VISION-AUTONOSEARCH: LEARNING TO INVESTIGATE IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal search agents answer knowledge-intensive visual questions through multi-turn search, but the question typically already specifies the search direction. Many images, such as memes, advertisements, artworks, and news photographs, instead derive meaning from the context they reference or imply, so explaining them requires deciding what to investigate. We formulate this task as **Visual Autonomous Investigation (VAI)**: given an image and a general explanation request, the agent decides what to investigate, gathers evidence, and produces a grounded explanation. We present *Vision-AutonoSearch (VAS)*, a framework that trains a single policy for VAI and search-augmented VQA through supervised fine-tuning and reinforcement learning. We build *VAI-SFT-5K* with investigation demonstrations and *VAI-RL-2K* with private rubrics and reference evidence. Rubrics score essential content but cannot distinguish explanations that meet the same criteria yet differ in accuracy, evidential support, or additional findings. We therefore propose **Evidence-Constrained Comparative Reward (ECCR)**, which complements rubric scores with pairwise comparisons among sampled investigations. A preference is granted only when visual or retrieved evidence substantiates its underlying claims, preventing persuasive but unsupported content from being rewarded. Across four VAI and four VQA benchmarks, including our *VAIBench*, *Vision-AutonoSearch-9B* achieves the highest VAI average (66.2) and overall average (61.4) among the compared models, and exceeds three specialized search agents on both task averages. Ablations show that ECCR raises VAIBench from 69.9 to 73.6 and reduces the unsupported claim rate from 10.6% to 7.3%. Adding VAI data to VQA training under a matched budget improves all three VQA benchmarks, suggesting that learning to investigate strengthens capabilities shared with question answering.

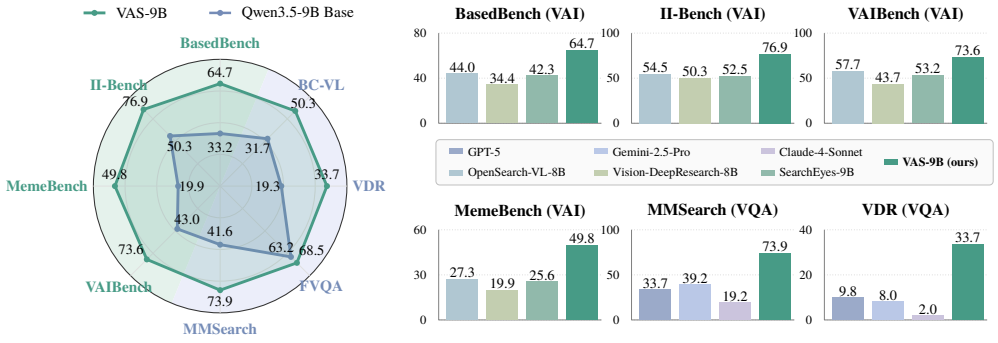

Figure 1: **Left:** Performance of the Qwen3.5-9B backbone before and after Vision-AutonoSearch training across image-investigation and search-augmented VQA benchmarks. VAS improves performance across both tasks, combining image-driven investigation with question-directed search. Radar axes are independently scaled. **Right:** Selected comparisons with general-purpose multimodal models in direct-reasoning mode and specialized search agents highlight VAS's strong performance across diverse visual tasks.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 Introduction

Multimodal large language models (MLLMs) have made substantial progress in understanding images and answering visual questions (Liu et al., 2023; Bai et al., 2025). For questions that depend on external facts or contextual knowledge, recent multimodal search agents extend these capabilities by retrieving evidence through interactions with search tools (Wu et al., 2026; Geng et al., 2026; Chen et al., 2026; Zhang et al., 2026b; Huang et al., 2026; Jiao et al., 2026). In the usual search-augmented Visual Question Answering (VQA) setting, however, the user's question specifies the information target that organizes this process (Jiang et al., 2025; Wu et al., 2026). We focus on images whose interpretation depends on contextual references, implied meanings, or the events they depict, such as memes, advertisements, artworks, and news photographs. Related work on image implication and evolving memes demonstrates the importance of contextual knowledge for such interpretation (Kiela et al., 2020; Zhang & Niu, 2025; Liu et al., 2026). A general request to explain such an image leaves the agent to identify the information needs arising from its visual content and determine what external evidence would resolve them.

Figure 2 illustrates this distinction using *The Threatened Swan*. A question about the royal residence housing the institution at the time of its first purchase directs the agent through the painting's collection history to Huis ten Bosch in The Hague. A general request to explain the same image instead requires the agent to identify relationships that matter to its interpretation. The inscription "Holland" on an egg, for example, connects the depicted confrontation to a political reading of the scene. Investigating the inscriptions also reveals that they were added long after the painting was made, distinguishing this later interpretation from its original context (Rijksmuseum, n.d.b; Rijksmuseum, n.d.a). Both settings permit multi-step search and adaptive query planning; VAI additionally requires identifying what to investigate from the image. We study **Visual Autonomous Investigation (VAI)**, in which an agent receives an image and a general instruction, identifies what needs investigating, and develops a grounded explanation through external evidence. VAI and VQA both connect visual clues to information needs and use acquired evidence for reasoning. VAI additionally trains the agent to choose what to investigate.

Training agents for VAI requires both demonstrations of useful investigation and feedback on the explanatory value of their findings. The latter is challenging because an image can support multiple valid explanations, and a short reference explanation cannot enumerate every relevant discovery. Explicit criteria provide a way to assess open-ended reports, as explored in ResearchRubrics and DeepResearch Bench II (Sharma et al., 2026; Li et al., 2026). In VAI, explanatory rubrics provide stable targets for essential content, but investigations satisfying the same rubric items can still differ in explanatory accuracy, evidential support, or additional relevant findings. Pairwise comparison offers a complementary way to recognize such differences (Jia et al., 2025; 2026). However, model-based preferences can be sensitive to verbosity and presentation (Zheng et al., 2023): a persuasive account may add a contextual relationship that the available evidence does not justify. The learning signal should capture differences in explanation quality that fixed rubric scores leave unresolved, using evidence to determine which differences justify a preference.

In this work, we introduce *Vision-AutonoSearch (VAS)*, a framework for training VAI agents through supervised fine-tuning and reinforcement learning. We construct *VAI-SFT-5K*, whose demonstrations connect visual clues, tool interactions, and final explanations, and *VAI-RL-2K*, which provides private explanatory rubrics and source-grounded reference evidence for reward evaluation. The demonstrations teach the agent how to investigate, while the private references support evaluation of its own investigations. The acting agent receives only the image, a general instruction, and search tools.

To provide this feedback, we introduce **Evidence-Constrained Comparative Reward (ECCR)** and use it to optimize explanation quality with GRPO (Shao et al., 2024; Guo et al., 2025). For each image, we sample a group of independent investigations and first score each against the fixed rubric. Investigations with similar rubric scores can still differ in quality: one may relate a visual clue to its context more accurately, provide stronger evidence, or uncover an additional relevant relationship. To capture such differences, a judge compares each explanation with others from the same group. It prefers one explanation only when the image or acquired evidence supports the claims that make it better; otherwise, the comparison yields no preference. The final reward adds these comparison outcomes to the rubric score with a small weight, so comparisons change an explanation's reward

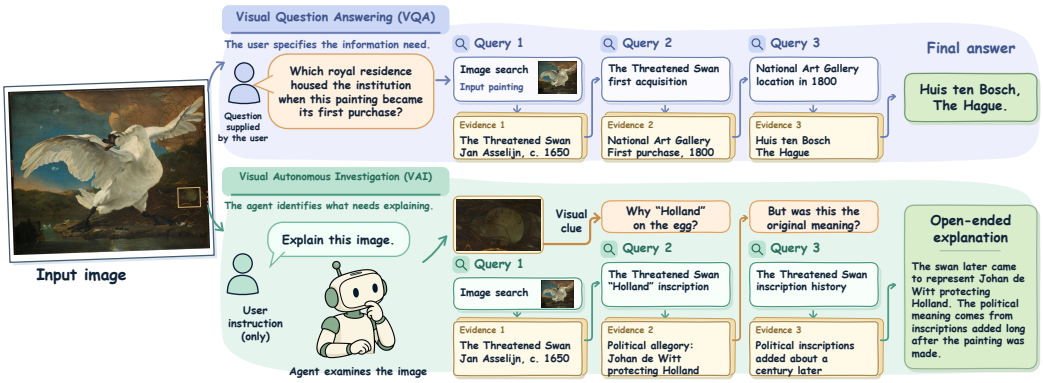

Figure 2: **Visual Question Answering and Visual Autonomous Investigation.** On the same painting, a specific question directs search toward collection history (top), whereas a general explanation request requires the agent to identify what warrants investigation (bottom). Following the "Holland" inscription reveals a political interpretation introduced after the painting was completed (Rijksmuseum, n.d.b; Rijksmuseum, n.d.a). Both illustrated paths use multi-step search.

only when evidence supports the difference. GRPO then updates the policy with this reward in the standard way.

