TACT: ALLEVIATING CREDIT MISASSIGNMENT IN GROUP-BASED POLICY OPTIMIZATION FOR SEARCH AGENTS WITH TOOL-AWARE CREDIT TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Mainstream group-based policy optimization methods broadcast trajectory-level advantages to all tokens, potentially penalizing useful steps in failed trajectories as heavily as unhelpful ones. We refer to this over-penalization of useful tool calls as credit misassignment. In search scenarios, many tool calls in failed trajectories have parameters similar to those of calls in successful trajectories for the same question, providing a signal for identifying and mitigating credit misassignment. Building on this insight, we propose Tool-Aware Credit Transfer (TACT), which leverages the parameter–intent correspondence of information-acquisition tools: similar call parameters provide probabilistic evidence of similar information-seeking intent, guiding conservative credit transfer. TACT constructs counterfactual references within the current training batch and attenuates potentially misassigned negative credit via confidence-gated conservative advantage correction. It requires no additional annotation, models, or sampling during training, and introduces negligible computational overhead. Across multiple search benchmarks, TACT delivers consistent, plug-and-play improvements over strong baselines for three group-based policy optimization algorithms (GRPO, GSPO, and SAPO). Our code and models will be publicly released upon acceptance.

1 INTRODUCTION

Agent-based reinforcement learning has emerged as the dominant paradigm for training tool-augmented multimodal agents, enabling models to iteratively invoke external tools—such as text search, image search, and code interpreters—and integrate tool outputs into coherent multi-step reasoning chains Zheng et al. (2025); Wu et al. (2025); Hong et al. (2025); Chng et al. (2025); Dong et al. (2025). Within this framework, Group Relative Policy Optimization (GRPO) Shao et al. (2024) has become the de facto training algorithm, owing to its elimination of a separate value network and its strong empirical performance on outcome-supervised tasks.

GRPO uniformly broadcasts the trajectory-level advantage to every token, implicitly assuming equal contribution from each generation step—an assumption that systematically breaks down in multi-step tool-use scenarios where different tool use contribute different informational value.

Credit Misassignment. As illustrated in Figure 1, two trajectories make identical tool calls but produce different final answers. In this example, GRPO assigns positive advantages to the parameter-matched Actions 1 and 2 in the successful trajectory and negative advantages to their counterparts in the failing trajectory. Thus, calls with the same information-seeking intent receive opposing trajectory-level signals within a single training update. However, an incorrect final answer does not imply that every preceding tool call was unhelpful. We refer to the potential over-penalization of such potentially useful calls in failing trajectories by uniformly assigned negative advantages as *credit misassignment*.

Prevalence of Parameter-Matched Calls. At $\theta \approx 0.5$, over half of failing trajectories contain a tool call matched by its parameters to a call in a successful trajectory for the same question, and over half of failing tool calls have such a match (Figure 2(a,b)). Higher parameter similarity is associated

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.

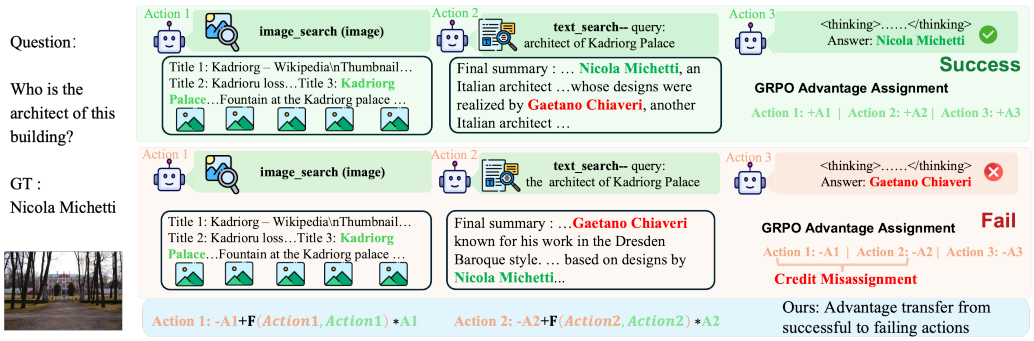

Figure 1: A concrete example of credit misassignment in group-based RL algorithms.

with greater coverage across successful trajectories (Figure 2(c)). A separate evaluation further finds that selected failing calls more often return information needed to answer the question (79.4% vs. 27.5%; §4.4). Together, these findings suggest that failing trajectories contain a substantial pool of structurally identifiable signals that may be useful for credit compensation.

