rich information. Our approach enables Vision Language Models to interact with search engines more effectively, significantly enhancing their reasoning and retrieval capabilities. Extensive evaluations on various benchmarks have demonstrated significant advantages in visual information reasoning, retrieval, and understanding with our model. For future work, we plan to introduce more actions that mimic how humans handle complex information, allowing the model to focus more on deep thinking. Additionally, we aim to reduce hallucinations by leveraging more advanced models, further improving the accuracy and reliability of our framework.

Acknowledgments

This work was supported by the Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC. The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Sciences.

References

- [1] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
- [2] Chen, J., Lin, H., Han, X., Sun, L.: Benchmarking large language models in retrieval-augmented generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 17754–17762 (2024)
- [3] Chen, L., Li, L., Zhao, H., Song, Y., Vinci: R1-v: Reinforcing super generalization ability in vision-language models with less than \$3. <https://github.com/Deep-Agent/R1-V> (2025), accessed: 2025-02-02
- [4] Chen, M., Li, T., Sun, H., Zhou, Y., Zhu, C., Wang, H., Pan, J.Z., Zhang, W., Chen, H., Yang, F., et al.: Research: Learning to reason with search for llms via reinforcement learning. arXiv preprint arXiv:2503.19470 (2025)
- [5] Chen, Z., Liu, K., Wang, Q., Liu, J., Zhang, W., Chen, K., Zhao, F.: Mindsearch: Mimicking human minds elicits deep ai searcher. arXiv preprint arXiv:2407.20183 (2024)
- [6] Chen, Z., Liu, K., Wang, Q., Zhang, W., Liu, J., Lin, D., Chen, K., Zhao, F.: Agent-flan: Designing data and methods of effective agent tuning for large language models. arXiv preprint arXiv:2403.12881 (2024)
- [7] Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
- [8] Cho, J., Mahata, D., Irsoy, O., He, Y., Bansal, M.: M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. arXiv preprint arXiv:2411.04952 (2024)
- [9] Du, Y., Li, C., Guo, R., Yin, X., Liu, W., Zhou, J., Bai, Y., Yu, Z., Yang, Y., Dang, Q., et al.: Pp-ocr: A practical ultra lightweight ocr system. arXiv preprint arXiv:2009.09941 (2020)
- [10] Fang, Y., Sun, T., Shi, Y., Gu, X.: Attentionrag: Attention-guided context pruning in retrieval-augmented generation. arXiv preprint arXiv:2503.10720 (2025)
- [11] Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., Colombo, P.: Colpali: Efficient document retrieval with vision language models. In: The Thirteenth International Conference on Learning Representations. pp. 1–26 (2025)
- [12] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 2, 1 (2023)
- [13] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
- [14] Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025)
- [15] Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al.: Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024)
- [16] Jiang, D., Zhang, R., Guo, Z., Wu, Y., Lei, J., Qiu, P., Lu, P., Chen, Z., Fu, C., Song, G., et al.: Mmsearch: Benchmarking the potential of large models as multi-modal search engines. arXiv preprint arXiv:2409.12959 (2024)
- [17] Jiang, P., Lin, J., Cao, L., Tian, R., Kang, S., Wang, Z., Sun, J., Han, J.: Deepretrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. arXiv preprint arXiv:2503.00223 (2025)

