




Technical Report



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

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen
Shihang Wang, Pengjun Xie, Fei Huang, Feng Zhao[†]

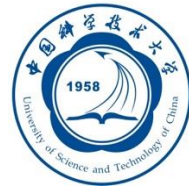
 Tongyi Lab, Alibaba Group

曹耘宁
2025/06/09

智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab



- 研究背景
- 研究方法
- 实验效果
- 总结



研究背景

3

□ 多模态RAG

- ⊙ 固定检索流程，难以端到端优化
- ⊙ 例如：question → query 改写 → 检索 → 过滤 → 拼接到LLM上下文 → ...

□ RL推理

- ⊙ 以往多模态推理任务简单的把图像插入上下文，推理过程只在文本中进行
- ⊙ 视觉任务中的推理如何定义合适的action，实现视觉动态感知？

□ RL+RAG

- ⊙ VLM 与 search engine进行多轮交互，由RL优化这一过程

研究背景

4

- 研究动机：模型根据问题需要，由粗到细主动感知视觉信息

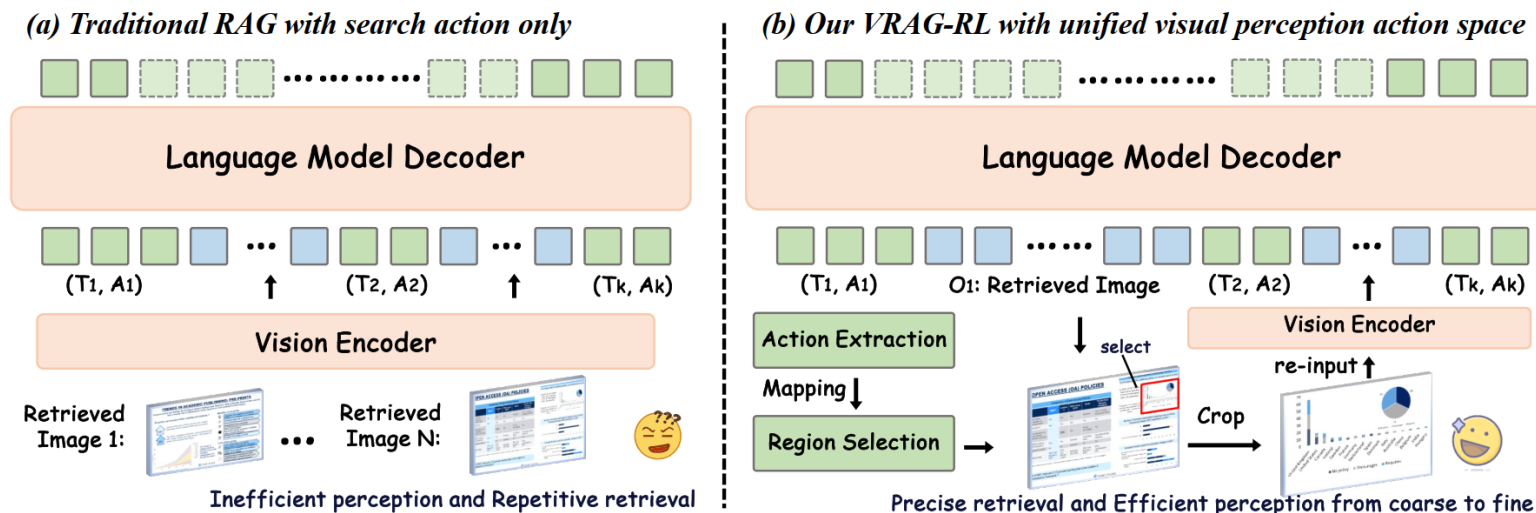


Figure 2: **Comparison between our VRAG-RL and the traditional RAG in terms of perception methods.** (a) Traditional methods lack effective perception, which easily leads to repetitive and ineffective retrieval calls and suboptimal outcomes. (b) Our VRAG-RL is efficient and accurate, enabling the model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