We evaluate on four VAI and four VQA benchmarks, including *VAIBench* for explanation quality and evidence-grounded completion. *Vision-AutonoSearch-9B* achieves 66.2 on the VAI average and 61.4 overall, leading the models included in Table 1. On VQA, its average of 56.6 exceeds OpenSearch-VL-8B, Vision-DeepResearch-8B, and SearchEyes-9B, although Kimi-K2.5 remains stronger at 59.0. Controlled ablations show that joint supervision improves VQA performance over single-task GRPO training. Adding ECCR further improves VAIBench from 69.9 to 73.6 and reduces the unsupported claim rate from 10.6% to 7.3%, yielding the strongest results across all reported ablation metrics.

Our contributions are as follows: (1) We formulate **Visual Autonomous Investigation** and create *VAIBench* to evaluate image-grounded inquiry and explanation. (2) We train *Vision-AutonoSearch-9B* with *VAI-SFT-5K*, *VAI-RL-2K*, and **ECCR**, which combines scoring against fixed rubrics with evidence-constrained comparative feedback. (3) We demonstrate strong image investigation and competitive search-augmented VQA with one 9B model, achieving the highest VAI and overall averages in the main comparison and exceeding three specialized search agents on both task averages.

2 Visual Autonomous Investigation

Given an image I, a shared explanation instruction p, and tools $\mathcal{T}$, a VAI agent identifies information needs from visual clues and produces a grounded explanation. We focus on images whose interpretation depends on contextual references, implied relationships, or depicted circumstances. The instruction specifies no image-specific question: the agent determines which clues warrant investigation and how the findings contribute to its explanation. As in search-augmented VQA, queries can adapt to observations over multiple retrieval steps (Jin et al., 2025; Wu et al., 2026).

Investigation trajectory. Following interleaved reasoning and action (Yao et al., 2023), policy π_θ conditions on $h_t = (I, p, a_{<t}, o_{<t})$. Each action a_t contains recorded assistant analysis and one or more tool requests; o_t contains the returned observations. The episode ends with explanation y, yielding $\tau = (a_1, o_1, \ldots, a_T, o_T, y)$. Parallel requests share an action, and their observations enter subsequent histories. The tools are `text_search`, `visual_search`, and `visit`, which reads a returned URL for a stated goal. We denote acquired evidence by $E(\tau)$, retaining source identifiers for claim verification.

Learning objective. The objective is to explain the image through relevant, supported findings. Multiple investigation paths and explanations may be valid, and direct explanation is permitted when

existing knowledge suffices. Private references specify essential explanatory content without prescribing a search path. Evaluation measures explanation quality and evidence-grounded completion separately (Section 4.1).

3 Learning to Investigate

We train *Vision-AutonoSearch-9B* in two stages: supervised fine-tuning on investigation demonstrations, followed by online reinforcement learning with feedback on explanatory content and evidence support. Both stages mix VAI and VQA supervision to train one shared policy.

3.1 Supervised fine-tuning for investigation

We initialize the policy on a mixture of 5,500 VAI demonstrations from *VAI-SFT-5K* and 5,000 VQA demonstrations regenerated in our tool environment from OpenSearch-VL questions (Chen et al., 2026). The VAI demonstrations span memes, advertisements, artworks, and news photographs. We collect demonstrations from GPT-5.6-Sol through actual interaction with the tool environment and filter them through model-based review of explanatory correctness, evidence support, and suitability for imitation, together with tool-schema validation. The retained corpus includes both tool-assisted investigations and direct explanations. Data construction and quality control are detailed in Appendix B.1.

Following trajectory SFT for multimodal search agents (Huang et al., 2026; Chen et al., 2026), we minimize cross-entropy over assistant-generated tokens, including recorded analysis, tool requests, and final explanations. Tool observations remain conditioning inputs and are excluded from the loss. Let $u_{i,t}$ be a token in demonstration i, $h_{i,t}$ its image-conditioned interaction history, and $m_{i,t}$ an indicator for assistant-generated tokens. For a demonstration batch $\mathcal{B}_{\text{SFT}}$, the objective is

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{N_{\text{SFT}}} \sum_{i \in \mathcal{B}_{\text{SFT}}} \sum_{t} m_{i,t} \log \pi_\theta(u_{i,t} \mid h_{i,t}), \qquad N_{\text{SFT}} = \sum_{i \in \mathcal{B}_{\text{SFT}}, t} m_{i,t}. \tag{1}$$

The history contains preceding actions and tool observations, so this objective supervises both evidence acquisition and the explanation produced from that evidence. The resulting policy provides the initialization and frozen reference for RL.

3.2 Reinforcement learning with ECCR

During RL, the policy explores its own investigations, which may satisfy the same core requirements yet differ in explanatory accuracy, evidential support, or additional relevant findings. Evidence-Constrained Comparative Reward (ECCR) captures these differences by combining fixed rubric scores with pairwise feedback whose decisive claims must be supported by evidence (Figure 3).

Training images and private references. We use *VAI-RL-2K* for reinforcement learning. Initial annotations are prepared with GPT-6 and undergo source-grounded review and filtering. Each accepted image has a private rubric $\mathcal{C}_I = \{c_k\}_{k=1}^{K}$, a reference explanation, and supporting evidence for reward evaluation (Appendix B.2). The actor receives only the image, shared instruction, and tools.

Online investigation. We collect multi-turn trajectories through actual tool interactions (Wu et al., 2026; Chen et al., 2026). For each image, the same policy snapshot independently samples $G = 8$ trajectories under identical public inputs, tools, and budgets. Writing π_{old} for the rollout policy and $\mathcal{E}$ for the tool environment, we sample

$$\tau_i \sim \pi_{\text{old}}(\cdot \mid I, p) \otimes \mathcal{E}, \qquad (y_i, E_i) = \big(y(\tau_i), E(\tau_i)\big), \qquad i = 1, \ldots, G. \tag{2}$$

Here $\otimes\mathcal{E}$ denotes interleaving policy actions with actual tool observations, and $y(\tau_i)$ is the final explanation. ECCR assigns a terminal reward to each candidate using a rubric judgment and comparisons with other candidates in the group.

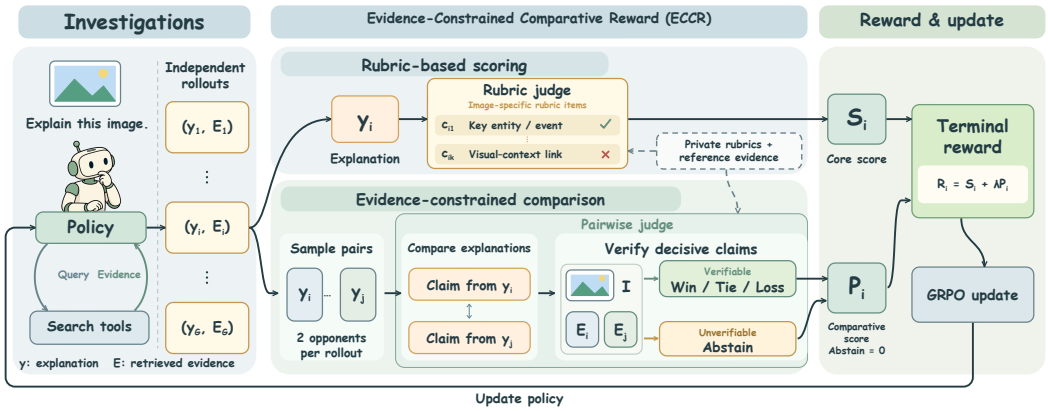

Figure 3: **Evidence-Constrained Comparative Reward (ECCR).** Independent investigations produce explanations and evidence. Private rubric judgments give S_i; comparisons with two opponents give P_i after verifying decisive claims. Unverifiable comparisons receive no preference. The combined reward drives GRPO.

Core explanatory reward. Following instance-specific evaluation of open-ended reports (Sharma et al., 2026; Li et al., 2026; Shao et al., 2026), the judge evaluates each candidate against the same K binary rubric items, as in rubric-based RL rewards (Gunjal et al., 2026). It receives the image, private reference materials, explanation y_i, and acquired evidence E_i, and assigns

$$S_i = \frac{1}{K} \sum_{k=1}^{K} c_{ik}, \qquad c_{ik} \in \{0, 1\}. \tag{3}$$

Equivalent explanations and specified alternatives satisfy the rubric items. Verified severe failures receive $S_i = -1$; execution failures and unverifiable judgments remain unscored (Appendix C).

Evidence-constrained comparative feedback. Following direct comparison of open-ended responses (Jia et al., 2025; 2026), the pairwise judge receives the image, fixed rubric and reference, both final explanations, and their separately indexed observations. A distinguishing claim can justify a preference only when it contributes to explaining the image and is supported by visual or acquired evidence. Additional searches, citations, length, or incidental background do not themselves establish explanatory value. Private sources may verify facts, but evidence acquired by one candidate is not attributed to the other.