- [18] Jin, B., Yoon, J., Han, J., Arik, S.O.: Long-context llms meet rag: Overcoming challenges for long inputs in rag. arXiv preprint arXiv:2410.05983 (2024)
- [19] Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H., Han, J.: Search-r1: Training llms to reason and leverage search engines with reinforcement learning. arXiv preprint arXiv:2503.09516 (2025)
- [20] Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey. *Journal of Artificial Intelligence Research* **4**, 237–285 (1996)
- [21] Lee, C., Roy, R., Xu, M., Raiman, J., Shoenybi, M., Catanzaro, B., Ping, W.: Nv-embed: Improved techniques for training llms as generalist embedding models. arXiv preprint arXiv:2405.17428 (2024)
- [22] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* **33**, 9459–9474 (2020)
- [23] Li, X., Dong, G., Jin, J., Zhang, Y., Zhou, Y., Zhu, Y., Zhang, P., Dou, Z.: Search-o1: Agentic search-enhanced large reasoning models. arXiv preprint arXiv:2501.05366 (2025)
- [24] Li, Y., Li, Y., Wang, X., Jiang, Y., Zhang, Z., Zheng, X., Wang, H., Zheng, H.T., Huang, F., Zhou, J., et al.: Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. arXiv preprint arXiv:2411.02937 (2024)
- [25] Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., Zhang, M.: Towards general text embeddings with multi-stage contrastive learning. arXiv preprint arXiv:2308.03281 (2023)
- [26] Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 26296–26306 (2024)
- [27] Liu, R., Wang, J., Shi, Y., Xie, Z., An, C., Zhang, K., Zhao, J., Gu, X., Lin, L., Hu, W., et al.: Attention as a compass: Efficient exploration for process-supervised rl in reasoning models. arXiv preprint arXiv:2509.26628 (2025)
- [28] Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785 (2025)
- [29] Ma, Y., Zang, Y., Chen, L., Chen, M., Jiao, Y., Li, X., Lu, X., Liu, Z., Ma, Y., Dong, X., et al.: Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. arXiv preprint arXiv:2407.01523 (2024)
- [30] Meng, F., Du, L., Liu, Z., Zhou, Z., Lu, Q., Fu, D., Shi, B., Wang, W., He, J., Zhang, K., et al.: Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. arXiv preprint arXiv:2503.07365 (2025)
- [31] Meng, Y., Xia, M., Chen, D.: Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems* **37**, 124198–124235 (2024)
- [32] Moreira, G.d.S.P., Osmulski, R., Xu, M., Ak, R., Schifferer, B., Oldridge, E.: Nv-retriever: Improving text embedding models with effective hard-negative mining. arXiv preprint arXiv:2407.15831 (2024)
- [33] OpenAI: Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/> (2024)
- [34] Pichai, S., Hassabis, D., Kavukcuoglu, K.: Introducing gemini 2.0: our new ai model for the agentic era (2024)
- [35] Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* **36**, 53728–53741 (2023)
- [36] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)

- [37] Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems* **37**, 8612–8642 (2024)
- [38] Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., Wu, C.: Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256* (2024)
- [39] Sutton, R.S., Barto, A.G., et al.: Reinforcement learning. *Journal of Cognitive Neuroscience* **11**(1), 126–134 (1999)
- [40] Tanaka, R., Nishida, K., Nishida, K., Hasegawa, T., Saito, I., Saito, K.: Slidevqa: A dataset for document visual question answering on multiple images. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. pp. 13636–13645 (2023)
- [41] Wang, Q., Ding, R., Chen, Z., Wu, W., Wang, S., Xie, P., Zhao, F.: Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017* (2025)
- [42] Wang, Z., Wang, K., Wang, Q., Zhang, P., Li, L., Yang, Z., Jin, X., Yu, K., Nguyen, M.N., Liu, L., et al.: Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073* (2025)
- [43] Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* **8**, 229–256 (1992)
- [44] Wu, J., Yin, W., Jiang, Y., Wang, Z., Xi, Z., Fang, R., Zhang, L., He, Y., Zhou, D., Xie, P., et al.: Webwalker: Benchmarking llms in web traversal. *arXiv preprint arXiv:2501.07572* (2025)
- [45] Wu, W., Huang, S., Jiang, Y., Xie, P., Huang, F., Zhao, H.: Unfolding the headline: Iterative self-questioning for news retrieval and timeline summarization. *arXiv preprint arXiv:2501.00888* (2025)
- [46] Xia, P., Zhu, K., Li, H., Zhu, H., Li, Y., Li, G., Zhang, L., Yao, H.: Rule: Reliable multimodal rag for factuality in medical vision language models. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. pp. 1081–1093 (2024)
- [47] Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115* (2024)
- [48] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: React: Synergizing reasoning and acting in language models. In: *The Eleventh International Conference on Learning Representations*. pp. 1–33 (2023)
- [49] Yu, E., Lin, K., Zhao, L., Yin, J., Wei, Y., Peng, Y., Wei, H., Sun, J., Han, C., Ge, Z., et al.: Perception-rl: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954* (2025)
- [50] Yu, R., Ma, X., Wang, X.: Introducing visual perception token into multimodal large language model. *arXiv preprint arXiv:2502.17425* (2025)
- [51] Yu, S., Tang, C., Xu, B., Cui, J., Ran, J., Yan, Y., Liu, Z., Wang, S., Han, X., Liu, Z., et al.: Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594* (2024)
- [52] Zeng, Y., Huang, W., Huang, S., Bao, X., Qi, Y., Zhao, Y., Wang, Q., Chen, L., Chen, Z., Chen, H., et al.: Agentic jigsaw interaction learning for enhancing visual perception and reasoning in vision-language models. *arXiv preprint arXiv:2510.01304* (2025)
- [53] Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., Ma, Y.: Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372* (2024)

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: The abstract and introduction clearly state the contributions and scope.

