

Abstract

Multimodal large language models (MLLMs) have achieved remarkable success across a broad range of vision tasks. However, constrained by the capacity of their internal world knowledge, prior work has proposed augmenting MLLMs by “reasoning-then-tool-call” for visual and textual search engines to obtain substantial gains on tasks requiring extensive factual information. However, these approaches typically define multimodal search in a naive setting, assuming that a single full-level or entity-level image query and few text query suffices to retrieve the key evidence needed to answer the question, which is unrealistic in real-world scenarios with substantial visual noise. Moreover, they are often limited in the reasoning depth and search breadth, making it difficult to solve complex questions that require aggregating evidence from diverse visual and textual sources. Building on this, we propose **Vision-DeepResearch**, which proposes one **new multimodal deep-research paradigm**, *i.e.*, performs multi-turn, multi-entity and multi-scale visual and textual search to robustly hit real-world search engines under heavy noise. Our Vision-DeepResearch supports dozens of reasoning steps and hundreds of engine interactions, while internalizing deep-research capabilities into the MLLM via cold-start supervision and RL training, resulting in a strong end-to-end multimodal deep-research MLLM. It substantially outperforming existing multimodal deep-research MLLMs, and workflows built on strong closed-source foundation model such as GPT-5, Gemini-2.5-pro and Claude-4-Sonnet. The code will be released in <https://github.com/Osilly/Vision-DeepResearch>.

1. Introduction

Multimodal large language models (MLLMs) have achieved substantial success on a wide range of real-world tasks (Liu et al., 2023; Huang et al., 2025; Bai et al., 2025). However, due to their limited internal world knowledge, complex fact-intensive VQA remains a major challenge (Jiang et al.; Wu et al., 2025a; Geng et al., 2025; Narayan et al., 2025). Recent multimodal deep-research MLLM works (Wu et al., 2025a; Geng et al., 2025; Narayan et al., 2025) have proposed equipping MLLMs with the “reasoning-then-tool-call” paradigm (*i.e.*, ReAct (Yao et al., 2022)), where models use external tools as one action after reasoning to query search engines and obtain factual observations. This substantially improves MLLMs’ performance on VQA problems that

require extensive real-world knowledge.

However, existing works face two key issues. First, *they formulate multimodal search under an overly simple setting*. Image queries are treated as full-image (or entirety-level) retrieval, and a small number of text queries is assumed to suffice. This overlooks a critical challenge in realistic, noisy search engines: the **hit and run problem**. This issue is evident in practice. As illustrated in Fig. 1 (A.1), full-image retrieval can be dominated by irrelevant visual noise. In real user scenarios, it is unrealistic to assume that the entire image content is aligned with the target information. Furthermore, search engines exhibit substantial hit-rate variability. Even when querying the same visual or textual entity, retrieval results can differ markedly across query scales, and obtaining the required evidence from real-world engines is often non-trivial.

Second, *existing methods are constrained in both reasoning depth and search breadth, making it difficult to perform complex multi-hop deep-research and to aggregate evidence from multiple information sources*. Rather than treating retrieval as a one-off operation, it should be modeled as a trial-and-error process that adaptively explores multi-scale search and iteratively refines queries based on intermediate results, enabling it to better handle noisy, unstable search environments and ambiguous inputs. As shown in Fig. 1 (A.2), our proposed Vision-DeepResearch substantially increases both the number of reasoning steps and the number of engine interactions compared to prior multimodal deep-research MLLMs and agentic workflows. This enables one 30B-A3B-scale and even a 8B-scale model to achieve State-of-The-Art (SoTA) performance across multiple multimodal factual benchmarks (see Fig 1 (bottom)).

To address the mentioned issues, *we design one new multimodal deep-research paradigm*. As presented in Fig. 1 (B), we design the highly automated factual VQA synthesis and multi-turn trajectory generation pipelines to obtain high-quality training data, and then integrate multi-turn, multi-entity, and multi-scale visual and textual search capabilities into a base MLLM through cold-start supervision and reinforcement learning (RL) training, yielding a strong multimodal deep-research MLLM, **Vision-DeepResearch**. We adopt an entity-level, stringent image verification and filtering pipeline for candidate question synthesis. To expand the multi-hop structure of the textual component, we further perform random walks over real search engines and real web pages, together with joint entity and answer obfuscation, to obtain high-quality factual VQA data. Moreover, we induce existing MLLMs to generate multi-entity, multi-scale visual region proposals to efficiently probe visual search engines and collect visual retrieval trajectories. We also introduce an obfuscated termination strategy to control the depth of visual retrieval. We then bridge modalities by leveraging a