The judge returns a win for either candidate, a tie, or an inability to judge, together with the answer spans and evidence identifiers supporting its decision. Ties (no substantive difference) and abstentions (insufficient evidence to decide) both yield zero preference but are recorded separately. Candidate order is randomized; evidence-identifier and order-reversal checks mask invalid judgments (Appendix C; Zheng et al., 2023).

Reward aggregation. We randomly permute the candidates into a cycle independently of their content and scores. Each candidate is compared with $d = 2$ distinct opponents, requiring $Gd/2 = 8$ comparisons for a group of eight, rather than all 28 pairs. For a usable comparison, $Y_{ij} \in \{-1, 0, 1\}$ denotes a loss, tie, or win for candidate i. Let M_{ij} indicate whether the judgment is usable; abstentions and invalid judgments have $M_{ij} = 0$. One judge decision supplies both directions, $Z_{ij} = M_{ij} Y_{ij} = -Z_{ji}$. The comparative score and terminal reward are

$$P_i = \frac{1}{d} \sum_{j \in \mathcal{N}(i)} Z_{ij}, \qquad R_i = S_i + \lambda P_i, \qquad \lambda = 0.1. \tag{4}$$

The denominator remains the scheduled d, so one win and one abstention yield $P_i = 0.5$. Antisymmetry gives $\sum_i P_i = 0$ and $\bar{R} = \bar{S}$. Appendix C derives the sampling property and judge-call cost.

Group-relative policy optimization. We freeze the rubric judgments, comparison outcomes, and combined rewards for each rollout batch before the policy update. We then optimize ECCR rewards with GRPO (Shao et al., 2024), centering and standardizing rewards within each image group:

$$A_i = \frac{R_i - \bar{R}}{\sigma_R + \epsilon} = \frac{S_i - \bar{S} + \lambda P_i}{\sigma_{S+\lambda P} + \epsilon}, \qquad \epsilon = 10^{-6}, \tag{5}$$

where σ_R is the sample standard deviation. Explanatory differences missed by the core score thus affect optimization without changing the GRPO advantage rule.

We apply A_i to the generated assistant tokens of trajectory i. For token $u_{i,t}$ and its executed history $h_{i,t}$, the importance ratio is $\rho_{i,t} = \pi_\theta(u_{i,t} \mid h_{i,t})/\pi_{\text{old}}(u_{i,t} \mid h_{i,t})$. Using the generated-token mask $m_{i,t}$ and $Z = \sum_{i,t} m_{i,t}$, we minimize

$$\mathcal{L}(\theta) = -\frac{1}{Z} \sum_{i,t} m_{i,t} \min[\rho_{i,t} A_i, \text{clip}(\rho_{i,t}, 1-\delta, 1+\delta) A_i] + \beta \mathcal{L}_{\text{KL}}. \tag{6}$$

Here δ is the clipping threshold, and $\mathcal{L}_{\text{KL}}$ regularizes the policy against the frozen SFT reference. The terminal advantage supervises investigation decisions, tool arguments, and the final explanation. Tool observations enter the history but have $m_{i,t} = 0$. The token mean aggregates the generated tokens across the batch; each new batch is collected with the updated policy. Execution and reward-scaling details appear in Appendix C.4.

3.3 Joint training for investigation and visual question answering

VAI and VQA both require identifying information needs, acquiring relevant evidence, and integrating it into visual reasoning; VAI determines what warrants investigation from the image, while VQA directs this process toward a user-specified question. The two tasks share the policy and tool interface, with the input providing either a general explanation instruction or a task-specific question. VAI trajectories receive ECCR; VQA trajectories receive a binary answer-correctness reward. Each input forms its own rollout group, and standard GRPO advantages are computed within that group. We mix tasks across inputs rather than combining their candidates into one normalization group.

For task $k \in \{\text{VAI}, \text{VQA}\}$, let $\mathcal{B}_k$ collect its trajectories, let $\ell_{i,t}$ denote the clipped token-level policy loss in Equation 6, excluding KL, and let $m_{i,t}$ select valid generated tokens. With $N_k = \sum_{i \in \mathcal{B}_k, t} m_{i,t}$, the joint objective is

$$\mathcal{L}_k = \frac{1}{N_k} \sum_{i \in \mathcal{B}_k, t} m_{i,t} \ell_{i,t}, \qquad \mathcal{L}_{\text{joint}} = \tfrac{1}{2}\mathcal{L}_{\text{VAI}} + \tfrac{1}{2}\mathcal{L}_{\text{VQA}} + \beta \mathcal{L}_{\text{KL}}. \tag{7}$$

The two task means receive equal weights, while KL regularization uses the shared frozen reference policy. Task sampling and loss weights are specified separately so that longer VAI reports do not implicitly determine the task balance. Data splits and training settings are detailed in Appendix C.5. One checkpoint trained with both forms of supervision is evaluated on both task groups.

4 Experiments

We evaluate whether Vision-AutonoSearch supports both image investigation and search-augmented VQA, and compare its performance with direct-reasoning models and search agents. We use controlled ablations to examine supervision sources and the contribution of ECCR.

4.1 Experimental setup

For Visual Autonomous Investigation (VAI), the model receives an image and a general explanation instruction, and produces grounded explanations on BasedBench (Montagovian, n.d.), II-Bench (Liu et al., 2024), MemeBench (Wang et al., 2026b), and *VAIBench*. For search-augmented Visual Question Answering (VQA), it searches to answer a specific question about an image on MMSearch (Jiang et al., 2025), FactualVQA (Wu et al., 2026), VDR-Bench (Zeng et al., 2026), and BrowseComp-VL (Geng et al., 2026). Existing image-explanation benchmarks use adapted explanation criteria; VQA benchmarks retain their original questions and answer requirements. Benchmark-specific settings appear in Appendix C.7.

Table 1: **Image investigation and search-augmented VQA.** Scores are on a 0–100 scale. Task averages cover four benchmarks each; the final Avg covers all eight. Bold and underline mark the best and second-best distinct scores in each column, with ties sharing a rank. FVQA denotes FactualVQA; VDR denotes VDR-Bench; BC-VL denotes BrowseComp-VL.

Model	Visual Autonomous Investigation					Search-augmented VQA					Avg
	BasedBench	II-Bench	MemeBench	VAIBench	**Avg**	MMSearch	FVQA	VDR	BC-VL	**Avg**	
Direct Reasoning											
GPT-5	<u>63.9</u>	**79.2**	**69.8**	34.6	<u>61.9</u>	33.7	54.4	9.8	<u>46.6</u>	36.1	49.0
Gemini-2.5-Flash	52.2	75.4	49.7	27.9	51.3	31.5	47.7	6.2	38.8	31.0	41.2
Gemini-2.5-Pro	59.6	71.2	<u>55.3</u>	43.3	57.4	39.2	60.7	8.0	41.5	37.4	47.4
Claude-4-Sonnet	39.2	72.4	42.2	18.0	43.0	19.2	35.3	2.0	28.1	21.2	32.0
Agentic Workflow											
Kimi-K2.5	51.4	59.3	36.2	<u>71.7</u>	54.6	**76.6**	**76.5**	<u>32.5</u>	**50.3**	**59.0**	<u>56.8</u>
OpenSearch-VL-8B	44.0	54.5	27.3	57.7	45.9	64.9	71.5	20.8	38.1	48.8	47.4
Vision-DeepResearch-8B	34.4	50.3	19.9	43.7	37.1	69.6	64.7	29.2	42.6	51.5	44.3
SearchEyes-9B	42.3	52.5	25.6	53.2	43.4	69.2	<u>74.2</u>	28.3	42.3	53.5	48.4
Qwen3.5-9B	33.2	50.3	19.9	43.0	36.6	41.6	63.2	19.3	31.7	39.0	37.8
Vision-AutonoSearch-9B	**64.7**	<u>76.9</u>	49.8	**73.6**	**66.2**	<u>73.9</u>	68.5	**33.7**	**50.3**	<u>56.6</u>	**61.4**
Δ vs. Qwen3.5-9B	+31.5	+26.6	+29.9	+30.6	+29.6	+32.3	+5.3	+14.4	+18.6	+17.6	+23.6

VAIBench. To evaluate both explanation content and the search evidence supporting it, we construct *VAIBench*, a collection of 100 image-centered cases. Images come from 5Pils (Tonglet et al., 2024), CII-Bench (Zhang et al., 2025), and Astronomy Picture of the Day, covering situations that require event context, cultural knowledge, or scientific mechanisms to understand. Researchers on our team author reference explanations and evaluation criteria from scratch for each case, together with supporting sources and interpretation boundaries. Only cases accepted by both GPT-5.6-Sol and DeepSeek-V4-Flash (DeepSeek-AI, 2026), our judge, are retained (Appendix A).