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: The paper includes potential limitations of the proposed method.

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: [\[Yes\]](#)

Justification: Theoretical results are accompanied by detailed assumptions and complete proofs.

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: The paper provides detailed instructions for reproducing the experimental results.

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: The paper provide data and detailed instructions for reproduction.

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: The paper provides a detailed description of the experimental setup.

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: The experimental results were conducted multiple times and averaged, providing data and error analysis.

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: The paper provides the details of them.

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: The research adheres to all guidelines outlined in the NeurIPS Code of Ethics.

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: The paper discusses both the potential positive and negative societal impacts.

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: The paper does not involve high-risk data or models.

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: All existing assets are properly credited.

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: New assets are well documented.

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 or 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: The paper does not involve research with human 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: The paper describe the usage of LLMs and VLMs.

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.

Appendix

A Model-Based Reward

We employ a model-based reward to evaluate the quality and relevance of generated responses. Specifically, we utilize Qwen2.5-7B-Instruct [47] as our reward model. This model is deployed on 4 NVIDIA A100 GPUs to enable efficient batch evaluation. The prompt used for the reward model is illustrated in Figure 12. Given the input query, reference answer, and generated response, the reward model assesses the correctness of the generated response and outputs a binary value (0 or 1) to represent the accuracy of the answer. Compared to the rule-based reward like exact match (EM) or Recall, used in previous work [19, 3], our model-based reward provides a more flexible and comprehensive evaluation of the generated response. This leads to higher training efficiency and better generalization to diverse datasets.

B The implementation of the search engine

To effectively support the retrieval-augmented generation tasks in our VRAG-RL framework, we implemented OCR-based and vision-based pipeline separately. The vision-based retriever is built upon the state-of-the-art embedding model ColPali [11], which is specifically designed for aligning textual queries with images. For the textual retrieval pipeline, we employ the PP-OCR [9] to extract text from images. We utilize the Llama-Index to ensure an efficient indexing and querying mechanism for large-scale image datasets. In our experiments, we deployed the search engine on a single NVIDIA A100 80G GPU, allowing us to handle large-scale queries efficiently. The use of batch querying further optimizes the retrieval speed, making it suitable for real-time applications.

C Reinforcement Learning Framework with GRPO

Our framework implements the Group Relative Policy Optimization (GRPO), which leverages the average reward of multiple sampled outputs as a baseline rather than relying on a learned value function. The policy model is optimized by maximizing the following objective function:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{x \sim \mathcal{D}, \{y_i\}_{i=1}^G \sim \pi_{\text{old}}(\cdot|x; \mathcal{V})} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{\sum_{t=1}^{|y_i|} I(y_{i,t})} \sum_{t=1: I(y_{i,t})=1}^{|y_i|} \min \left(\frac{\pi_{\theta}(y_{i,t}|x, y_{i,<t}; \mathcal{V})}{\pi_{\text{old}}(y_{i,t}|x, y_{i,<t}; \mathcal{V})} \hat{A}_{i,t}, \right. \right. \\ \left. \left. \text{clip} \left(\frac{\pi_{\theta}(y_{i,t}|x, y_{i,<t}; \mathcal{V})}{\pi_{\text{old}}(y_{i,t}|x, y_{i,<t}; \mathcal{V})}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right) - \beta \mathbb{D}_{KL} [\pi_{\theta} || \pi_{\text{ref}}] \right]$$

where rollout module samples a group of trajectories $\{y_1, y_2, \dots, y_G\}$ from the reference policy π_{ref} for each input question x by interacting with the external environment \mathcal{V} . $\hat{A}_{i,t}$ represent the advantage, computed based on the relative rewards of outputs within each group.