A natural remedy is fine-grained credit assignment, but existing methods are all prohibitively costly Lightman et al. (2023); Wang et al. (2024); Luo et al. (2024); Zhang (2026): manual annotation, dedicated PRM training, and Monte Carlo continuation sampling are expensive and none guarantees accurate credit assignment Zhang (2026). Critically, all of them treat tool-use steps as ordinary generation steps, overlooking a structural property unique to tool calls that makes zero-cost, approximate credit correction possible *without any additional cost*.

Key Insight: Intent Consistency of Parameter-Matched Calls. For information-acquisition tools, the parameters specify the evidence-seeking target: the image under search, the region to inspect, or the textual query. Across trajectories for the same question, calls with matching parameters therefore exhibit consistent information-seeking intent even if tool responses and subsequent reasoning differ. This structural correspondence gives successful calls a useful, batch-internal reference role: when a failing call pursues the same evidence target, its uniform negative advantage may be overly punitive. TACT exploits that signal to attenuate negative advantages on selected failing calls.

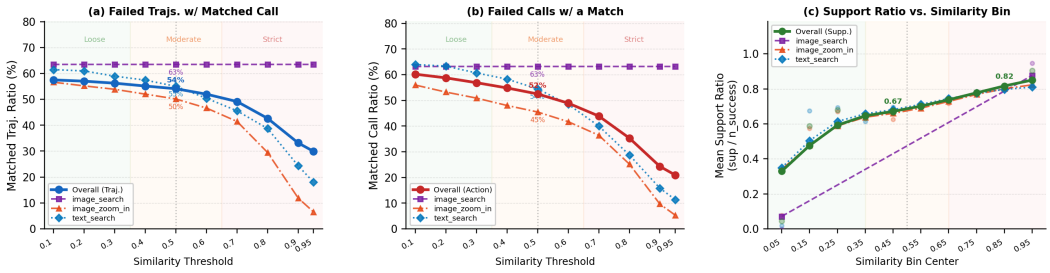

Figure 2: Parameter matching in failing trajectories. A failing call is counted as matched if a successful trajectory for the same question contains a call of the same tool type with parameter similarity at least θ. (a) Fraction of failing trajectories with at least one matched call. (b) Fraction of failing tool calls with a match. (c) Mean successful-trajectory coverage by similarity bin.

Our Method: TACT. Building on this insight, we propose **Tool-Aware Credit Transfer (TACT)**, a plug-and-play advantage-correction mechanism, which constructs a counterfactual reference library from successful trajectories in the current batch, uses a confidence score (parameter similarity $\times$ coverage) to gate credit transfer, and applies a conservative clamp ensuring no failing step is globally encouraged—requiring no additional models, annotation, or sampling, and integrating as a drop-in replacement for the advantage assignment step of any group-based RL algorithm. TACT

naturally achieves computational efficiency, difficulty adaptivity, and safe degradation to standard GRPO, and is instantiated on three tool types: image search, region zoom-in, and text search.

Our main contributions are as follows:

- We identify and formally characterize **credit misassignment** as a systematic failure mode of GRPO in tool-augmented reasoning, and empirically quantify its prevalence and scale.
- We propose **TACT**, a plug-and-play advantage-compensation mechanism that leverages the intent consistency of parameter-matched calls to compensate misassigned negative credit at zero marginal cost, requiring no additional models, annotation, or sampling.
- TACT consistently improves performance across multiple search benchmarks and three RL algorithms with negligible computational overhead.