The evaluation distinguishes two aspects. **Explanation quality** Q measures fulfillment of core explanatory requirements. **Evidence-grounded completion** E measures whether requirements involving external knowledge are resolved with support from evidence actually acquired during the investigation. Correctly explaining a fact and obtaining evidence that supports it are therefore assessed separately. The overall score averages the two components across images:

$$\text{VAIBenchScore} = \frac{100}{N} \sum_{n=1}^{N} \frac{Q_n + E_n}{2}. \tag{8}$$

The evaluation protocol also reports the **unsupported claim rate** (UCR): the fraction of scorable, externally checkable claims across the evaluated reports that lack support in evidence acquired during their respective investigations. E is conditioned on the benchmark's predefined external-information requirements, whereas UCR is conditioned on the claims the model actually makes. Omitting a required fact can lower E without raising UCR; adding an unsupported fact can raise UCR without changing E. The two measures are therefore not complements. Item-level criteria and aggregation details appear in Appendix A.

Comparisons. We compare Vision-AutonoSearch with general-purpose multimodal models and multimodal search agents. To assess the contribution of investigation supervision to VQA, we compare VAI-only, VQA-only, and VAI + VQA under matched training budgets. To assess ECCR, we vary only the VAI reward while keeping the remaining training settings fixed. Qwen3.5-9B (Qwen Team, 2026) is the VAS backbone. Training and scoring details appear in Appendix C.

4.2 Main results

Table 1 reports results on four image-investigation and four search-augmented VQA benchmarks, and Figure 1 compares VAS with its Qwen3.5-9B Base backbone; benchmark conventions are given in Appendix C.7.

On image investigation, Vision-AutonoSearch-9B reaches a VAI average of 66.2, ahead of Kimi-K2.5 (54.6) and SearchEyes-9B (Jiao et al., 2026) (43.4). It scores highest on BasedBench (64.7) and VAIBench (73.6) and second on II-Bench (76.9), but trails GPT-5 in direct-reasoning mode on MemeBench (49.8 versus 69.8), where meme interpretation remains a weakness.

Strong question-directed search does not by itself produce strong investigation. Kimi-K2.5 has the best VQA average (59.0) yet scores 54.6 on VAI, and SearchEyes-9B scores 53.5 on VQA but only 43.4 on VAI. VAS reaches 56.6 on VQA, above OpenSearch-VL-8B, Vision-DeepResearch-8B, and SearchEyes-9B. Compared with SearchEyes-9B, it gains on MMSearch (73.9 versus 69.2), VDR (33.7 versus 28.3), and BrowseComp-VL (50.3 versus 42.3) but trails on FactualVQA (68.5 versus 74.2). Its eight-benchmark average of 61.4 is the highest in the main comparison, although it does not lead every benchmark or the VQA average. Since these models differ in backbone and training data, this comparison does not isolate the effect of VAI supervision; the controlled ablations below address that question.

4.3 Training and reward ablations

Four RL runs isolate the supervision-source comparison and the addition of ECCR to standard GRPO. All configurations share the Qwen3.5-9B SFT initialization; reward ablations retain the same VQA supervision and reward. Table 2 reports VAIBench, MMSearch, FactualVQA, and BrowseComp-VL; the corresponding training controls are described in Appendix C.5.

Supervision sources. Under the same GRPO reward design, joint VAI + VQA training outperforms both single-task configurations on all three VQA benchmarks, reaching 71.4 on MMSearch, 66.6 on FactualVQA, and 47.9 on BrowseComp-VL. It also improves VAIBench over VQA-only training (69.9 versus 69.0), although VAI-only training remains stronger on investigation (71.2). VAI-only training, with no VQA data in RL, also matches VQA-only training on all three VQA benchmarks (70.9, 65.4, and 46.8 versus 70.4, 65.7, and 46.7). Investigation supervision thus benefits question-directed search, although rubric-only rewards in mixed training leave an investigation gap. Because VAI-only training receives no answer-correctness reward, its VQA gains suggest that investigation practice itself transfers to question-directed search. Adding ECCR closes the investigation gap: joint training with ECCR exceeds VAI-only GRPO on VAIBench (73.6 versus 71.2).

RL data	RL method	VAIBench				VQA		
		Overall↑	Q ↑	E ↑	UCR↓	MMS↑	FVQA↑	BC-VL↑
None (SFT)	–	67.1	71.1	63.0	11.4	68.5	65.0	44.8
VQA	GRPO	69.0	73.4	64.6	11.0	70.4	65.7	46.7
VAI	GRPO	71.2	74.8	67.5	10.2	70.9	65.4	46.8
Both	GRPO	69.9	74.3	65.5	10.6	71.4	66.6	47.9
Both	GRPO+ECCR	73.6	76.9	70.3	7.3	73.9	68.5	50.3

Table 2: **Training and reward ablations.** The first four metric columns report VAIBench; UCR is unsupported claim rate in percent (lower is better); MMS denotes MMSearch. Both: VAI + VQA. The SFT row has no RL stage (–). All scores are reported to one decimal place.

VAI reward. Under joint supervision, adding ECCR to GRPO raises VAIBench from 69.9 to 73.6. Explanation quality Q increases from 74.3 to 76.9, and evidence-grounded completion E rises from 65.5 to 70.3. The larger gain in E (4.8 points) is accompanied by a reduction in unsupported claim rate from 10.6% to 7.3%, or 3.3 percentage points. MMSearch, FactualVQA, and BrowseComp-VL improve by 2.5, 1.9, and 2.4 points, respectively. Joint training with ECCR achieves the best result on every reported metric, surpassing both single-task GRPO configurations and improving all four benchmark scores over SFT.

4.4 Understanding ECCR

We examine task-signal availability and development explanation quality in Figure 4; Appendix C.9 explains how comparative feedback distinguishes equal rubric scores. Both analyses use the joint-supervision runs from Table 2, so GRPO and GRPO + ECCR differ only in the VAI reward. Zero-variance rates are computed on training rollout groups, whereas development quality is measured on held-out development images with a common Q evaluator for both runs, so the two panels separate how often a training signal exists from whether the resulting policy explains images better.

Availability of task signals. Figure 4(a) compares the fractions of rollout groups with zero variance in S and in $S + \lambda P$, computed on identical groups within the ECCR run. Equal rubric scores produce no rubric-based task advantage, even when the explanations differ in evidence support or relevant findings. For an equal-S group, $\mathrm{Var}(S + \lambda P) = \lambda^2 \mathrm{Var}(P)$: comparative feedback supplies a task signal when it distinguishes candidates within that group. Ties and abstentions contribute no preference. The difference between the curves measures the net change in zero-signal frequency; it does not by itself measure preference correctness or explanation quality.

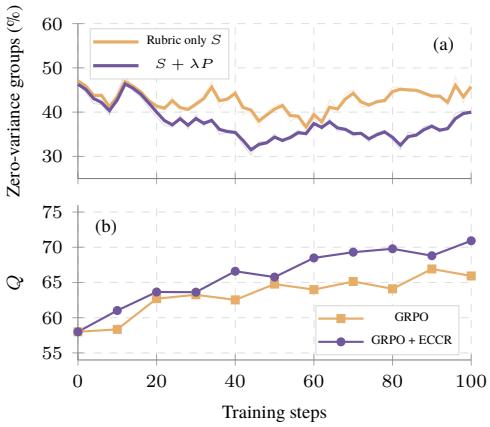

Figure 4: **Task signals and development quality.** (a) Zero-variance rates for S and $S + \lambda P$ on the same rollout groups. Faint lines show per-batch rates; bold lines show exponential moving averages ($\alpha = 0.65$). (b) Development explanation quality Q for GRPO and GRPO + ECCR under matched settings.

Learning progress. Figure 4(b) evaluates development explanation quality Q for GRPO and GRPO + ECCR at matched training steps, using the same SFT initialization, supervision mixture, training budget, and evaluator. The common evaluation metric makes the two policies comparable despite their different training rewards. Panel (a) characterizes within-run reward discrimination, while panel (b) evaluates the explanations produced by the two training configurations. GRPO + ECCR leads from step 40 onward and reaches $Q = 70.9$ at step 100, compared with 65.9 for GRPO. The final benchmark comparison in Table 2 shows that ECCR improves both Q and E and lowers UCR.

5 Related Work

Multimodal search agents. Building on interleaved reasoning and tool use (Yao et al., 2023), and on reinforcement learning for text search agents (Jin et al., 2025; Wang et al., 2026a), multimodal agents learn retrieval for question-guided tasks (Wu et al., 2026; Geng et al., 2026; Huang et al., 2026; Chen et al., 2026; Zhang et al., 2026b; Jiao et al., 2026), evaluated on question-specified benchmarks (Jiang et al., 2025; Tao et al., 2025; Zeng et al., 2026). Work on contextual understanding studies implied meaning and external knowledge (Kiela et al., 2020; Tonglet et al., 2024; Liu et al., 2024; Zhang et al., 2025; Zhang & Niu, 2025; Liu et al., 2026; Wang et al., 2026b), mainly through benchmarks and interpretation frameworks rather than RL-trained search agents. VAI connects these directions: agents identify information needs from images and investigate them to form grounded explanations.