D Expert Trajectories Collection

Data Collection. To train our model effectively, we collected expert trajectories using Qwen-VL-max-latest for prompt-based data collection. Specifically, we utilized the React-based prompt to gather data, ensuring that the model could perform complex reasoning tasks. During the data collection process, whenever grounding was required to focus on specific regions of interest within images, we employed Qwen2.5VL-72B to perform the grounding tasks. This was done under the guidance of the historical trajectories.

Data Proportions. To ensure that our model could perform diverse multi-step reasoning during Reinforcement Learning (RL), we carefully balanced the training data. Specifically, we balanced the trajectories based on the number of steps (2-6) and the types of actions involved (search and perception). This approach ensured that the model was exposed to a wide range of reasoning tasks and could learn to handle different types of interactions with the environment effectively.

E Dataset Information

We evaluate our method on three visually rich document datasets: SlideVQA, ViDoSeek, and MMLongbench.

1. **SlideVQA** [40] is a dataset for document visual question answering focused on understanding slides. It contains over 2,600 slide decks with more than 52,000 slide images and 14,500 questions that require complex reasoning skills such as single-hop, multi-hop, and numerical reasoning. The dataset is designed to support various reasoning types and includes annotated arithmetic expressions for numerical questions to enhance reasoning capabilities.
2. **ViDoSeek** [41] is a dataset specifically designed for visually rich document retrieval-reason-answer tasks. It aims to evaluate the performance of RAG systems on large-scale document collections. Unlike traditional VQA datasets that focus on single images or documents, ViDoSeek contains queries with unique answers across a collection of approximately 6,000 images, covering diverse content types such as text, charts, tables, and layouts. This dataset provides a more comprehensive and challenging benchmark for evaluating the retrieval and reasoning capabilities of RAG models in real-world scenarios.
3. **MMLongbench** [29] is a dataset designed to evaluate the document understanding capabilities of VLMs with an emphasis on long-context, multi-modal documents composed of text, images, charts, tables, and layout structures.

Table 4: Statistics of datasets.

Dataset	Total Questions	Corpus Size	Visual Elements
SlideVQA-Test	2020	8000	Text, Chart, Table, Layout
SlideVQA-Train	12268	44359	Text, Chart, Table, Layout
ViDoSeek	1142	5400	Text, Chart, Table, Layout
MMLongBench	847	6492	Text, Chart, Figure, Table, Layout

F Compared Baselines

Here we detailedly introduce the baselines we compare with and our re-produce details.

1. **Vanilla RAG**. There are two types of Vanilla RAG: text-based and visual-based. Text-based Vanilla RAG uses text as the retrieval corpus, which is reflected in text search engines and text modality generation. During the retrieval phase, it directly uses the original question to search for relevant text, which is then inserted into the context to answer the question. Visual-based Vanilla RAG uses images as the corpus. During the retrieval phase, it directly uses the original question to search for relevant images, which are then inserted into the context to answer the question.
2. **ReAct RAG** [48]. The method incorporates Chain-of-Thought (COT) prompting in RAG agent tasks with a format of a Thought-Action-Observation loop. The main difference between text-based and visual-based approaches lies in the retrieval corpus of the search engine and the modality of the information inserted.
3. **Search-R1** [19]. The method introduces multi-turn reasoning RL into the text RAG. We used our framework for reproducing, which includes multi-turn interactions and rule-based rewards.
4. **Search-R1-VL**. This is a vision-based baseline implemented on our framework based on search-R1. We used the same reward and post-process methods and trained models based on cold start with the same dataset as VRAG-RL.

G Hyperparameters

The detailed hyperparameters we use during training are shown in Table 5 and Table 6. We employ identical hyperparameters for different models.

Table 5: Key hyperparameters for SFT.

Name	Value
Finetuning type	Full
Freeze vision tower	True
Freeze multi-modal projector	True
Freeze language model	False
Cutoff len	16384
Epochs	3
Batch size	16
Gradient accumulation steps	2
Learning rate	1.0e-5
LR scheduler type	cosine
Warmup ratio	0.1

Table 6: Key hyperparameters for RL.