- 研究背景
- 研究方法
- 实验效果
- 总结

总体框架

6

- 多轮强化训练；视觉感知动作空间；反馈

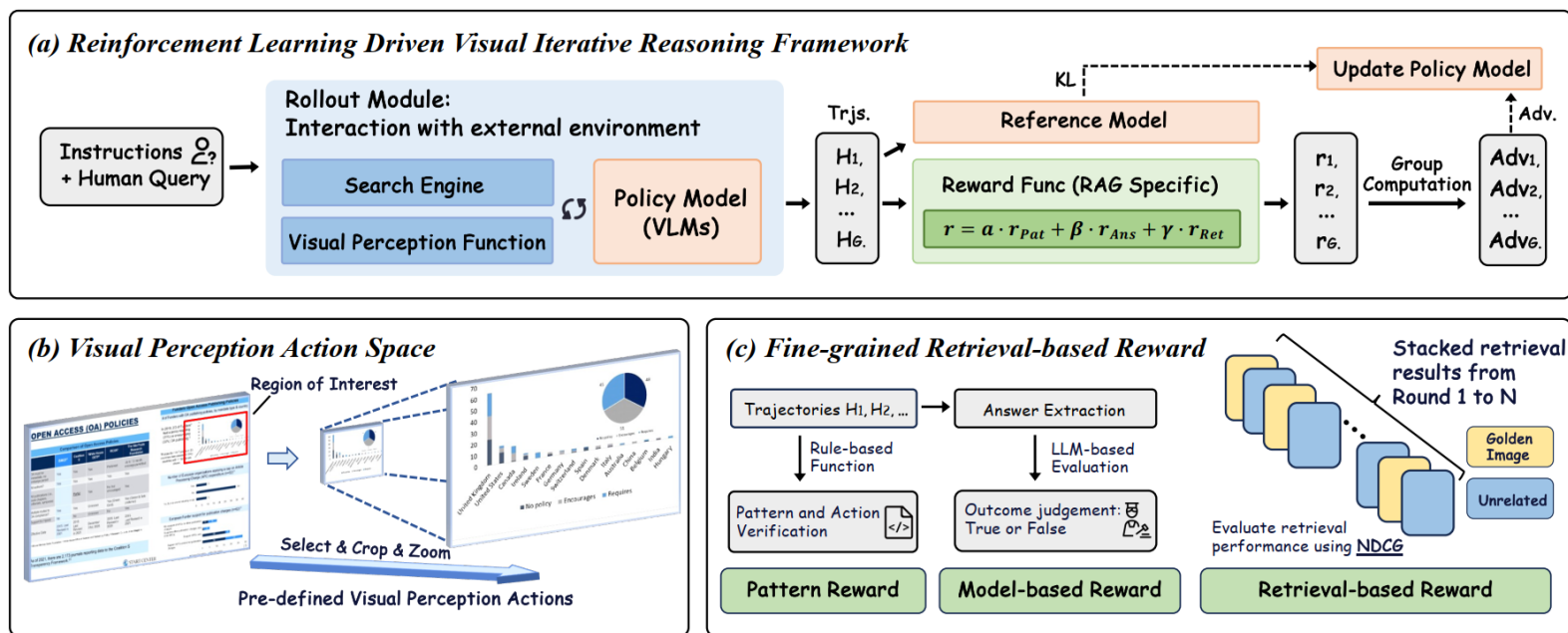


Figure 1: **Overall Framework of our Reinforcement Learning Framework.** (a) demonstrates the interaction process between the model and the external environment, as well as the implementation of the GRPO algorithm. (b) shows the proposed visual perception action space which allows the model to extract 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.



Action-主动感知信息密集区域

7

- Search
- Region/clip/zoom
 - ⊙ 得到裁剪坐标，映射回原图并裁剪，重新输入

Thought-Action-Observation ($\mathcal{T}, \mathcal{A}, \mathcal{O}$)

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

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

- answer



SFT-Trajectory Data Scaling-Up Based on Multi-Expert Sampling

8

□ Sft阶段需要更加丰富的数据

- ⊙ 多专家采样：结合大模型与小模型的优势，生成多样化、高质量的推理轨迹。
- ⊙ 大模型（ π_{LM} ）：擅长全局推理与多模态理解，负责推理路径和工具选择。
- ⊙ 小专家模型（ π_{EM} ）：专注于细节，如视觉区域定位，负责具体坐标标注。
- ⊙ 通过多轮协作，采集出覆盖多种推理方式的丰富数据。

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

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

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

Reward-Retrieval efficiency reward

9

- 检索越早、越准确地召回相关信息，奖励越高
- NDCG reward

1. 定义

- 设一条推理轨迹中，模型共检索到文档序列 $D_{trj} = [d_1, d_2, \dots, d_n]$
- 相关文档集合记为 D_{rel}
- 每个检索到的文档 d_i 有相关性得分 s_i ，如果 $d_i \in D_{rel}$ 则 $s_i = 1$ ，否则 $s_i = 0$

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

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

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

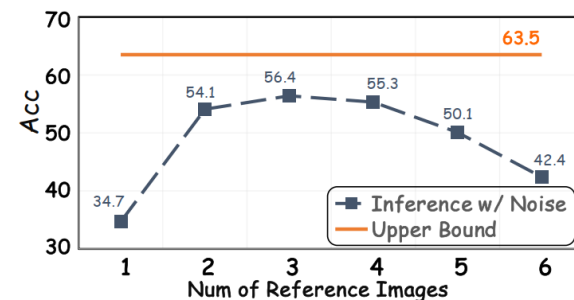


Figure 3: Experiments on the impact of context length on model performance.