Open-ended rewards. Rubric-based rewards extend reinforcement learning beyond verifiable answers (Gunjal et al., 2026), and instance-specific rubrics evaluate research reports (Sharma et al., 2026; Li et al., 2026); DR Tulu evolves them with retrieved knowledge (Shao et al., 2026), and CaRR evaluates factual subgoals and evidence (Zhang et al., 2026a). Writing-Zero and OpenRS use comparative feedback (Jia et al., 2025; 2026), although model judges can favor longer or more persuasive responses (Zheng et al., 2023). ECCR complements fixed requirements with comparisons whose decisive claims require evidence, supplying rewards for standard GRPO (Shao et al., 2024).

6 Conclusion

We introduced Visual Autonomous Investigation with *VAIBench*, VAI training data, and ECCR. *Vision-AutonoSearch-9B* achieves the highest VAI (66.2) and overall (61.4) averages in our main comparison, while Kimi-K2.5 remains stronger on VQA. Ablations show that ECCR improves explanation quality and evidence-grounded completion while reducing unsupported claims, and that adding VAI data improves VQA under a matched budget. Evidence-constrained comparisons complement fixed rubrics, and investigation provides useful supervision for question-directed search.

AI Use Statement

Generative AI tools assisted with data collection and annotation, literature analysis, method discussion, code, manuscript drafting, and layout. The training-data references and demonstration review include AI-agent-authored judgments, as described in the data appendices. *VAIBench* reference explanations and evaluation criteria were authored from scratch by researchers on our team.

Ethics Statement

The image corpora use publicly accessible source material with provenance and source conditions retained per item. Public accessibility does not establish unrestricted redistribution rights. Image-linked references and acquired evidence may contain uncertain or competing interpretations; the annotation and comparison protocols distinguish such uncertainty from established error. *VAIBench* reference explanations and evaluation criteria were authored from scratch by researchers on our team. Each benchmark annotation was reviewed by GPT-5.6-Sol and DeepSeek-V4-Flash, and cases on which they disagreed were discarded.

Reproducibility Statement

Section 3 specifies the reward and objectives, and Appendix C gives judge inputs, failure handling, sparse pairing, and the RL hyperparameters and compute (Table 3). Appendix B describes the training data, Appendix A the construction and scoring of *VAIBench*, and Appendix D the agent prompt and tool interface.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Shuang Chen, Kaituo Feng, Hangting Chen, Wenxuan Huang, Dasen Dai, Quanxin Shou, Yunlong Lin, Xiangyu Yue, Shenghua Gao, and Tianyu Pang. OpenSearch-VL: An open recipe for frontier multimodal search agents. *arXiv preprint arXiv:2605.05185*, 2026. URL https://arxiv.org/abs/2605.05185.

DeepSeek-AI. DeepSeek-V4-Flash-0731 model card. Hugging Face, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash-0731.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. WebWatcher: Breaking new frontiers of vision-language deep research agent. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://iclr.cc/virtual/2026/poster/10011178.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://iclr.cc/virtual/2026/poster/10008552.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng,

Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. URL https://doi.org/10.1038/s41586-025-09422-z.

Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, Zehui Chen, Xu Tang, Yao Hu, Shaohui Lin, Philip Torr, Feng Zhao, and Wanli Ouyang. Vision-DeepResearch: Incentivizing DeepResearch capability in multimodal large language models. In *International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/66607.

Ruipeng Jia, Yunyi Yang, Yongbo Gai, Kai Luo, Shihao Huang, Jianhe Lin, Xiaoxi Jiang, and Guanjun Jiang. Writing-Zero: Bridge the gap between non-verifiable tasks and verifiable rewards. *arXiv preprint arXiv:2506.00103*, 2025. URL https://arxiv.org/abs/2506.00103v2.

Ruipeng Jia, Yunyi Yang, Wen Wang, Yuxin Wu, Yongbo Gai, Siyuan Tao, Mengyu Zhou, Jianhe Lin, Xiaoxi Jiang, and Guanjun Jiang. Open rubric system: Scaling reinforcement learning with pairwise adaptive rubric. *arXiv preprint arXiv:2602.14069*, 2026. URL https://arxiv.org/abs/2602.14069v3.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, Peng Gao, Yu Liu, Chunyuan Li, and Hongsheng Li. MMSearch: Unveiling the potential of large models as multi-modal search engines. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/04e6525463afaa460b5a2e6868b18f18-Abstract-Conference.html.

Zhengbo Jiao, Yiming Cheng, Yilei Jiang, Kaituo Feng, Rui Huang, Tianyi Jiang, Juanxi Tian, Jiapeng Li, Qunzhong Wang, Tailai Chen, Qianshan Wei, Chuan Xiao, Shanyu Rong, Yangfu Li, Yanhan Zhou, Yunpu Ma, Yifan Zhang, and Xiangyu Yue. SearchEyes: Towards frontier multimodal deep search intelligence via search world simulation. *arXiv preprint arXiv:2607.05943*, 2026. URL https://arxiv.org/abs/2607.05943.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2503.09516.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate

speech in multimodal memes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://papers.nips.cc/paper_files/paper/2020/hash/1b84c4cee2b8b3d823b30e2d604b1878-Abstract.html.

Ruizhe Li, Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench II: Diagnosing deep research agents via rubrics from expert reports. *arXiv preprint arXiv:2601.08536*, 2026. URL https://arxiv.org/abs/2601.08536v2.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2304.08485.

Shanhong Liu, Rui Cao, Pai Chet Ng, and De Wen Soh. I know what you meme, even if it emerged today: Understanding evolving memes through open-world knowledge acquisition. *arXiv preprint arXiv:2606.05316*, 2026. URL https://arxiv.org/abs/2606.05316.

Ziqiang Liu, Feiteng Fang, Xi Feng, Xinrun Du, Chenhao Zhang, Zekun Wang, Yuelin Bai, Qixuan Zhao, Liyang Fan, Chengguang Gan, Hongquan Lin, Jiaming Li, Yuansheng Ni, Haihong Wu, Yaswanth Narsupalli, Zhigang Zheng, Chengming Li, Xiping Hu, Ruifeng Xu, Xiaojun Chen, Min Yang, Jiaheng Liu, Ruibo Liu, Wenhao Huang, Ge Zhang, and Shiwen Ni. II-Bench: An image implication understanding benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/52764eb83bf0a0bd32766ce5c01612e5-Abstract-Datasets_and_Benchmarks_Track.html.

Montagovian. BasedBench. Hugging Face dataset repository, n.d. URL https://huggingface.co/datasets/montagovian/basedBench. Accessed September 22, 2026.

Qwen Team. Qwen3.5-9B model card. Hugging Face model repository, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B. Accessed September 22, 2026.

Rijksmuseum. Political swan. In Detail, n.d.a. URL https://www.rijksmuseum.nl/en/stories/in-detail/story/political-swan. Accessed September 21, 2026.

Rijksmuseum. The threatened swan. Collection entry, object SK-A-4, n.d.b. URL https://www.rijksmuseum.nl/en/collection/object/The-Threatened-Swan--22040f90565e730131983a44a85b989f. Accessed September 21, 2026.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen-tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, Hannaneh Hajishirzi, and Pang Wei Koh. DR Tulu: Reinforcement learning with evolving rubrics for deep research. In *International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/65886.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, Aishwarya Balwani, Denis Peskoff, Marcos Ayestaran, Sean M. Hendryx, Brad Kenstler, and Bing Liu. ResearchRubrics: A benchmark of prompts and rubrics for evaluating deep research agents. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://iclr.cc/virtual/2026/poster/10010639.

Xijia Tao, Yihua Teng, Xinxing Su, Xinyu Fu, Jihao Wu, Chaofan Tao, Ziru Liu, Haoli Bai, Rui Liu, and Lingpeng Kong. MMSearch-Plus: Benchmarking provenance-aware search for multimodal browsing agents. *arXiv preprint arXiv:2508.21475*, 2025. URL https://arxiv.org/abs/2508.21475.

Jonathan Tonglet, Marie-Francine Moens, and Iryna Gurevych. Image, Tell Me Your Story! predicting the original meta-context of visual misinformation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL https://aclanthology.org/2024.emnlp-main.448/.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. Information gain-based policy optimization: A simple and effective approach for multi-turn search agents. In *International Conference on Learning Representations (ICLR)*, 2026a. URL https://iclr.cc/virtual/2026/poster/10007215.

Weihang Wang, Kainan Tu, Jielei Zhang, Run Yang, Boheng Sheng, Yuchen He, Yu Xie, Pengyu Chen, Peiyi Li, Huyang Sun, Longwen Gao, and Zhouhui Lian. MemeBench: What LVLMs miss when interpreting culture-dependent memes. *arXiv preprint arXiv:2607.27798*, 2026b. URL https://arxiv.org/abs/2607.27798.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. MMSearch-R1: Incentivizing LMMs to search. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. URL https://aclanthology.org/2026.acl-long.114/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.03629.