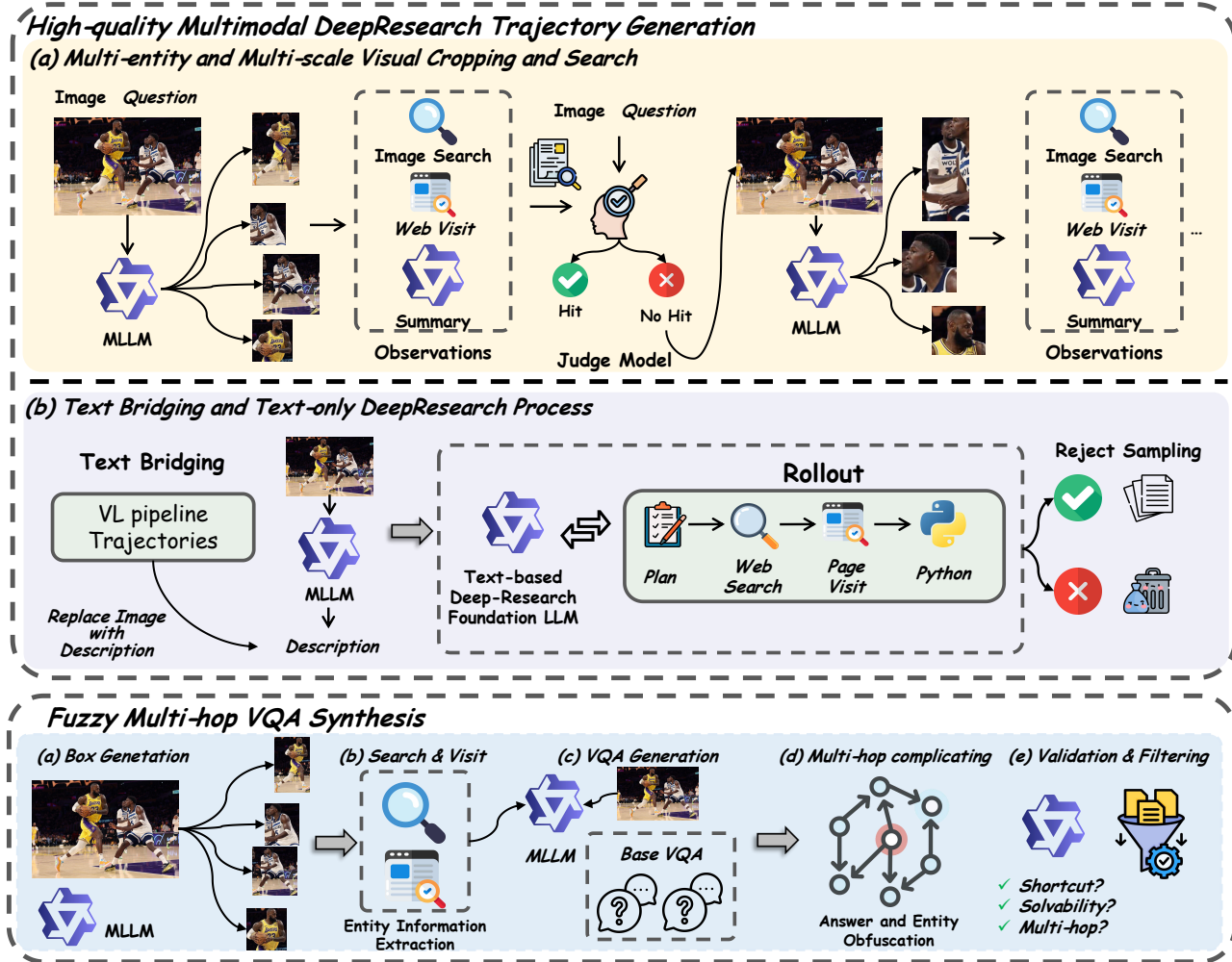


Figure 2. Our Data Pipeline. As shown in the top panel, we construct a complete multimodal deep-research synthesis pipeline. Leveraging the capabilities of an MLLM and a text-based DeepResearch foundation LLM, we generate long-horizon, multi-tool trajectories. As shown in the bottom panel, we obtain high-quality factual VQA instances via a rigorous verification and obfuscation procedure, which are then used for trajectory synthesis and RL training.

strong deep-research LLM foundation model to produce the corresponding text-search trajectories, ultimately yielding complete multimodal deep-research cold-start trajectories.

Extensive experiments show that our proposed Vision-DeepResearch significantly outperforms all prior multimodal deep-research MLLMs on six factual benchmarks, and even surpasses agent workflows built on strong closed-source models such as GPT-5, Gemini-2.5-Pro, and Claude-4-Sonnet. *We believe our work can inspire the research community deeply.*

2. Method

In this section, we present the pipeline for constructing our vision deep research agent, which is capable of performing long-horizon vision-language reasoning in realistic and noisy web environments. We first discuss the design motivation (Sec. 2.1), then describe the data generation process

in detail (Sec. 2.2), and finally outline the training strategies (Sec. 2.3).

2.1. Motivation

Realistic multimodal deep-research systems require gather information under noisy web conditions. However, as shown in Sec. 1, existing works perform search tools to short-horizon or single-shot operations, leading to brittle retrieval behavior and premature convergence in complex multimodal reasoning tasks. We highlight two key issues below.

Coarse search strategy. In real-world web environments, search environment is inherently noisy and unstable. Images often contain multiple visual entities with cluttered backgrounds and large scale variations, while search engines may return inconsistent results even for visually similar queries, making reliable retrieval particularly challenging. Humans naturally cope with such conditions through an iterative

and exploratory search process: when initial attempts fail, they progressively refine their queries by cropping different regions, adjusting scales, or focusing on alternative visual cues until sufficient evidence is accumulated to identify the target entity. The same holds for textual search: reaching the desired target via a search engine often requires multiple attempts. Even minor word-level modifications to a query can lead to substantially different retrieved content.

A robust multimodal deep-research system can obtain substantial benefits by mirroring this human search behavior. Instead of treating retrieval as a one-shot operation, it should model retrieval as a **trial-and-justice** process, adaptively exploring multi-scale regions and refining queries based on intermediate results to better handle noisy, unstable search environments and ambiguous inputs. Unfortunately, in visual search side, this problem is particularly even worse. Most prior work adopts a single-pass, full-image or entity retrieval paradigm, where the model issues a single visual query to the search engine using the original image. This approach is highly fragile in real-world web environments containing multiple visual entities.

Lack of optimization for long-horizon engine interaction.

Most existing training dataset synthesis in multimodal deep-research MLLMs (Jiang et al.; Wu et al., 2025a; Geng et al., 2025; Narayan et al., 2025) contain search trajectories that are limited to short contexts or an average of fewer than five retrieval rounds. As a result, models trained on such data tend to prematurely terminate exploration in complex tasks, settling for partially relevant evidence rather than systematically exploring alternative visual or textual queries. However, there currently exists no systematic framework for synthesizing trajectories that involve dozens of reasoning steps and both visual and textual search engine interactions, which are essential for authentic deep research scenarios.

In contrast, deep-research LLMs (Team et al., 2025) specialized for text-based deep research have demonstrated strong ReAct-style capabilities, enabling them to iteratively plan, search, and reflect over long contexts, often involving tens of tool invocations. Nevertheless, current MLLMs struggle to transfer such long-horizon reasoning behavior from text-only tasks to multimodal settings.