Reward-Pattern consistency+model based reward

10

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}



算法流程

11

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
-



- 研究背景
- 研究方法
- 实验效果
- 后续工作
- 总结

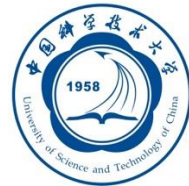


数据集

13

□ 视觉文档数据集

数据集名称	简介	主要内容类型	规模/特点
SlideVQA	针对幻灯片文档视觉问答的数据集，聚焦于理解幻灯片内容	文本、表格、图表、布局等	2,600+套幻灯片， 52,000+图片， 14,500+复杂推理问题
ViDoSeek	针对视觉丰富文档的检索-推理-回答任务而设计	文本、图表、表格、布局等	约6,000张图片，问题需跨文档检索唯一答案
MMLongBench	评测VLM对长上下文、多模态文档理解能力的数据集	文本、图片、图表、表格、布局	强调长上下文和多模态内容，包含多种视觉元素



对比的baseline方法

14

方法名称	类别	简要说明	主要特征/流程
Vanilla RAG	文本/视觉	最基础的RAG方法，分为文本型和视觉型。文本型用文本作为检索语料，视觉型用图片作为检索语料。	直接用原始问题检索，插入上下文后直接回答。
ReAct RAG	文本/视觉	引入“思考-行动-观察”循环的RAG方法，融合Chain-of-Thought (COT) 提示。	模型先推理，再检索，再推理，支持多轮交互。
Search-R1	文本	多轮推理RL方法，基于Search-R1论文实现。	多轮交互+规则奖励，原生为文本RAG，本文复现时采用相同RL框架。
Search-R1-VL	视觉	Search-R1的视觉版，基于本文框架实现。	多轮交互+规则奖励，检索和插入均为图片，冷启动训练。
VRAG-RL (本文方法)	视觉	本文提出的新RL方法，支持视觉感知动作空间和细粒度奖励。	多轮推理，支持区域选择、裁剪、缩放等感知动作，奖励结构更复杂。

实验对比

15

□ 平均有20%~30%提升

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	11.8	9.9	5.3	7.4	6.5	3.7	3.9	5.2	2.5	8.4
🌀 Search-R1	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	15.7	10.9	6.7	14.2	2.7	3.6	3.4	3.1	5.1	10.9
👁 Search-R1-VL	26.3	20.1	20.1	29.8	8.5	7.8	7.9	9.3	7.6	21.3
👁 VRAG-RL	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	21.2	13.3	14.3	21.3	5.9	5.1	7.3	5.5	1.7	15.8
🌀 Search-R1	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	34.8	20.4	27.5	42.1	10.1	12.4	10.2	6.2	7.1	26.9
👁 Search-R1-VL	48.3	42.3	40.5	50.3	19.9	13.4	12.9	11.4	10.2	37.4
👁 VRAG-RL	69.3	43.1	60.6	74.8	26.1	26.3	24.8	25.9	21.2	57.1

消融实验



16

□ 对

Case s

17



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>**



<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>

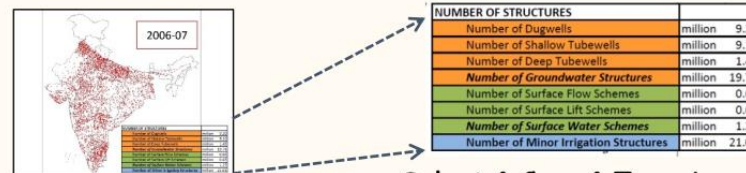
<search> Detailed data on Dugwells and Shallow Tubewells data **</search>**



<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>**



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.

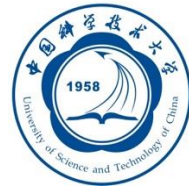
</think>

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





- 研究背景
- 研究方法
- 实验效果
- 总结



总结

19

□ 多模态RL的独特性:

- ⊙ Action的定义: 增强视觉域的交互
- ⊙ Reward设计: 对中间检索过程的奖励提高训练稳定性

□ 工具调用/agent相关工作:

- ⊙ Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning
- ⊙ WebDancer: Towards Autonomous Information Seeking Agency
- ⊙ PANGU DEEPDIVER: ADAPTIVE SEARCH INTENSITY SCALING VIA OPEN-WEB REINFORCEMENT LEARNING
- ⊙ Agentic Reasoning and Tool Integration for LLMs via Reinforcement Learning
- ⊙ RAGEN: Understanding Self-Evolution in LLM Agents via Multi-Turn Reinforcement Learning