2 Related Work

Multimodal Search Agents. To mitigate the limitations of static knowledge bases, large language models have evolved from static retrieval mechanisms into dynamic tool-augmented reasoning systems Chng et al. (2025). In the multimodal search domain, MMSearch-R1 Wu et al. (2025) and WebWatcher Geng et al. (2025) build multimodal search agents by designing image search and text search tools. DeepEyesV2 Hong et al. (2025) and SenseNova-MARS Chng et al. (2025) further introduce image processing tools to enable fine-grained multimodal search and understanding. Vision-DeepSearch Zeng et al. (2026) and MTA-Agent Peng et al. (2026) further advance agent performance through more sophisticated data construction and training pipelines. However, these multimodal search methods focus exclusively on improving overall performance, overlooking the intrinsic credit assignment problem inherent in multi-step tool use.

Fine-Grained Credit Assignment in Agentic RL. PRM-based methods Lightman et al. (2023); Wang et al. (2024) and step-level sampling approaches Luo et al. (2024); Yue et al. (2026) can provide step-wise reward signals, but at prohibitive cost and without guaranteeing accuracy. More recently, lighter-weight alternatives have emerged: GiGPO Feng et al. (2025) constructs step-level rewards via state grouping; HCAPO Tan et al. (2026) and Belief-RL Auzina et al. (2026) assess step quality through hindsight probability shifts and belief estimation, respectively. However, these methods predominantly target tool-integrated reasoning in the text domain. Critically, these methods all treat tool-use steps as ordinary generation steps, overlooking the *intent consistency of parameter-matched calls* in information-acquisition tools: similar call parameters can correspond to actions with similar information-seeking intent, making it possible to compare action credit across successful and failing trajectories.

3 Method

3.1 Problem Formulation

Tool-augmented reasoning trajectories. Let a multimodal tool-augmented reasoning trajectory be:

$$\tau^j = \left(o_1^j,\ r_1^j,\ o_2^j,\ r_2^j,\ \ldots,\ o_K^j,\ r_K^j,\ o_{\text{ans}}^j\right)$$

where o_k^j denotes the k-th reasoning and tool-use token segment, r_k^j the corresponding tool-response segment, and o_{ans}^j the final answer segment; $\text{args}(o_k^j)$ and $\text{type}(o_k^j)$ denote the call arguments and tool type of o_k^j, respectively. GRPO computes group-normalized advantages from within-group rewards and broadcasts them to every response token in τ^j; the policy gradient objective follows the standard PPO-clip form. We distinguish successful and failing trajectories by final-answer correctness and consider credit transfer only for tool-use steps in failing trajectories with negative advantages.

3.2 Formal Analysis of Credit Misassignment

Group-relative credit discrepancy. TACT operates entirely on the group-normalized advantage $\hat{A}_j = (R_j - \mu)/(\sigma + \varepsilon)$ that GRPO broadcasts to every token in trajectory τ^j; it does not estimate the