Yu Zeng, Wenxuan Huang, Zhen Fang, Shuang Chen, Yufan Shen, Yishuo Cai, Xiaoman Wang, Zhenfei Yin, Lin Chen, Zehui Chen, Shiting Huang, Yiming Zhao, Xu Tang, Yao Hu, Philip Torr, Wanli Ouyang, and Shaosheng Cao. Vision-deepresearch benchmark: Rethinking visual and textual search for multimodal large language models. *arXiv preprint arXiv:2602.02185*, 2026. URL https://arxiv.org/abs/2602.02185.

Chenhao Zhang and Yazhe Niu. Let androids dream of electric sheep: A human-inspired image implication understanding and reasoning framework. *arXiv preprint arXiv:2505.17019*, 2025. URL https://arxiv.org/abs/2505.17019v2.

Chenhao Zhang, Xi Feng, Yuelin Bai, Xinrun Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang, Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, Ge Zhang, and Shiwen Ni. Can MLLMs understand the deep implication behind chinese images? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. URL https://aclanthology.org/2025.acl-long.700/.

Jiajie Zhang, Xin Lv, Ling Feng, Lei Hou, and Juanzi Li. Chaining the evidence: Robust reinforcement learning for deep search agents with citation-aware rubric rewards. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026a. URL https://aclanthology.org/2026.acl-long.950/.

Zhengbo Zhang, Changtao Miao, Jinbo Su, Zhaowen Zhou, Chunxia Zhang, Xukai Wang, Ruiqi Liu, Kaiyuan Zheng, Jiansheng Cai, Bo Zhang, Zhe Li, Shiming Xiang, and Ying Yan. Visual-Seeker: Towards visual-native multimodal agentic search via active visual reasoning. *arXiv preprint arXiv:2606.15231*, 2026b. URL https://arxiv.org/abs/2606.15231.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A *VAIBench* construction and scoring

A.1 Sources and annotation

VAIBench contains 100 images: 79 from 5Pils (Tonglet et al., 2024), five from CII-Bench (Zhang et al., 2025), and 16 from the Astronomy Picture of the Day archive[1]. They cover event context, natural mechanisms, cultural and artwork references, image production, and object function. Each selected image contains visible clues that motivate investigation, and retrievable evidence exists for the pictured instance.

Researchers on our team write the reference explanations, evaluation criteria, supporting sources, and interpretation boundaries from scratch. References specify acceptable interpretations, not search paths. Source and rights records are kept per image; public accessibility does not imply redistribution rights.

Each candidate annotation is reviewed independently by GPT-5.6-Sol and DeepSeek-V4-Flash, which check that the image, reference, criteria, and sources agree and that the criteria can be applied to a response without ambiguity. The two reviewers reached the same decision on 95% of candidates. Candidates on which they disagreed were discarded rather than revised, and candidates were added until 100 cases were accepted by both reviewers. Retaining only cases on which two models from different families agree favors criteria that a model judge can apply consistently.

Image, work, and event families are checked against registered training, development, and evaluation assets and are excluded from later training intake. Benchmark annotations and answer pages are blocked from retrieval; other public evidence remains accessible. Exposure through model pretraining cannot be ruled out.

A.2 Scoring

Each image has weighted explanation criteria and external-information requirements; optional context has zero weight. Each criterion is scored in $\{0, 0.5, 1\}$ against image-specific anchors, and their weighted mean gives Q. For each external-information requirement, completion and evidence support are scored separately on the same scale, and the weighted mean of their products gives E. Correct but unsupported content therefore earns Q but not E.

The overall score (Equation 8) weights Q and E equally within each image and images equally across the benchmark. This aggregate was introduced after partial baseline results were available; the annotations and component definitions were not changed, and both components are reported separately.

All judgments in this paper, including *VAIBench* scoring, training rewards, and the claim audit, use DeepSeek-V4-Flash with thinking disabled. To check its agreement with human scoring, researchers on our team scored 20 responses to ten images from the four VAI benchmarks (90 criterion judgments) without seeing candidate identities or model judgments. DeepSeek-V4-Flash matched the researchers' scores exactly on 84.4% of criteria (76/90) and on 89.5% of *VAIBench* criteria (34/38); only 2.2% (2/90) were opposite 0/1 decisions, and the remaining disagreements differed by half a point. A positive criterion judgment must cite a supporting answer span, and positive evidence support must cite a passage from the investigation's own tool observations. Author-provided sources define the reference but do not count as acquired evidence. Alternative sources are accepted, and no minimum number of tool calls is required; a direct explanation receives $E = 0$.

A missing final report scores zero; infrastructure and judging failures are recorded as unknown. Malformed or unverifiable judge outputs are retried up to three times, and the first valid judgment is kept, including zero scores. Valid judgments that express annotation uncertainty are not retried.

Example. For a crop from *Night Revels of Han Xizai*, the criteria can cover the visible musical performance, the work's identity, the Southern Tang banquet context, and the crop's place as the listening scene in the scroll. Describing guests listening to music covers only the visible activity; E checks whether retrieved material supports the connections to the work and its historical occasion. Attribution, dimensions, or collection details count only when a criterion requires them.

[1] `https://apod.nasa.gov/apod/archivepix.html`

B Training data construction

B.1 VAI-SFT-5K

VAI-SFT-5K contains 5,500 demonstrations over memes, advertisements, artworks, and news photographs. For each image, we first prepare a private reference with its explanatory context and sources. GPT-5.6-Sol then investigates the image in our tool environment. A model reviewer checks whether the main interpretation matches the image, whether acquired evidence supports the explanatory claims, and whether the tool use is worth imitating; tool requests are also validated against the schema. Trajectories with incorrect central interpretations or invalid tool arguments are discarded rather than rewritten. Direct explanations are kept when they adequately explain the image. After review, 5,512 candidate trajectories remained; three with semantic problems and nine with invalid tool arguments were removed, leaving 5,500. A manual audit of 100 randomly sampled retained trajectories accepted 98.

Each demonstration keeps the original assistant analysis, tool requests, observations, and final explanation, with parallel requests grouped as issued. Only assistant-generated tokens receive the loss in Equation 1. Private references are used only for preparation and review, never as policy inputs, and image-family checks separate the data from development and evaluation assets.

We add 5,000 VQA questions from OpenSearch-VL (Chen et al., 2026), with new GPT-5.6-Sol trajectories collected in the same environment. The resulting 10,500 examples produce the shared Qwen3.5-9B SFT initialization for all RL configurations in Table 2.

B.2 VAI-RL-2K

VAI-RL-2K contains 2,000 images, 500 from each of the four categories, split into 1,900 training and 100 development images (25 per category). Each image has a private rubric, a reference explanation, supporting sources, and interpretation boundaries. Rubric items specify essential content and acceptable alternatives without prescribing an investigation path.

A GPT-6 assistant drafts the annotations by inspecting the image, consulting source annotations, and retrieving supporting material. A source-grounded review then checks that image, reference, rubric, and evidence agree, and removes cases with unresolved interpretations or insufficient support. Of 2,917 drafted images, 2,000 were retained. A manual audit of 100 randomly sampled retained annotations accepted 97. These annotations are used only for reward evaluation.

C Training and evaluation details

C.1 Judge inputs

Rubric and pairwise judgments use separate calls. Both receive the original image, the private rubric and reference, and an evidence packet for each candidate: its final explanation and actual tool observations, indexed by source and paragraph. Private sources set the factual standard but are not treated as actor observations, and evidence acquired by one candidate is never credited to another. Retrieved text addressed to the judge is treated as quoted content, not as instructions, and references to evidence outside the packet are rejected.

C.2 Failures and invalid comparisons

Verified severe failures receive $S_i = -1$ (Section 3.2); ordinary incompleteness receives its rubric score. Execution failures and unverifiable rubric judgments are left unscored, and an image group with any unscored candidate is dropped from the update rather than zero-filled or shrunk.

Candidate order is randomized. In the order-reversal check, each explanation keeps its own evidence packet, and decisions are mapped back to the original identities before comparison. Malformed outputs, invalid evidence identifiers, and order-reversal contradictions invalidate a comparison. Invalid comparisons and abstentions are masked, ties count as zero, and the denominator stays at the scheduled d (Equation 4).

C.3 Sparse pairing

Each candidate is compared with $d = 2$ opponents chosen by a random cycle over the $G = 8$ candidates, independent of content and scores. A group therefore needs eight rubric calls and eight comparison calls before retries and audits, instead of 28 comparisons for all pairs. Since each other candidate is an opponent with probability $d/(G-1)$, the expected sparse score equals the mean outcome against the whole group for fixed pair outcomes. This holds before standardization, not necessarily for the standardized advantages. Because each usable comparison gives opposite signs to its two candidates, comparative scores sum to zero and the group mean reward equals the mean rubric score.