To address these challenges, we design a new data and training paradigm consisting of the following components:

- **Multi-entity and multi-scale visual cropping and search.** We enable robust trial-and-error visual retrieval in noisy environments by performing visual search over multiple scales and regions, thereby increasing the hit rate of target visual entities.
- **Long-horizon trajectory construction.** We leverage the strong localization capabilities of MLLMs together with

the deep retrieval behaviors of text-based deep research models, and seamlessly transfer their long-horizon ReAct-style reasoning to the visual domain via image-description-based context window sharing.

2.2. Data Pipeline

As shown in Fig 2, our data pipeline constructs *long-horizon multimodal deep-research trajectories* by combining visual search with text-based deep-research reasoning, bridged through image descriptions.

2.2.1. HIGH-QUALITY MULTIMODAL DEEPRESEARCH TRAJECTORY GENERATION

Multi-entity and Multi-scale Visual Cropping and Search.

Given an input image I , the question q and the prompt p_v to induce one MLLM to generate single-turn ReAct-style context for visual toll call, we first generate one reasoning R and then localize regions relevant to the query and generate multiple bounding boxes $S_b = \{I_b^1, \dots, I_b^n\}$, including fine-grained entity-level regions I_b . Each cropped regions set in t -th step defines a visual action $A^t = \text{Tool-Call}(S_b^t)$, which is submitted to the visual search tool pipeline, where $t_v \in \{1, \dots, T_v\}$ and T_v is the final step of the visual tool pipeline.

We denote the observation returned by the visual tool pipeline Vision-Pipeline at step t as:

$$\mathcal{O}^{t_v} = \text{Vision-Pipeline}(A^{t_v}), \quad (1)$$

and the cumulative visual evidence up to step t_v as:

$$\mathcal{V}^{t_v} = \{\mathcal{O}^1, \dots, \mathcal{O}^{t_v}\}, \quad (2)$$

where the tool list include three sequential execution tools, *i.e.*, the visual search tool, website visit tool and website summary tool. The visual search tool takes a cropped image region as input and returns the matched webpage URL, we then use a website visit tool to fetch the page content in markdown format. The returned content is typically long and cluttered with image links, and passing it directly to the MLLM can easily exceed the context window. To mitigate this, we use an auxiliary MLLM to summarize the webpage and verify the correspondence between the cropped image query and the matched images on the page, so as to extract the most relevant evidence while filtering out irrelevant content.

When we finish the single-turn ReACT-style context, we use the induction prompt p_v again to generate the next-turn context. Finally, the visual tool trajectory is represented as:

$$\mathcal{C}_{\text{vision}} = \{I, q, p_v, R^1, A^1, \mathcal{O}^1, \dots, p_v, R^{T_v}, A^{T_v}, \mathcal{O}^{T_v}\}, \quad (3)$$

To control visual search depth, an external judge model evaluates whether the accumulated evidence \mathcal{V}_t is sufficient

to support downstream text-based reasoning steps. Conditioned on the original image I , the question q , the ground truth a_{true} and the collected evidence, the judge outputs a binary *hit signal*.

$$h^{t_v} = \text{Judge}(I, q, \mathcal{V}^{t_v}, a_{\text{true}}) \in \{0, 1\}, \quad (4)$$

In the Judge stage, we ask the model to make a relatively lenient determination of whether the currently available visual evidence contains sufficient information to answer the question. If $h^{t_v} = 0$, the pipeline continues with additional visual search actions A^{t_v+1} , while if $h^{t_v} = 1$, the final step of the visual tool pipeline T_v is set to current t_v and the visual tool process terminates. With this strategy, we aim to retrieve sufficiently informative evidence from the visual search engine to support the subsequent text-based reasoning process.

Text Bridging and Text-only DeepResearch Process. To leverage the strong ReAct-style capability of the text-based deep-research foundation LLM, we bridge the above visual trajectory $\mathcal{C}_{\text{vision}}$ to the text-only context. We first generate a detailed textual description D for the input image I , while replacing I with D and removing the induction prompts p which in $\mathcal{C}_{\text{vision}}$, keep the rest of the trajectory unchanged as the bridged context, *i.e.*, reasoning R^{t_v} , actions A^{t_v} and observations \mathcal{O}^{t_v} remain the same. We then sent the bridged context and the corresponding prompt p_t to induce the text-based deep-research foundation LLM yields the subsequent text-based trajectory. The textual tool trajectory of text-based deep-research foundation LLM can denote as:

$$\begin{aligned} \mathcal{C}_{\text{text}} = \\ \{D, q, R^1, A^1, \mathcal{O}^1, \dots, R^{T_v}, A^{T_v}, \mathcal{O}^{T_v}, \\ p_t, R^{T_v+1}, A^{T_v+1}, \mathcal{O}^{T_v+1}, \dots, R^{T_v+T_t}, A^{T_v+T_t}, a_{\text{output}}\}, \end{aligned} \quad (5)$$

where T_t is the number of steps generated by the foundation LLM and a_{output} is the model’s final answer output. The textual tool list involving web search, website visit&summary and python code.

Finally, we merge the trajectories $\mathcal{C}_{\text{vision}}$ and $\mathcal{C}_{\text{text}}$ to obtain the full multimodal deep-research trajectory:

$$\begin{aligned} \mathcal{C}_{\text{multimodal}} = \\ \{I, q, R^1, A^1, \mathcal{O}^1, \dots, R^{T_v}, A^{T_v}, \mathcal{O}^{T_v}, \\ R^{T_v+1}, A^{T_v+1}, \mathcal{O}^{T_v+1}, \dots, R^{T_v+T_t}, A^{T_v+T_t}, a_{\text{output}}\}, \end{aligned} \quad (6)$$

We then apply rejection sampling to select trajectories $\mathcal{C}_{\text{multimodal}}$ for cold-start training, *i.e.*, an LLM verifies whether the final trajectory output a_{output} matches the ground-truth answer a_{true} . Consistent trajectories are retained in the training set, while inconsistent ones are discarded. Furthermore, we also incorporate text-only deep-research trajectories generated by the original question q into the foundation LLM.