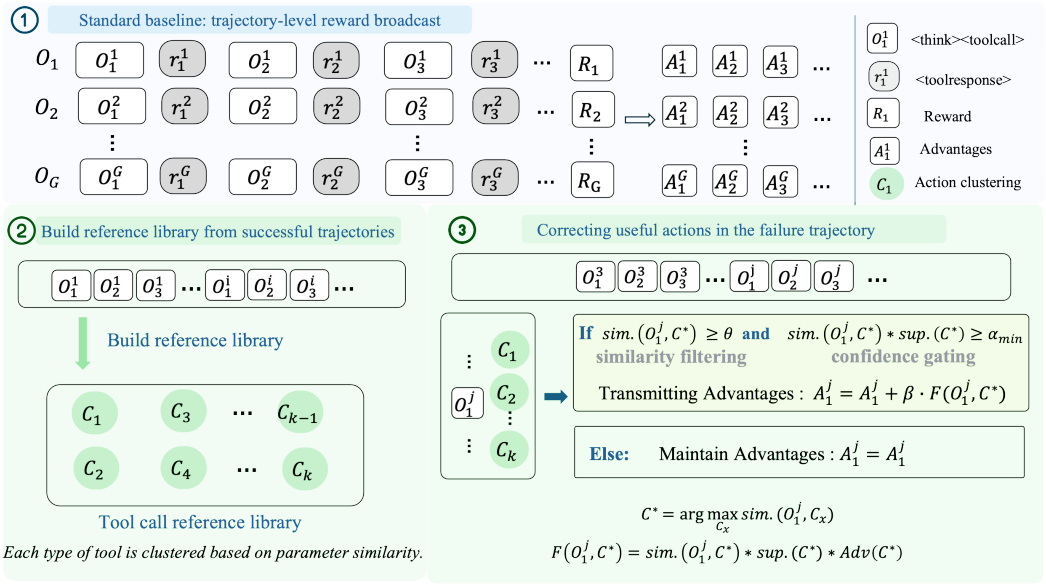

Figure 3: Overview of TACT. ① Standard GRPO broadcasts trajectory-level advantages uniformly to all tokens. ② Tool-use steps from successful trajectories are clustered into reference groups. ③ Each failing tool-use step is matched to the most similar reference group, and advantage transfer is applied via a two-level gate (θ and $\alpha_{\min}$).

true state-action value Q^π or the policy advantage A^π. Let C_k denote the class of same-question, same-tool-type calls matched to o_k^j by parameter similarity, and let $Y \in \{0,1\}$ denote trajectory correctness, with

$$p_k^+ \triangleq \Pr(Y=1 \mid o \in C_k)$$

the success probability of trajectories containing a call in this parameter-matched class (defined over trajectories, not within a single trajectory). By the law of total expectation, the group-relative signal aggregated over C_k decomposes as

$$\mathbb{E}\big[\hat{A}_j \,|\, o{\in}C_k\big] = p_k^+ \,\mathbb{E}\big[\hat{A}_j \,|\, o{\in}C_k,\, Y{=}1\big] + (1{-}p_k^+)\,\mathbb{E}\big[\hat{A}_j \,|\, o{\in}C_k,\, Y{=}0\big]. \tag{1}$$

Relative to the conditional signal $\mathbb{E}[\hat{A}_j \mid o \in C_k,\, Y{=}0]$ that GRPO actually broadcasts to a failing step in class C_k, the *group-relative credit discrepancy* is

$$\Delta_k^{\text{credit}} \triangleq \mathbb{E}\big[\hat{A}_j \,|\, o{\in}C_k\big] - \mathbb{E}\big[\hat{A}_j \,|\, o{\in}C_k,\, Y{=}0\big] = p_k^+\Big(\mathbb{E}\big[\hat{A}_j \,|\, o{\in}C_k,\, Y{=}1\big] - \mathbb{E}\big[\hat{A}_j \,|\, o{\in}C_k,\, Y{=}0\big]\Big). \tag{2}$$

This rewriting only compares different conditional expectations of the *same* empirical signal $\hat{A}_j$, avoiding any subtraction between Q^π, A^π, and $\hat{A}$. The objective of TACT is to mitigate the group-relative credit discrepancy Δ_k^{credit} for each failing tool-use step, without introducing any annotation or model.

3.3 Parameter-Matched Intent and Credit Transfer

Intent Consistency of Parameter-Matched Calls. For information-acquisition tools, if the parameters of two calls satisfy the similarity threshold $\text{sim}(\text{args}(o_k^j), \text{args}(o_{k'}^i)) \geq \theta$, the two calls correspond to similar *information-acquisition actions*: even when the execution environment is noisy, they target similar evidence needs and can therefore carry comparable action-level informational value. Intent consistency here requires that the information sought by the action is primarily specified by the call parameters, rather than by the surrounding trajectory context. This property naturally extends to tools where fixed parameters specify the underlying information-acquisition operation, and is concretely instantiated for the three tool types used in this work (see §C for details).

Counterfactual Witness. If a failing tool call o_k^j matches a call in a successful trajectory τ^i, then τ^i serves as a **counterfactual witness** that the matched information-seeking intent can support a correct answer. Uniformly penalizing the failing call may therefore misassign credit and warrants compensation. Matching relies only on tool arguments in the current batch, requiring no additional sampling, models, or annotation and providing approximate process supervision at zero marginal cost.