Name	Value
Number of agent groups	5
Warmup steps ratio	0.285
Mini batch size	64
Micro batch size per GPU	2
Learning rate (Actor)	1.0e-6
KL loss coefficient	0.01 (optional)
Tensor model parallel size	4
Total epochs	1
Max prompt length	8192
Max response length	2048
GPU memory utilization	0.6

H Case Study

In Figure 7 and 8, we list the trajectories of our VRAG-RL to illustrate how our model reasons and interacts with the environment. These cases highlight two challenges in visually rich information RAG: (1) accurately retrieving relevant images, and (2) the reference information often requires higher-resolution perception. In Figure 7, we can observe that the model demonstrated reflective capability, and eventually identified subtle clues in the relevant images. Moreover, as shown in Figure 8, the model engages in visual perception actions only when required, showcasing human-like reasoning instead of simply replicating patterns from its training data.

I Limitations

Despite our best efforts, this paper still has some limitations: (1) Retriever Accuracy: The accuracy of the current retriever still needs improvement to better support the RAG system. (2) Base Model Capability: Enhancing the capabilities of the base model can contribute to the overall improvement of the system. (3) Real-time Requirements: The current multi-turn interactions may not meet the requirements of high real-time applications.

J Ethics Statement

Our research complies with the NeurIPS Code of Ethics. We used publicly available datasets (as detailed in Appendix E), avoiding ethical concerns related to privacy, confidentiality, or misuse of personal biological information. The VRAG-RL framework aims to improve visual information understanding for beneficial applications. We emphasize responsible use and remain vigilant against misuse.

K Prompts

In this section, we illustrate all the prompts used in our paper. Part of our prompts are taken from Search-R1 [19].

K.1 Vanilla RAG Prompt

See Figure 11.

K.2 Search-R1 Prompt

See Figure 10.

K.3 ReAct RAG Prompt

ReAct RAG uses the same prompt as Search-R1, as shown in Figure 10.

K.4 VRAG-RL Prompt

See Figure 9.

K.5 Model-based Reward Prompt

See Figure 12.



Question: How many more Dugwells are there than Shallow Tubewells in the years of India's Energy Divide whose map was prepared by IWMI based on Minor Irrigation Census?



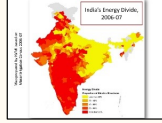
<think> To answer this question, I need to find the specific data related to Dugwells and Shallow Tubewells in India's Energy Divide map prepared by IWMI based on Minor Irrigation Census. This requires searching for the relevant information.

</think>

<search> India's Energy Divide IWMI Minor Irrigation Census Dugwells Shallow Tubewells **</search>**



Image Retrieval



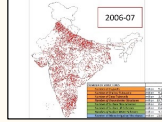
<think> The image provided is a map showing the proportion of electric structures in India's Energy Divide. It indicates the proportion of electric structures using different color codes. However, it does not provide specific numbers for Dugwells and Shallow Tubewells. I need to search for more detailed data on Dugwells and Shallow Tubewells from the Minor Irrigation Census.

</think>

<search> Detailed data on Dugwells and Shallow Tubewells data **</search>**



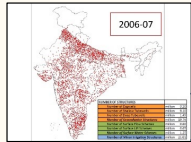
Image Retrieval



<think> The image contains a table with the number of structures, including Dugwells and Shallow Tubewells. I need to crop this part of the image to focus on the relevant data.

</think>

<region> [361, 366, 721, 517] **</region>**



NUMBER OF STRUCTURES	
Number of Dugwells	million 9.20
Number of Shallow Tubewells	million 9.10
Number of Deep Tubewells	million 1.45
Number of Groundwater Structures	million 19.76
Number of Surface Flow Schemes	million 0.60
Number of Surface Lift Schemes	million 0.65
Number of Surface Water Schemes	million 1.25
Number of Minor Irrigation Structures	million 21.01

Select & Crop & Zoom in



<think> The cropped image provides the specific numbers for Dugwells (9.20 million) and Shallow Tubewells (9.10 million). To find how many more Dugwells there are than Shallow Tubewells, I need to subtract the number of Shallow Tubewells from the number of Dugwells.

</think>

<answer> There are 0.10 million (or 100,000) more Dugwells than Shallow Tubewells. **</answer>**

Figure 7: Case 1 for VRAG-RL.



Question: Which H2 antagonist has a lower bioavailability: the drug that inhibits several CYP-450 isoenzymes and reduces hepatic blood flow or Famotidine?



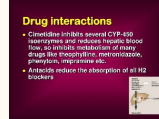
<think> To answer this question, I need to identify the H2 antagonist that inhibits several CYP-450 isoenzymes and reduces hepatic blood flow. Then, I will compare its bioavailability with that of Famotidine.