2.2.2. VERIFIED FACTUAL VQA GENERATION

Factual VQA Verification. First, we curate images from multiple open-source datasets, focusing on real-world, high-quality, complex images with multiple entities. We filter out images smaller than 224×224 , and then use an MLLM as a selector, guided by predefined criteria, to identify high-quality real-world images and remove overly trivial cases. We then further filter the candidates along two dimensions. We directly feed the VQA instance to an MLLM. If it can answer correctly without external evidence, we discard the sample. Moreover, we submit the full image to an image search engine. If the retrieved results perfectly match the full-image query, we also discard the sample.

Applying these criteria yields a higher-quality factual VQA dataset. We use a subset of the resulting samples for trajectory synthesis (Sec. 2.2.1) and RL training (Sec. 2.3.2), while using the remaining images (discarding their original QAs) for Fuzzy Multi-hop VQA Synthesis.

Fuzzy Multi-hop VQA Synthesis. For each retained image, we first prompt an MLLM to propose a set of entity-level candidate bounding boxes. We then crop these regions at multiple scales and perform image search. We match the retrieved images against the corresponding crops. If they refer to the same entity, we retain the box and its associated entity. We then use an MLLM to generate a simple, unambiguous entity-level question, *e.g.*, ‘What is the name of the cat in the image?’.

The entity-level questions obtained above are often overly explicit and simplistic, deviating from real user queries. We therefore further obfuscate both the entity and the question in two ways.

1. *Answer obfuscation*, which increases the required reasoning depth by chaining relations around the answer (*e.g.*, ‘‘What is the name of the teacher of the cat owner’s daughter?’’).
2. *Entity obfuscation*, a technique commonly used in text-only deep-research LLMs (Wu et al., 2025c;b; Li et al., 2025; Tao et al., 2025b), where we perform random walks over webpages to replace the original entity with related entities along multi-hop links (*e.g.*, ‘‘The cat’s owner works at A, and the owner’s daughter studies at B. So what is the cat’s name?’’).

However, repeatedly increasing complexity purely via answer chaining can lead to rigid, templated reasoning patterns. In contrast, entity obfuscation alone can introduce cross-source shortcuts, where the final answer can be inferred via consistency checks over multiple textual entities, without requiring any visual evidence. We therefore adopt an interleaved obfuscation strategy that alternates between answer and entity obfuscation.

Table 1. Benchmark results across different Settings with improvement (Δ , compared with base MLLM in agentic workflow setting). The best results are highlighted in **bold**, and the second-best results are underlined.

Model	VDR	FVQA	MMSearch+	MMSearch	LiveVQA	BC-VL	Avg.
Direct Answer							
GPT-5	9.8	57.3	19.1	33.3	57.5	47.2	37.4
Gemini-2.5 Pro	8.0	60.7	14.5	39.8	60.3	43.1	37.7
Gemini-2.5 Flash	6.2	47.7	8.1	30.4	51.0	37.1	30.1
Claude-4-Sonnet	2.0	35.3	4.0	18.7	38.5	29.3	21.3
Claude-3.7-Sonnet	4.6	36.7	4.0	21.1	38.0	32.3	22.8
Qwen3-VL-8B-Instruct	2.8	28.0	3.2	15.2	41.0	25.1	19.2
Qwen3-VL-8B-Thinking	5.6	24.0	2.7	15.8	43.3	25.1	19.4
Qwen3-VL-30B-A3B-Instruct	3.8	34.7	3.2	18.7	42.7	29.6	22.1
Qwen3-VL-30B-A3B-Thinking	4.4	32.7	4.5	19.3	49.0	34.6	24.1
RAG Workflow							
Gemini-2.5-flash	–	–	–	43.9	41.3	12.1	–
Claude-3.7-Sonnet	–	–	–	32.7	30.3	10.0	–
Qwen-2.5-VL-72B	–	–	–	29.2	35.7	10.2	–
Agent Workflow							
GPT-5	20.4	<u>69.0</u>	17.2	63.7	73.3	46.1	48.3
Gemini-2.5 Pro	18.8	68.3	22.2	<u>69.0</u>	76.0	49.9	<u>50.7</u>
Gemini-2.5 Flash	16.3	68.0	19.9	64.0	73.0	44.6	47.6
Claude-4-Sonnet	13.6	<u>69.0</u>	<u>23.1</u>	67.2	69.7	48.6	48.5
Claude-3.7-Sonnet	27.2	67.3	17.2	63.7	72.0	<u>50.4</u>	49.6
Qwen3-VL-8B-Thinking	17.6	51.3	12.2	45.6	56.3	37.1	36.7
Qwen3-VL-30B-A3B-Thinking	23.2	63.0	13.6	53.2	62.0	44.1	43.2
Multimodal DeepResearch MLLM							
MMSearch-R1-7B	–	58.4	–	53.8	48.4	–	–
Webwatcher-7B	–	–	–	49.1	51.2	20.3	–
Webwatcher-32B	–	–	–	55.3	58.7	26.7	–
Ours							
Qwen3-VL-8B-Instruct (Agentic)	17.0	58.7	11.3	52.0	63.0	38.6	40.1
Vision-DeepResearch-8B (Ours)	<u>29.2</u>	64.7	20.4	69.6	<u>76.7</u>	42.6	<u>50.5</u>
Δ	+12.2	+6.0	+9.1	+17.6	+13.7	+4.0	+10.4
Qwen3-VL-30B-A3B-Instruct (Agentic)	20.2	57.7	10.0	55.0	60.0	42.6	40.9
Vision-DeepResearch-30B-A3B (Ours)	37.8	74.2	28.5	69.6	77.6	53.7	56.9
Δ	+17.6	+16.5	+18.5	+14.6	+17.6	+11.1	+16.0

Moreover, in real-world question design, humans typically (1) decide the assessment target, (2) search to verify answerability, (3) draft multiple candidate questions and select the best one, and (4) attempt to solve it themselves. We emulate this process with an automated pipeline: an MLLM extracts retrieval keywords from the current entity, queries external sources to collect additional information, another MLLM proposes multiple candidate questions conditioned on different retrieved evidence, and a judge MLLM selects the most reasonable and objective question. We iterate this pipeline during interleaved answer/entity obfuscation, ultimately producing a complex question paired with its answer. Finally, we merge the original image and the complex question-answer pair to obtain the fuzzy multi-hop VQA problems. The resulting fuzzy multi-hop VQA instances are treated as higher-quality samples, and are used both for multimodal deep-research trajectory synthesis and for RL training.