3.4 Confidence-Gated Credit Transfer

Reference construction and matching. For each question, we cluster calls from successful trajectories by tool type and parameter similarity into reference groups $C_1, C_2, \ldots, C_k$ (Figure 3 ②). Each group's reference advantage $\mathrm{Adv}(C_x)$ is the mean GRPO-normalized trajectory advantage of its member calls, rather than the mean raw reward. For a failing step o_k^j, we select the group with the highest mean parameter similarity (Figure 3 ③):

$$C^* = \arg\max_{C_x} \ \text{sim.}(o_k^j,\ C_x), \qquad \text{sim.}(o_k^j, C_x) = \frac{1}{|C_x|} \sum_{o_{k'}^i \in C_x} \text{sim}(o_k^j,\ o_{k'}^i) \tag{3}$$

where sim measures parameter similarity. A failing call receives no transfer if its best-match similarity $\text{sim.}(o_k^j, C^*)$ falls below the matching threshold θ.

Credit transfer function. Given the selected C^*, the credit transfer function is:

$$F(o_k^j, C^*) = \underbrace{\text{sim.}(o_k^j, C^*)}_{\text{param. match}} \times \underbrace{\text{sup.}(C^*)}_{\text{coverage}} \times \mathrm{Adv}(C^*) \tag{4}$$

where $\text{sup.}(C^*) = n(C^*)/G^+$, with $n(C^*)$ the number of successful trajectories containing a step in C^* and G^+ the total number of successful trajectories. The confidence score $\alpha = \text{sim.} \times \text{sup.}$ jointly captures parameter match quality and signal coverage: higher similarity indicates more consistent information-seeking intent, while higher support indicates that the matched call appears across more successful trajectories, providing a more robust reference signal. Credit transfer requires both $\text{sim.}(o_k^j, C^*) \geq \theta$ and $\alpha(o_k^j, C^*) \geq \alpha_{\min}$: θ first filters by parameter similarity, and $\alpha_{\min}$ then gates on the joint confidence score.

3.5 Conservative Advantage Compensation

For each failing tool-use step o_k^j with $\alpha(o_k^j, C^*) \geq \alpha_{\min}$, the compensated advantage is:

$$A_k^j \leftarrow \min\Big(\hat{A}_j + \beta \cdot F(o_k^j, C^*),\ 0\Big) \tag{5}$$

where $\beta > 0$ is the transfer coefficient; otherwise $A_k^j \leftarrow \hat{A}_j$. Clipping at 0 prevents any tool step in a failing trajectory from receiving positive advantage, preserving the overall optimization direction. TACT therefore requires no modification to the optimizer or sampling procedure and serves as a drop-in replacement for GRPO's advantage assignment. It also provides two benefits: difficulty adaptivity, since on hard questions ($G^+ \ll G$), rare successes yield larger normalized advantages and stronger credit transfer, while on easy questions ($\hat{A}_{\text{success}} \approx 0$), TACT reduces to standard GRPO; and entropy preservation, since partially canceling negative gradients on tool calls with matched information-seeking intent sustains exploration and mitigates entropy collapse, consistent with the entropy trends observed across all three RL algorithms (Figure 4).

4 Experiments

4.1 Implementation Details

Tool Setup. Following the setup of SenseNova-MARS Chng et al. (2025), we instantiate three representative tool types: (1) *Image search* performs reverse image search and returns image captions; (2) *Text search* executes a query-specified web search and returns a Qwen3-32B-generated summary; (3) *Region zoom-in* crops a specified bounding-box region and returns the patch. Parameter similarity definitions for each tool type and a per-type training analysis are provided in Appendix C.3 and D.4.