</think>

<search> H2 antagonist that inhibits several CYP-450 isoenzymes **</search>**



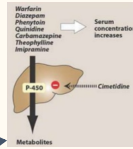
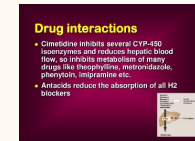
Image Retrieval



<think> The image provides information about drug interactions, specifically mentioning Cimetidine's effect on CYP-450 isoenzymes and hepatic blood flow. This aligns with the search query. I need to look at the lower right corner of the image to confirm this.

</think>

<region> [556, 337, 722, 505] **</region>**



Select & Crop & Zoom in



<think> The diagram confirms that Cimetidine inhibits the P-450 enzymes, leading to increased serum concentrations of drugs metabolized by these enzymes. Now, I need to find information on the bioavailability of Cimetidine and Famotidine to determine which has lower bioavailability.

</think>

<search> bioavailability of Cimetidine and Famotidine **</search>**



Image Retrieval

Comparison of H ₂ antagonists				
	Cimetidine	Famotidine	Famotidine	Cimetidine
Bioavailability	80	40	40	<40
Relative Potency	1	0.15	30	0.15
Half-life (hrs)	1.5-2.5	1.5-2.5	2.5-4	1.1-1.8
LD50 (mg/kg)	4	4	12	4
Protein of CYP450	1	0.1	0	0
Mean log (SD)	400	150	30	150



<think> Based on the provided table, Cimetidine has a bioavailability of 80%, while Famotidine has a bioavailability of 40%. Therefore, Famotidine has a lower bioavailability compared to Cimetidine. The drug that inhibits several CYP-450 isoenzymes and reduces hepatic blood flow is Cimetidine, as indicated in previous information.

</think>

<answer> Famotidine has a lower bioavailability (40%) compared to Cimetidine (80%). Therefore, the H2 antagonist with lower bioavailability is Famotidine. **</answer>**

Figure 8: Case 2 for VRAG-RL.

VRAG-RL Prompt.

System Prompt:

Answer the given question. You must conduct reasoning inside `<think>` and `</think>` first every time you get new information. After reasoning, if you find you lack some knowledge, you can call a search engine by `<search> query </search>` and user will return the searched results. Every time you retrieve an image, you have the option to crop it to obtain a clearer view, the format for coordinates is `<bbox>[x1, y1, x2, y2]</bbox>`. You can search as many times as your want. If you find no further external knowledge needed, you can directly provide the answer inside `<answer>` and `</answer>`, without detailed illustrations. For example, `<answer> Beijing </answer>`.

User Prompt:

Query: {Query Description}

Figure 9: Prompt of VRAG-RL.

Search-R1(-VL) Prompt.

System Prompt:

Answer the given question. You must conduct reasoning inside `<think>` and `</think>` first every time you get new information. After reasoning, if you find you lack some knowledge, you can call a search engine by `<search> query </search>` and user will return the searched results. You can search as many times as your want. If you find no further external knowledge needed, you can directly provide the answer inside `<answer>` and `</answer>`, without detailed illustrations. For example, `<answer> Beijing </answer>`.

User Prompt:

Query: {Query Description}

Figure 10: Prompt of Search-R1(-VL) and ReAct RAG.

Vanilla RAG Prompt.

System Prompt:

Answer the given question. You must conduct reasoning inside `<think>` and `</think>` first every time you get new information. After reasoning, you should directly provide the answer inside `<answer>` and `</answer>`, without detailed illustrations. For example, `<answer> Beijing </answer>`.

User Prompt:

Query: {Query Description}

Reference: {Retrieved Images / Text Tokens}

Figure 11: Prompt of Vanilla RAG.

Reward Model Prompt.

System Prompt:

Character Introduction

You are an expert evaluation system for a question answering chatbot.

You are given the following information:

- the query
- a generated answer
- a reference answer

Your task is to evaluate the correctness of the generated answer.

Response Format

Your response should be formatted as following: <judge>True or False</judge>

If the generated answer is correct, please set "judge" to True. Otherwise, please set "judge" to False.

Please note that the generated answer may contain additional information beyond the reference answer.

User Prompt:

Query: {Query Description}

Reference Answer: {Reference Answer}

Generated Answer: {Generated Answer}

Figure 12: Prompt of Reward Model.