2.3. Training

We use the generated multimodal deep-research trajectories and verified VQA to train our Vision-DeepResearch model using a combination of supervised fine-tuning (SFT) and

reinforcement learning (RL).

2.3.1. SUPERVISED FINE-TUNING

We first perform SFT to teach the model the fundamental behaviors required of a multimodal deep-research MLLM. Following the pipeline described in Sec. 2.2, we collect 30K high-quality trajectories. Each trajectory $\mathcal{C}_{\text{multimodal}}$ consists of: a question, an initial image and the multi-turn deep-research steps (including both multimodal and text-only trajectories). For data mixing, we sample verified fact-centric VQA problems from existing VQA datasets and augment them with multimodal deep-research trajectories, yielding 16K instances. We also sample text-only QA problems and generate corresponding text-only trajectories, resulting in 8K instances. Finally, we sample 6K fuzzy VQA instances and similarly populate them with trajectories. The VQA filter process and fuzzy VQA synthesis described in Sec. 2.2.2.

The model is trained by minimizing the standard autoregressive cross-entropy loss (CE loss). The goal of SFT is to guide the model to learn multi-turn, multi-entity and multi-scale patterns, integrate visual and textual evidence during reasoning, and develop long-horizon planning behaviors

Table 2. Ablation study on rollout pipeline.

Setting	VDR	MMS+	BC-VL	Avg.
Direct Answer	4.8	3.6	27.6	12.0
WIS	11.8	10.0	26.1	16.0
WIS+TS	16.0	23.5	48.4	29.3
CIS	15.4	22.7	30.8	23.0
CIS+TS	37.8	28.5	53.7	40.0

Table 3. Ablation results on training data and methods.

Model	VDR	MMS+	BC-VL	Avg.
Qwen3-VL-30B-Instruct	20.2	10.0	42.6	24.3
+16K VQA traj. (SFT)	24.4	23.5	50.9	32.9
+8K QA traj. (SFT)	27.0	23.5	50.1	33.5
+6K fuzzy VQA traj. (SFT)	33.2	26.0	51.4	36.9
+RL training	37.8	28.5	53.7	40.0

for visual tasks that are analogous to those exhibited by text-based deep research models.

2.3.2. REINFORCEMENT LEARNING

High-throughput Asynchronous Rollout Architecture Design. Multi-turn agentic-reasoning RL is typically bottlenecked by rollouts, *i.e.*, long reasoning horizons and frequent tool calls introduce substantial latency, and naive synchronous rollouts can severely stall the event loop. To this end, we design a high-throughput multi-threaded asynchronous rollout pipeline building on rLLM framework (Tan et al., 2025). It dispatches tasks via a queued scheduler and maintains a tool pool to support concurrent multi-tool calls within a single action (*e.g.*, querying search results for multiple image crops in parallel), returning observations asynchronously. By offloading potentially blocking operations for event loop, our asynchronous design achieves over $10\times$ higher rollout throughput than synchronous rollouts, to achieve the efficient RL training of our Vision-DeepResearch.

Training Recipe. To further refine the agent’s behavior, we apply RL training with Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Guo et al., 2025) with Leave-One-Out trick (Ahmadian et al., 2024; Luo; Chen et al., 2025) on the post-SFT Vision-DeepResearch model. We perform reinforcement learning using 15K high-quality VQA instances, where 10K filtered from existing VQA dataset and 5K obtained from fuzzy VQA synthesis pipeline. During training, the model interacts with a real online search environment (including visual search, text search, and website visiting) and samples long-horizon rollout trajectories. We cap the maximum horizon, context length and single-turn response length at 50 turns, 64K and 4K tokens, respectively.

For reward design, we adopt LLMs-as-Judge paradigm, where a judge model determines whether the final answer matches the reference answer and provides the corresponding reward signal. During training, we use a pure accuracy reward, *i.e.*, the model receives a reward of 1.0 if the gener-

ated answer is correct, and 0.0 otherwise.

Engineering Trick. During RL training, we explore many tricks to keep stable.

1. Trajectory Rollout Interrupted. During RL training, we encounter a severe long-tail issue, where a small fraction of trajectories dominates the wall-clock time of an entire rollout batch. This long tail typically arises in two cases. First, repetitive text: the model may enter degenerate loops, repeatedly generating near-duplicate responses until hitting the context limit. We mitigate this via an n-gram-based repetition detector with a minimum character-length threshold; once repetition is detected and the response exceeds the threshold, we immediately terminate the trajectory. Second, cascading format/tool-call failures: the model may repeatedly produce invalid formats or incorrect tool invocations, persisting until the maximum turn or length budget is exhausted. For this case, we track consecutive errors and terminate the trajectory once the error count reaches three. For other cases (*e.g.*, isolated formatting mistakes or occasional tool-call errors), we return an explicit observation (*e.g.*, “format error, please try again”) and allow the model to continue the rollout.

2. Mask Trajectory. We observe a severe negative-gradient issue during training. Concretely, for anomalous trajectories, for example, those terminated by the safeguards above, those exceeding the turn/length budget, or those dominated by format/tool-call errors (*e.g.*, accounting for more than half of the steps)—naively assigning a 0.0 reward is overly punitive. This can suppress otherwise correct steps within these trajectories, injecting substantial negative signal and ultimately destabilizing training. To address this, we mask such trajectories from gradient updates (*i.e.*, exclude them from backpropagation), while still including them in advance computation.