Type	Model	Average	MMSearch	HR-MMSearch	FVQA-test	InfoSeek	SimpleVQA	LiveVQA	MAT-Search
				Direct Answer					
Open-source	Qwen2.5-VL-7B-Instruct Bai et al. (2025)	27.70	7.60	0.58	26.28	31.95	47.88	19.63	60.00
	Qwen3-VL-8B-Instruct Team (2025)	29.24	11.70	12.13	24.22	23.15	42.94	23.18	67.33
	Qwen2.5-VL-32B-Instruct Bai et al. (2025)	32.01	11.70	3.93	30.50	36.65	48.57	21.40	71.33
	Qwen3-VL-32B-Instruct Team (2025)	35.22	16.96	19.02	32.17	28.95	45.90	31.59	72.67
Proprietary	GPT-4o-mini Hurst et al. (2024)	33.08	15.79	1.31	36.83	35.95	44.42	24.63	72.66
	Gemini-2.5-Flash Comanici et al. (2025)	40.87	21.64	7.54	43.78	44.10	55.48	31.57	82.00
	GPT-4o Hurst et al. (2024)	42.38	23.39	13.11	48.00	52.90	51.73	28.18	79.33
	GPT-5 OpenAI (2025)	50.24	35.09	22.62	54.39	54.15	61.70	44.39	79.33
	GPT-5.2 OpenAI (2025)	50.92	43.27	24.92	50.94	50.40	59.92	47.00	80.00
	Gemini-3-Flash Google (2025)	53.68	57.31	21.97	56.50	54.85	63.57	38.90	82.67
	Gemini-3-Pro DeepMind (2025)	55.87	62.57	26.89	59.22	56.30	64.07	40.06	82.00
				Agentic Model (zero-shot)					
Open-source	Qwen2.5-VL-7B-Instruct Bai et al. (2025)	35.50	32.16	19.34	36.00	28.80	42.35	22.52	67.33
	Qwen3-VL-8B-Instruct Team (2025)	51.33	47.37	27.87	53.61	46.15	62.29	39.37	82.67
	Qwen2.5-VL-32B-Instruct Bai et al. (2025)	53.45	49.71	33.44	52.22	50.10	65.15	42.17	81.33
	Qwen3-VL-32B-Instruct Team (2025)	53.82	49.12	34.43	54.28	49.85	64.17	42.87	82.00
Proprietary	GPT-4o-mini Hurst et al. (2024)	45.65	38.60	26.23	50.00	42.35	50.84	31.54	80.00
	GPT-4o Hurst et al. (2024)	55.09	49.12	30.16	66.34	59.55	63.67	40.09	76.67
	Gemini-2.5-Flash Comanici et al. (2025)	58.05	59.06	40.00	61.72	53.70	68.81	47.75	75.33
	GPT-5 OpenAI (2025)	60.12	52.63	38.36	62.61	55.95	70.58	56.02	84.67
	Gemini-3-Flash Google (2025)	61.26	62.57	41.64	64.89	61.10	67.92	48.06	82.67
	GPT-5.2 OpenAI (2025)	67.64	66.08	48.20	68.78	65.55	78.18	65.99	80.67
	Gemini-3-Pro DeepMind (2025)	69.06	74.27	48.52	72.61	66.45	75.91	59.69	86.00
				Agentic Model					
Open-source	Visual-ARFT Liu et al. (2025)	40.13	34.50	24.92	41.72	37.95	42.45	25.40	74.00
	DeepMMSearch-R1 Narayan et al. (2025)	–	–	–	–	47.51	55.87	–	–
	MMSearch-R1 Wu et al. (2025)	52.49	53.80	20.33	58.40	55.10	57.40	48.40	74.00
	DeepEyesV2 Hong et al. (2025)	–	63.70	–	60.60	51.10	59.40	–	–
	SkyWork-R1V4 Zhang et al. (2025)	–	66.10	–	67.20	–	–	–	–
	SenseNova-MARS Chng et al. (2025)	63.65	66.67	40.33	67.11	61.70	70.19	56.22	**83.33**
	Qwen3-VL-8B-Instruct+SAPO	62.48	64.33	38.69	65.61	59.55	70.38	56.86	82.00
	Qwen3-VL-8B-Instruct+SAPO + TACT	**65.21**	**67.21**	**40.66**	**69.89**	**62.70**	**73.64**	**59.71**	82.67

Table 1: Performance on search-oriented benchmarks under Direct Answer and Agentic Model workflows.