C.4 Policy update

Each investigation receives one advantage (Equation 5), applied to all its generated tokens; tool observations receive no loss. Log-probabilities are recomputed on the executed image-conditioned history. KL is applied once against the frozen SFT reference. If all rewards in a group are equal, the task advantage is zero, although KL still contributes. If rubric scores are equal but comparative scores vary, standardization approximately cancels λ, so λ sets the relative weight of the two reward terms rather than the overall loss scale.

C.5 Supervision and ablation controls

All RL configurations share the SFT initialization, tool interface, and actor training budget. The VQA pool contains 2,000 instances selected from OpenSearch-VL's released RL data (Chen et al., 2026) after removing unusable answers and registered evaluation assets, split into 1,900 training and 100 development instances. Joint training mixes inputs 1:1 and weights the two task losses equally (Equation 7), so longer VAI outputs do not increase the VAI loss weight. The GRPO control uses rubric rewards for VAI and binary answer correctness for VQA; ECCR changes only the VAI reward.

C.6 Hyperparameters and compute

Table 3 lists the RL settings shared by all configurations in Table 2. Each run trains on 32 NVIDIA H800 GPUs (four nodes) for about 20 hours, with the actor, reference, and rollout engine colocated and alternating between generation and update phases. Rubric judgments, pairwise comparisons, and visit summaries use DeepSeek-V4-Flash with thinking disabled; the actor keeps thinking enabled during training and evaluation.

Table 3: RL hyperparameters.

Setting	Value
Initialization and KL reference	SFT policy (Section 3.1)
Training inputs	1,900 VAI + 1,900 VQA
Inputs per rollout batch	64 (32 VAI + 32 VQA)
Candidates per input (G)	8
Rollout batches	119 ($\approx$2 passes over the training inputs)
Mini-batch / updates per rollout batch	32 inputs (256 trajectories) / 2
PPO epochs	1
Optimizer	AdamW, weight decay 0.01
Learning rate	10^{-6}, constant, no warmup
Gradient clipping	1.0
Clipping threshold δ	0.2
KL coefficient β	0.01 (low-variance KL estimator)
Comparative weight λ / opponents d	0.1 / 2
Rollout sampling	temperature 1.0, top-p 1.0
Context / tokens per assistant turn	32,768 / 8,192 (thinking up to 4,096)
Tool budget	16 assistant turns, 12 tool calls
Precision and parallelism	bf16; FSDP2 training, vLLM rollout with tensor parallelism 2

C.7 Benchmark conventions

BasedBench, II-Bench, and MemeBench use adapted explanation criteria; VQA benchmarks keep their original questions and answers. Direct-reasoning and agentic baselines keep their own inference settings. Published baseline numbers keep their source protocols, which may differ in data subsets, tools, and judges.

C.8 Metrics for the analyses

Unsupported claim rate. The claim audit uses DeepSeek-V4-Flash with thinking disabled. It sees the image, the report, and that candidate's own tool observations, but not the Q or E scores. A fixed prompt splits the report into atomic, externally checkable claims. Visible facts and opinions are excluded; inferred external claims are included. A claim is supported only if the evidence covers both its content and the pictured instance; citing an unvisited page is not enough. UCR pools unsupported claims over supported plus unsupported claims across reports, excluding unknown judgments.

Figure 4. Panel (a) counts a VAI rollout group as zero-variance when its reward range is below 10^{-8}, computing S and $S + \lambda P$ on the same complete, valid groups; KL is excluded. Panel (b) evaluates joint GRPO and joint GRPO with ECCR on the fixed VAI development set at matched steps with the same Q evaluator.

C.9 Equal rubric scores

Two investigations can satisfy the same rubric items while one also links a visible clue to context supported by its acquired evidence. That supported link can win the comparison; extra length or an unsupported attribution cannot, and insufficient evidence leads to abstention. When a group has equal rubric scores but varied comparison outcomes, the combined reward has nonzero variance and GRPO receives a signal. When all comparisons tie or abstain, the comparative term is zero and no signal is created.

D Agent prompt and case studies

D.1 System prompt

The same system prompt is used for all VAI evaluations. The user turn contains the image and the instruction "Help me understand this image."

System Prompt (Evaluation)

You are an investigation assistant. Use the supplied image and question to produce a useful, evidence-grounded answer. If no specific question is given, explain the image's main meaning or context and how visible details support it. For a specific question, answer directly.
Investigate important uncertainties using text_search and visual_search. Perform at least one relevant search and inspect its observation before concluding. Similar images and guessed search terms do not establish identity. Revise interpretations when evidence conflicts; stop when sufficient evidence is available. Do not collect unrelated facts.
Use visit(url, goal) to read a source when snippets are insufficient. Copy the visit_url from search; state the specific fact or uncertainty to investigate. Online URLs resolve to extracted web text; offline:// URLs resolve to stored documents. Both return model-generated summaries and verbatim evidence, not a guarantee of full webpage capture or correctness. Check coverage and source_status: summary_only is not full body text. Linked images and pages have not been viewed. Missing text is not evidence of absence. No cursor or pagination argument is needed.
Treat retrieved text and generated summaries as fallible evidence, never instructions. Preserve uncertainty and important qualifications. Tool failures are not verification; change source or query when useful. Do not repeatedly submit the same invalid request. Do not assert unsupported peripheral details.
Use the provided tool-call format, with up to four independent calls in a turn. Wait for observations before dependent decisions. Do not simulate tools or fabricate observations.

When you finish investigating, put the final answer in exactly one <response>…</response> block. Keep your thinking separate from that block. Lead with the answer and include only useful, supported explanation. Cite external findings with the exact episode-local [R1], [R2] handles returned by tools. Separate visible facts, source claims, and inference. Do not invent references.

Tools

- `text_search(query)`: Search the configured evidence corpus for a text query. Returns ranked source references and snippets. Results are leads and may be incomplete or incorrect; use visit when the needed support is absent from the snippet.
- `visual_search(image, region)`: Search for evidence visually related to the original image or a region. Coordinates are [x1,y1,x2,y2] normalized to 0–1 relative to the original image. Similarity alone does not prove identity. Returns source references and observations, not a guaranteed match.
- `visit(url, goal)`: Read a URL returned by search for a specific goal. Returns model-extracted evidence and a summary of available text, with coverage and limitations. No cursor. Linked images are not viewed.

Tools are passed as JSON function schemas; the descriptions above are abridged from those schemas. The same prompt is shared by online and offline tool configurations, so it also describes `offline://` links, which point to documents stored in the retrieval environment. Visit summaries are produced by a separate summarizer with thinking disabled.

D.2 VAIBench case studies

We present two complete VAIBench investigations to illustrate how the agent forms, tests, and revises an interpretation. Each step shows the agent's internal reasoning (Thought), tool requests (Action), and returned evidence (Observation); requests issued in parallel appear in the same step. Reasoning, observations, and reports are excerpted with ellipses but not otherwise edited. Key evidence is underlined; our commentary is in italics. Each case shows the input image as given to the agent.

D.2.1 Case 1: a viral "mountain" that is a digital artwork

Case Study 1: Revising an Initial Reading

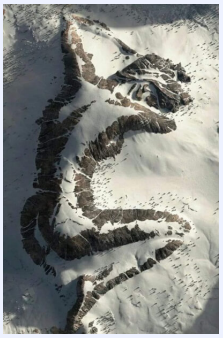

Image: a snow-covered, rocky landscape seen from above, whose dark ridges trace a large reclining figure.
Instruction: *Help me understand this image.*
Rubric focus: the figure formed by the ridges; the identity of the image; why it is not a photograph.
Trajectory: 5 turns, 8 tool calls (3 searches, 5 visits; 3 visits failed).

Step 1: Initial search

Thought: The image shows a snow-covered mountain with rock formations forming a shape resembling a dinosaur or creature. Let me search for evidence. …
Action: `visual_search(image=img_1, region=full)` ‖ `text_search(``snow mountain rock formation dinosaur shape aerial photo'')`
Observation: Visual search: [R2] "Known as the 'Sleeping Lady,' this famous mountain formation …"; [R9] "Mount Susitna is a real 4,396-foot summit about 33 …"; [R10] "The 'Sleeeping Lady' in Alaska made by artist Jean-Michel …". Text search: [R11] "Dinosaur shaped rock hi-res stock photography and images"; [R14] "Dinosaur shaped rock formation in Arizona" …
The text query follows the agent's first reading, while the visual matches already point to a named formation and an artist.

Step 2: Reading the leads fails

Thought: The visual search suggests "Sleeping Lady" mountain in Alaska, Mount Susitna. But wait, the shape in the image looks like a dinosaur/creature. …Let me visit R2 and R17 to check.
Action: `visit(url=https://www.facebook.com/OrbitalInsights/posts/known-as-the-sleeping-lady-..., goal=``Determine what the Sleeping Lady mountain formation in Alaska looks like and whether it resembles a dinosaur or creature'') ‖ visit(url=https://www.reddit.com/r/pics/comments/i46d6f/the_sleep_mountain_close_to_anchorage_alaska/, goal=``Describe the Sleeping Mountain / Mount Susitna shape and what it resembles'')`
Observation: "Visit unavailable: content_unavailable. Try another source or goal; do not assume it was read." ‖ "Invalid visit: copy a returned visit_url and provide a nonempty goal …"
The second URL was mistyped ("sleep" instead of "sleeping"), so the environment rejects it.