3. BF16 vs. FP16. Recent study (Qi et al., 2025) suggests that RL training in FP16 can be more effective and stable than BF16. We also explored FP16. However, our rollouts are long (up to a 64K context), which led to numerical overflow and training instability under FP16. We therefore train in BF16.

In this section, we do not introduce algorithmic innovations, instead, these engineering techniques are crucial for ensuring stable large-scale agentic-reasoning RL training.

3. Experiments

In this section, we present a comprehensive experimental study. We first compare our approach with existing methods across a range of multimodal retrieval and reasoning benchmarks (Sec. 3.1). Then, we conduct ablation studies to analyze the contributions of key components, including both

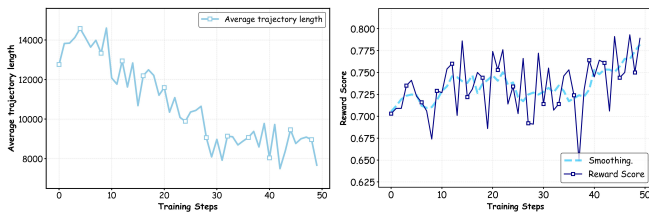


Figure 3. RL Curves of Mean Trajectory Length and Reward.

pipeline modules and data choices (Sec. 3.2 and Sec. 3.3). Finally, we analysis the RL training in Sec. 3.4. The experiment settings are presented in Appendix B.

3.1. Main Results

Tab. 1 benchmarks proprietary and open MLLMs under three paradigms. Tool-free direct answering is consistently weak on open-domain multimodal deep-research tasks (*e.g.*, Qwen3-VL-30B-A3B-Thinking: 24.1% Avg.). ReAct-style agent workflows yield large improvements for most models (*e.g.*, Gemini-2.5 Pro: 50.7% Avg.), while naive RAG workflows provide limited gains on the reported settings.

Our Vision-DeepResearch models achieve the best performance among open models and are competitive with strong proprietary agentic systems. Under the same agentic backbone, Vision-DeepResearch-8B improves over Qwen3-VL-8B-Instruct (Agentic) by +10.4% Avg., with notable gains on MMSearch (+17.6%) and LiveVQA (+13.7%). Scaling to Vision-DeepResearch-30B-A3B further boosts results to 56.9 Avg. (+16.0), with consistent improvements on VDR (+17.6%), FVQA (+16.5%), and MMSearch-Plus (+18.5%), indicating that our data and training better instill long-horizon “reason-then-tool-call” behavior beyond generic agent prompting.

3.2. Pipeline Ablation

We conduct a systematic ablation on 30B-A3B model to isolate the effects of multi-scale visual cropping and retrieval strategies (Tab. 2). We compare: Direct Answer (no retrieval), WIS (whole-image visual search), WIS+TS (whole-image visual + text search), CIS (multi-scale cropped-image visual search), and CIS+TS (multi-scale cropping with both visual and text search).

As shown in Tab. 2, removing retrieval (Direct Answer) yields very poor performance (Avg. 12.0%), confirming that external evidence is essential for open-domain multimodal reasoning. Whole-image visual search provides limited gains (Avg. 16.0%) and even degrades BC-VL (27.6%→26.1%), suggesting that single-shot retrieval is often distracted by background clutter and fails to surface long-tail knowledge. Adding text search on top of whole-image retrieval (WIS+TS) substantially improves overall accuracy (Avg. 29.3%), with a large boost on BC-VL

(48.4%), highlighting the complementarity between visual grounding and textual evidence. Introducing multi-scale cropping for visual retrieval (CIS) dramatically improves VDR (4.8%→15.4%), validating the importance of localized object-centric anchors. However, without text search it remains suboptimal on knowledge-heavy benchmarks (MMS+ 22.7%; BC-VL 30.8%). The full pipeline (CIS+TS) achieves the best and most balanced results across all benchmarks (VDR 37.8%, MMS+ 28.5%, BC-VL 53.7%; Avg. 40.0%), indicating that multi-scale visual retrieval and text search are jointly necessary: cropping provides precise visual anchors, while text search supplies the missing long-tail factual evidence.

3.3. Data Ablation

Tab. 3 shows that the base Qwen3-VL-30B-Instruct is insufficient for deep-research VQA (Avg. 2 4.3%; MMS+ 10.0%), indicating missing long-horizon tool-use and evidence grounding. SFT with tool-augmented trajectories brings major gains: adding verified VQA trajectories boosts MMS+ to 23.5% and BC-VL to 50.9%, while text-only QA trajectories achieve similar improvements (MMS+ 23.5%; BC-VL 50.1%), validating effective transfer via our vision-to-text bridging pipeline. Adding fuzzy multi-hop VQA trajectories further improves MMS+ (26.0%) and BC-VL (51.4%), suggesting better coverage of long-tail, multi-hop settings. RL on top of SFT yields the best overall results (VDR 37.8%, MMS+ 28.5%, BC-VL 53.7%), showing that online interaction is crucial for refining long-horizon decision making beyond offline supervision.

3.4. RL Training

In Fig. 3, we visualize the RL training curves of average trajectory length and reward. The model initially tends to produce longer trajectories; as training proceeds, it learns to use the available tools more **ineffectively**, exhibiting shorter trajectories while achieving higher rewards. Consistent with the last two rows of Tab. 3 (post-SFT vs. post-RL), RL improves performance by an average of 3.1% on three challenging benchmarks, demonstrating the effectiveness of the large-scale RL training described in Sec. 2.3.2.

Moreover, due to API cost and wall-clock constraints, we do not exhaustively scale RL training. We expect the model to further benefit from larger-scale RL optimization.

4. Conclusion

We propose Vision-DeepResearch, which scales multimodal deep-research trajectories to dozens of reasoning steps and hundreds of search-engine interactions, yielding substantially stronger performance. We believe our work provides useful insights for the community.