Method	MMSearch	HR-MMSearch	FVQA-test	SimpleVQA	LiveVQA	MAT-Search	InfoSeek	AVG
GRPO	53.22	29.51	62.17	67.11	51.42	78.00	52.45	56.27
+TACT	54.97	33.11	65.56	68.31	54.05	82.00	58.85	59.55
Δ	+3.30%	+12.21%	+5.46%	+1.78%	+5.12%	+5.13%	+12.20%	+5.83%
GSPO	53.22	31.15	66.50	68.21	53.55	80.67	56.40	58.53
+TACT	60.82	31.48	66.61	70.98	56.14	83.33	59.40	61.25
Δ	+14.28%	+1.05%	+0.17%	+4.06%	+4.82%	+3.31%	+5.32%	+4.65%
SAPO	53.97	30.16	64.50	67.52	51.75	81.33	55.10	57.76
+TACT	56.73	31.48	67.17	71.27	55.36	81.33	59.95	60.47
Δ	+5.10%	+4.35%	+4.13%	+5.56%	+6.97%	+0.00%	+8.80%	+4.69%

Table 2: TACT across base RL algorithms (Qwen3-VL-4B, $\beta = 0.25$). All values are accuracy (%) on seven multimodal search benchmarks. Δ denotes the relative improvement (%) of TACT over the corresponding base algorithm.

Training. We train with the VeRL framework Sheng et al. (2024) without any SFT warm-up, using a global batch size of 128, a learning rate of 1×10^{-6}, and $G = 8$ rollout trajectories per question on approximately 4K multimodal search samples. The reward function combines an accuracy reward and a format reward (see Appendix C.2 for details). To reduce training cost, text search during training is implemented via a local Wikipedia-based RAG system; full training configuration is provided in Appendix C.1.

Evaluation. We evaluate on seven multimodal search benchmarks (MMSearch, HR-MMSearch, FVQA-test, InfoSeek, SimpleVQA, LiveVQA, MAT-Search) using the Serper API; benchmark details are in Appendix C.4. We compare against top closed-source models (GPT-5, Gemini 3, in both direct-answer and agentic modes), leading open-source models (Qwen3-VL), and competitive search agents (DeepEyesV2 Hong et al. (2025), WebWatcher Geng et al. (2025), SenseNova-MARS Chng et al. (2025)).

β	MMSearch	HR-MMSearch	FVQA-test	SimpleVQA	LiveVQA	MAT-Search	InfoSeek	AVG
Qwen3-VL-4B	43.86	25.57	50.50	63.57	37.34	79.33	43.25	49.06
GRPO	53.22	29.51	62.17	67.11	51.42	78.00	52.45	56.27
TACT β=0.1	54.97	30.49	66.00	69.79	53.66	78.67	60.05	59.09
TACT β=0.25	54.97	33.11	65.56	68.31	54.05	82.00	58.85	59.55
TACT β=0.4	58.48	33.44	64.83	69.89	54.47	84.00	59.05	60.59
TACT β=0.6	54.97	34.10	68.28	71.17	55.89	80.00	61.70	60.87
TACT β=0.8	61.41	33.11	66.44	69.10	56.30	84.67	57.25	61.18
TACT β=1.0	57.31	30.49	66.00	69.79	53.66	80.67	60.05	59.71
TACT β=1.0 w/o sup.	52.63	20.66	59.22	67.72	51.83	81.33	51.60	55.00
TACT β=1.0 w/o clip.	56.73	29.84	65.00	69.10	54.50	80.00	58.30	59.07
Δ vs. Qwen3-VL-4B	+40.01%	+33.33%	+35.20%	+11.96%	+50.78%	+6.73%	+42.66%	+24.71%
Δ vs. GRPO	+15.40%	+15.56%	+9.83%	+6.05%	+9.50%	+8.55%	+17.64%	+8.74%

Table 3: Ablation on the transfer coefficient β (GRPO+TACT). All values are accuracy (%) on seven multimodal search benchmarks. Δ vs. Qwen3-VL-4B: improvement of the best-β result over the Qwen3-VL-4B base model; Δ vs. GRPO: improvement over vanilla GRPO training.