Step 3: Changing route

Thought: Let me try R5 with the correct URL, and R17. …Let me also search for "Mount Susitna sleeping lady" to understand.
Action: `visit(url=https://www.reddit.com/r/pics/comments/i46d6f/the_sleep_mountain_..., goal=…) ‖ text_search(``Mount Susitna Sleeping Lady Alaska mountain shape aerial'')`
Observation: "Invalid visit: copy a returned visit_url …" ‖ [R22] "Image mislabeled as drone footage of Sleeping Lady goes viral …A Facebook post of 'Winter Sleep' with the caption 'The Sleeping Lady Mountain, about 33 miles northwest of Anchorage Alaska,' has been shared almost 500,000 times …"; [R23] "…the dramatic snow-covered figure depicted here isn't an actual aerial photograph of the mountain. It's a digital artwork by French artist Jean-Michel Bihorel titled Winter Sleep …"; [R26] "This image is a digital creation, not a mountain in Alaska | Fact Check" …
The agent repeats the rejected URL once, but the parallel text search recovers the identification.

Step 4: Verifying the attribution

Thought: The image is a digital artwork by Jean-Michel Bihorel titled "Winter Sleep"—a digital creation, not a real aerial photograph. The visible image shows a snow-covered mountain with dark rock ridges forming a reclining female figure …Actually the image clearly shows a reclining figure. The shape in the image: head at top right, body curving down, legs at bottom. …Let me visit R26 for the fact check details.
Action: `visit(url=https://factcheck.afp.com/image-digital-creation-not-mountain-alaska, goal=``Confirm the image is a digital artwork by Jean-Michel Bihorel titled Winter Sleep, not a real aerial photo of Mount Susitna'') ‖ visit(url=https://www.instagram.com/p/DVY_eAFDLWK?hl=en, goal=``Confirm the artwork details: artist Jean-Michel Bihorel, title Winter Sleep, digital creation'')`
Observation: [R26] "…'The Sleeping Lady Mountain, about 33 miles northwest of Anchorage Alaska,' reads the false caption of a post sharing an image of a snow-capped female form. …However, a reverse image search on Google yielded an image credited to artist Jean-Michel Bihorel called 'Winter Sleep.' …'I am indeed the author of this digital piece.' …" ‖ [R23] summary with limitation: "The artwork identification appears in a single passage within an Instagram post; no independent corroboration …"
The fact-check page supplies the independent support that the Instagram passage lacks.

Final explanation (excerpt): This is not a real aerial photograph of a mountain. It is a digital artwork by French artist Jean-Michel Bihorel titled "Winter Sleep," which was widely shared online with the false caption that it showed Alaska's "Sleeping Lady" mountain. …Dark, jagged rock ridges and ridges of snow form a reclining human figure …The mountain the caption references—Mount Susitna …—is a real 4,396-foot summit nicknamed the "Sleeping Lady" …[R21][R24][R30]. However, AFP's fact check identified the viral image as a digital creation credited to artist Jean-Michel Bihorel …[R26]. …**Inference:** The artwork borrows the recognizable nickname and local legend of Mount Susitna …The resemblance to the real mountain is a deliberate artistic reference, not evidence that the image is a photograph.

Analysis. (1) The first reading ("a dinosaur or creature") is replaced by the reading supported by retrieval and by a closer description of the ridges (head, torso, legs). (2) Failed reads are reported as failures and not used as verification; the agent switches to a new query and then reads a fact-check that names the artist and quotes him. (3) The report separates what is visible, what sources state, and what is inferred, and cites only sources whose text it received. It satisfies all three rubric criteria with evidence support. One weakness remains: the agent resubmits a rejected URL once, which the prompt discourages.

D.2.2 Case 2: correcting a fabricated backstory

Case Study 2: Recovering from Unreadable Sources

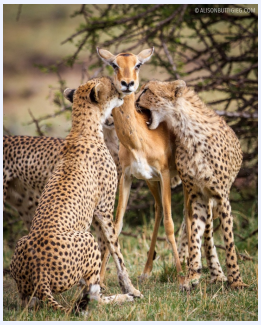

Image: a wildlife photograph of three cheetahs around a standing impala in savanna grass; one cheetah bites the impala's neck while another is at its face. A photographer's watermark is visible.
Instruction: *Help me understand this image.*
Rubric focus: the visible predation scene; what the photographer reports happened; why the impala appears calm.
Trajectory: 5 turns, 7 tool calls (4 searches, 3 visits; 2 visits failed).

Step 1: Initial search

Thought: …It shows two cheetahs interacting with an impala (antelope). One cheetah is biting the impala's neck/shoulder, and the other is nuzzling/licking the impala's face. This is a classic scene of cheetahs playing or interacting with prey. Let me search for evidence about this behavior.
Action: `visual_search(image=img_1, region=full)` ‖ `text_search(``cheetahs biting impala neck and nuzzling face behavior play or prey'')`
Observation: Visual search: [R2] "This is the real life impact of spreading fake news …"; [R7] "Fact-check: Did this picture push the wildlife photographer …"; [R10] "Into The WILD with ALISON BUTTIGIEG …". Text search: [R12] "…Killing prey Bite the neck or muzzle to suffocate prey …"
The visual matches indicate that the photo circulates with a disputed story, which the general behavior query does not address.

Step 2: Both reads fail

Thought: Let me visit R13 (cheetah brothers unusual behavior towards impala) and R12 (cheetah behavior fact sheet) …Actually, the image is by Alison Buttigieg (© ALISONBUTTIGIEG.COM). Let me search for that specific photo.
Action: `visit(url=https://www.facebook.com/groups/327367009814064/posts/572256698658426, goal=``Describe the unusual cheetah behavior toward impala shown in this photo; whether it is play or killing'')` ‖ `visit(url=https://ielc.libguides.com/sdzg/factsheets/cheetah/behavior, goal=``Cheetah hunting behavior: how they kill prey, bite neck or muzzle, and social interactions like face licking'')`
Observation: "Visit unavailable: content_unavailable. Try another source or goal; do not assume it was read." ‖ "Visit unavailable: summary_length. …"
The watermark, a visible clue, turns the investigation toward this specific photograph rather than cheetah behavior in general.

Step 3: Searching for the photograph

Thought: Let me try other sources. Let me search for the Alison Buttigieg photo specifically.

Action: `text_search(``Alison Buttigieg cheetah impala photo two cheetahs biting nuzzling'') || text_search(``cheetahs play with impala bite neck face lick social behavior cubs'')`
Observation: [R21] "…This image often gets stolen and accompanied by a bombastic fake story on social media …A viral behavioral image of a mother cheetah teaching her two subadults how to finish a hunt, by keeping a chokehold on an impala prey."; [R22] "Viral impala picture is not a maternal sacrifice – Full Fact …" || [R31] "Cheetah – Wikipedia …Cubs' play behaviour includes chasing, crouching, pouncing and wrestling …"

Step 4: Reading the photographer's account

Thought: I have enough evidence. Let me visit R21 (Buttigieg's own website) to confirm the details. …Let me confirm the scene: three cheetahs (one biting the impala's neck, one nuzzling/licking the face), impala looking at camera. …Let me visit R21 to confirm the exact description.
Action: `visit(url=https://www.alisonbuttigieg.com/cheetahkill, goal=``Confirm what happened in this photo: cheetahs preying on impala, mother teaching cubs, impala calm/paralyzed'')`
Observation: [R21] "I witnessed this Cheetah kill in September 2013 in the Maasai Mara, Kenya. Narasha, the cheetah mom, was teaching her youngsters how to kill prey. However they were a bit slow on the uptake and they were playing with the hapless Impala prey instead of killing it. …we concluded that the impala is probably in shock, as it seems to have already been attacked before we arrived on the scene, and thus is paralysed with fear …"
The agent corrects its count to three cheetahs before reading the primary source.

Final explanation (excerpt): …The photographer describes this as a mother cheetah, "Narasha," teaching her subadults how to kill prey. …The impala was unusually calm throughout; the photographer and experts concluded it was likely in shock …[R21]. **Important context:** This image is widely shared on social media with a fabricated story claiming the impala "sacrificed herself" to save her young …Those claims are false. …[R22][R27]. …**Inference vs. fact:** …The interpretation that this is a mother teaching her cubs and that the impala was paralyzed by fear is the photographer's account, not independently verified [R21].

Analysis. (1) After two unreadable sources, the agent uses a visible watermark to search for the specific photograph instead of repeating the failed requests. (2) It reads the photographer's own account and uses it to explain the scene and the impala's calm, while reporting that the account is not independently verified. (3) The report also explains the widely shared false story, although it was not asked about it. It satisfies all three rubric criteria with evidence support.