References

- Ahmadian, A., Cremer, C., Gallé, M., Fadaee, M., Kreutzer, J., Pietquin, O., Üstün, A., and Hooker, S. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.
- Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-VL Technical Report. *arXiv e-prints*, art. arXiv:2511.21631, November 2025. doi: 10.48550/arXiv.2511.21631.
- Chen, K., Cusumano-Towner, M., Huval, B., Petrenko, A., Hamburger, J., Koltun, V., and Krähenbühl, P. Reinforcement learning for long-horizon interactive llm agents. *arXiv preprint arXiv:2502.01600*, 2025.
- Comanici, G., Bieber, E., Schaeckermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- Fu, M., Peng, Y., Liu, B., Wan, Y., and Chen, D. Livevqa: Live visual knowledge seeking. *arXiv preprint arXiv:2504.05288*, 2025.
- Geng, X., Xia, P., Zhang, Z., Wang, X., Wang, Q., Ding, R., Wang, C., Wu, J., Zhao, Y., Li, K., et al. Webwatcher: Breaking new frontier of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*, 2025.
- Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.
- Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., and Lin, S. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.
- Jiang, D., Zhang, R., Guo, Z., Wu, Y., Qiu, P., Lu, P., Chen, Z., Song, G., Gao, P., Liu, Y., et al. Mmsearch: Unveiling the potential of large models as multi-modal search engines. In *The Thirteenth International Conference on Learning Representations*.
- Li, K., Zhang, Z., Yin, H., Zhang, L., Ou, L., Wu, J., Yin, W., Li, B., Tao, Z., Wang, X., et al. Websailor: Navigating super-human reasoning for web agent. *arXiv preprint arXiv:2507.02592*, 2025.
- Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- Luo, M. Deepswe: Training a fully open-sourced, state-of-the-art coding agent by scaling rl, jul 2025. *URL https://www.together.ai/blog/deepswe*.
- Narayan, K., Xu, Y., Cao, T., Nerella, K., Patel, V. M., Shiee, N., Grasch, P., Jia, C., Yang, Y., and Gan, Z. Deepmmsearch-r1: Empowering multimodal llms in multimodal web search. *arXiv preprint arXiv:2510.12801*, 2025.
- Qi, P., Liu, Z., Zhou, X., Pang, T., Du, C., Lee, W. S., and Lin, M. Defeating the training-inference mismatch via fp16. *arXiv preprint arXiv:2510.26788*, 2025.
- Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.
- Tan, S., Luo, M., Cai, C., Venkat, T., Montgomery, K., Hao, A., Wu, T., Balyan, A., Roongta, M., Wang, C., Li, L. E., Popa, R. A., and Stoica, I. rllm: A framework for post-training language agents, 2025. Notion Blog.
- Tao, X., Teng, Y., Su, X., Fu, X., Wu, J., Tao, C., Liu, Z., Bai, H., Liu, R., and Kong, L. Mmsearch-plus: Benchmarking provenance-aware search for multimodal browsing agents. *arXiv preprint arXiv:2508.21475*, 2025a.
- Tao, Z., Wu, J., Yin, W., Zhang, J., Li, B., Shen, H., Li, K., Zhang, L., Wang, X., Jiang, Y., et al. Webshaper: Agentic data synthesizing via information-seeking formalization. *arXiv preprint arXiv:2507.15061*, 2025b.
- Team, T. D., Li, B., Zhang, B., Zhang, D., Huang, F., Li, G., Chen, G., Yin, H., Wu, J., Zhou, J., et al. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*, 2025.
- Wang, P., Wu, Q., Shen, C., Dick, A., and Van Den Hengel, A. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017.

Wu, J., Deng, Z., Li, W., Liu, Y., You, B., Li, B., Ma, Z., and Liu, Z. Mmsearch-r1: Incentivizing llms to search. *arXiv preprint arXiv:2506.20670*, 2025a.

Wu, J., Li, B., Fang, R., Yin, W., Zhang, L., Tao, Z., Zhang, D., Xi, Z., Fu, G., Jiang, Y., et al. Webdancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*, 2025b.

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*, 2025c.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., and Cao, Y. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

Zeng, Y., Huang, W., Fang, Z., Chen, S., Shen, Y., Cai, Y., Wang, X., Yin, Z., Chen, L., Chen, Z., Huang, S., Zhao, Y., Hu, Y., Torr, P., Ouyang, W., and Cao, S. Vision-deepresearch benchmark: Rethinking visual and textual search for multimodal large language models. *preprint*, 2026.

A. Related Work

A.1. Text-only DeepResearch LLMs

Early deep-research LLMs primarily focused on text-only environments. A series of works, including Tongyi-DeepResearch (Team et al., 2025), WebDancer (Wu et al., 2025b), WebSailor (Li et al., 2025), and WebShaper (Tao et al., 2025b), formulate complex information retrieval as an iterative loop of reasoning–tool call–re-reasoning, in which agents autonomously generate search queries, browse web pages, and iteratively integrate evidence to produce long chains of reasoning. This paradigm has led to substantial performance gains in open-domain question answering and knowledge-intensive reasoning tasks, demonstrating that “reasoning-then-tool-call” is an effective pathway toward more capable general intelligence.

However, these models rely almost exclusively on textual retrieval and access, where critical information is recalled through keyword-based text search. As a result, they lack essential capabilities such as visual perception, entity localization, and cross-modal consistency verification. Given that real-world information is inherently multimodal, text-based deep research agents struggle to support many high-value applications, including complex web page understanding, GUI comprehension, and product or visual entity recognition. Consequently, multimodal deep research agents capable of multi-round search, understanding, and decision-making in noisy visual environments are widely regarded as a key research direction toward more powerful artificial general intelligence (AGI).

A.2. Multimodal DeepResearch MLLMs

Recent studies have begun to explore deep research capabilities in visual environments. WebWatcher (Geng et al., 2025) transforms text-based question answering into visual question answering via reverse image search, and constructs supervised fine-tuning datasets that enable agents to retrieve images and perform multi-step reasoning over them. MMSearch-R1 (Wu et al., 2025a) employs Group Relative Policy Optimization (GRPO) to incentivize models to actively invoke both image and text search tools, optimizing multimodal search strategies in an end-to-end manner. DeepMMSearch-R1 (Narayan et al., 2025) further introduces external grounding and cropping modules to isolate key regions in complex scenes before retrieval, partially mitigating background noise and improving retrieval effectiveness.

Despite these advances, existing multimodal deep-research MLLMs still suffer from critical limitations. (1) Coarse search strategy. Most prior work depends on one-shot full-level or entity-level image retrieval, making it difficult to reliably identify fine-grained visual entities through visual search engines in realistic web settings. The same for textual search, *i.e.*, reaching the desired target via a search engine often requires multiple attempts. (2) Insufficient optimization for long-horizon visual–text interaction. Current training paradigms typically focus on short-context or fewer-than-five-round retrieval scenarios, lacking systematic data and objective design for long-horizon (tens of rounds) visual and textual search. As a result, models often exhibit premature termination or shallow search behavior in complex tasks.

B. Experiment Setups

Training Data and Models. Our training data and procedures follow the pipeline described in Sec. 2.2. We conduct both supervised fine-tuning (SFT) and reinforcement learning (RL) on Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025), and only perform SFT on Qwen3-VL-8B-Instruct (Bai et al., 2025). For SFT, we use 30K high-quality visual deep research trajectories, while the detailed constitutions are presented in Sec. 2.3.2. Autoregressive supervision is applied at each step of every trajectory, covering the `<think>` reasoning, `<tool_call>` actions, and the final `<answer>`. This stage teaches the model multi-round visual cropping, search, and reasoning behaviors. We adopt Ms-Swift as the training framework. For RL training, we use 15K high-quality VQA instances (described in Sec. 2.3.2) and sample complete trajectories through interaction with a real online search environment. An LLM-as-Judge is employed to evaluate answer correctness and provide reward signals, which are further combined with format constraints (*e.g.*, adherence to the ReAct template) for reward computation and policy optimization. This stage is implemented using the rllm framework (Tan et al., 2025).

Benchmarks. We evaluate our method on 6 challenging benchmarks, including VDR-Bench (Zeng et al., 2026), FVQA (Wang et al., 2017), MMSearch-Plus (Tao et al., 2025a), MMSearch (Jiang et al.), LiveVQA (Fu et al., 2025), and BrowseComp-VL (BC-VL) (Geng et al., 2025).

Specifically, we use the test-mini split of VDR-Bench and the full split of BC-VL. We evaluate on all VQA instances in MMSearch and on the single-image subset of MMSearch-Plus. For FVQA and LiveVQA, we *randomly select* 300 samples

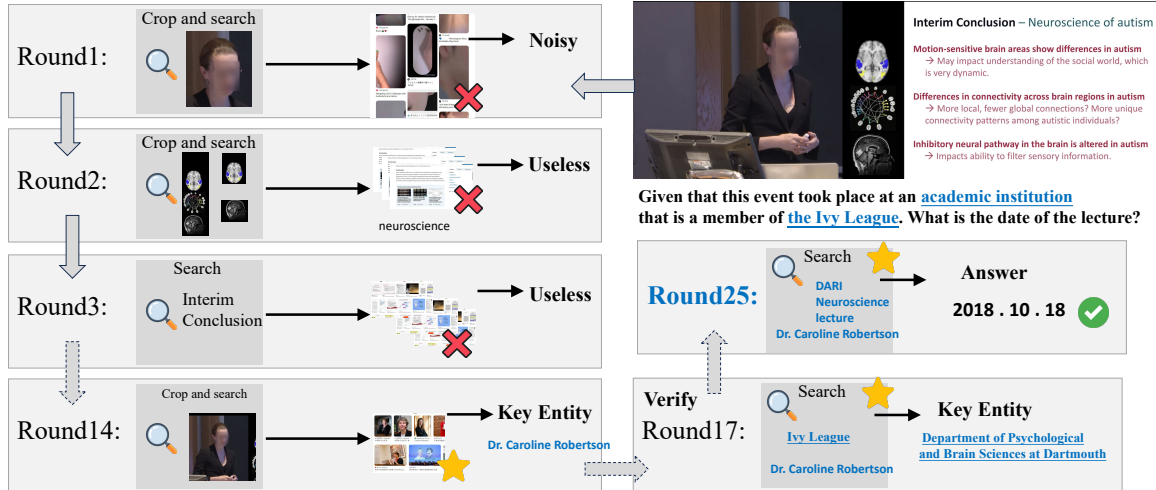


Figure 4. Our Data Pipeline. As shown in the top panel, we construct a complete multimodal deep-research synthesis pipeline. Leveraging the capabilities of an MLLM and a text-based DeepResearch foundation LLM, we generate long-horizon, multi-tool trajectories. As shown in the bottom panel, we obtain high-quality factual VQA instances via a rigorous verification and obfuscation procedure, which are then used for trajectory synthesis and RL training.

per benchmark.

Baselines. We compare our method with a diverse set of proprietary and open-source multimodal models, covering both direct inference and agentic reasoning settings. The evaluated models include the Gemini-2.5 series (Comanici et al., 2025), Claude-Sonnet-3.7/4 models, GPT5 (Singh et al., 2025), the Qwen3-VL family (8B and 30B variants, Instruct and Thinking) (Bai et al., 2025), as well as recent open-source agents such as WebWatcher (Geng et al., 2025) and MMSearch-R1 (Wu et al., 2025a). These models are evaluated under two distinct reasoning paradigms. Direct answer refers to single-pass generation of the answer without calling any external tools or performing information retrieval. ReAct-style agentic reasoning follows a “reasoning-then-tool-call” paradigm, where the model iteratively performs reasoning, calls tools (including multi-scale image cropping, image search, text search, and web browsing) and integrates the retrieved evidence over multiple steps. To ensure a fair comparison, all models are equipped with a unified multimodal toolset and are evaluated using a consistent judge prompt for answer assessment.

C. Case Study

As shown in Fig. 4,