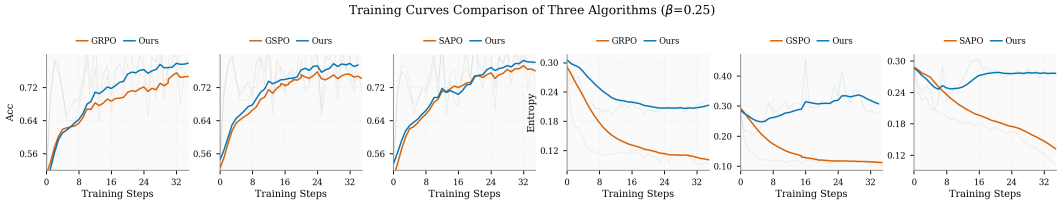

Figure 4: Ablation on generality across RL algorithms.

4.2 Main Results

As shown in Table 1, TACT achieves consistent improvements across all multimodal search benchmarks. We adopt SAPO as the RL backbone for our Qwen3-VL-8B main experiments; it raises the base model's average accuracy from 51.33% to 62.48%. TACT further improves upon this strong baseline by **4.4%** relatively, demonstrating the additional headroom unlocked by addressing credit misassignment alongside RL training. Compared to the strong search agent baseline SenseNova-MARS, which shares the same data and tool configuration, TACT further achieves a **2.5%** relative improvement. Notably, these gains stem entirely from applying TACT on top of the RL backbone—a pure advantage-correction mechanism—requiring no higher-quality data, additional tools, or increased model capacity. Moreover, the computational overhead is minimal: building the reference library and propagating advantages accounts for only **0.06%** of total training time (under 1 second per step). This highlights the effectiveness of correcting credit misassignment as an orthogonal, lightweight optimization that integrates seamlessly into any group-based RL algorithm.

4.3 Ablation Study

Generalization across RL algorithms. To verify that TACT is algorithm-agnostic, we apply it to three representative RL algorithms: GRPO, GSPO, and SAPO. All experiments use Qwen3-VL-4B for one training epoch, with the transfer coefficient fixed at $\beta = 0.25$. Training curves and evaluation results are shown in Figure 4 and Table 2. As shown in the upper row of Figure 4, TACT consistently improves training accuracy over the baselines for all three algorithms. The lower row further shows that TACT mitigates entropy collapse across all three algorithms, preserving exploration. Quantitatively, TACT improves average accuracy by 4–6% relative to the corresponding baselines across the three algorithms (Table 2), confirming its generality as a plug-and-play component.

Generality on Text-Only Search. As a plug-and-play mechanism, TACT applies to group-based RL pipelines where tool parameters specify the information sought. We test this extension in *text-only* search, using the same training configuration as GiGPO Feng et al. (2025) for all four backbones. This setting realizes the extension anticipated in §3.3 and separates the mechanism from multimodal-specific structures, since all transfer signals come from the least deterministic of the three tool types used in our multimodal search setting. Table 4 shows that TACT improves the av-

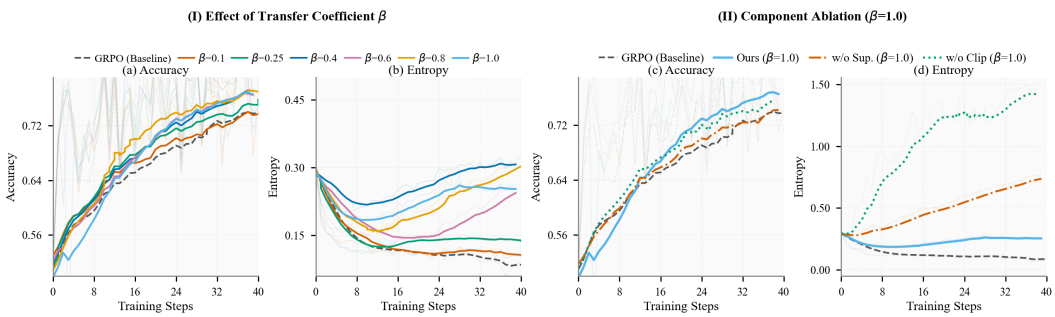

Figure 5: Robustness of TACT to the transfer coefficient β.

erage accuracy of all four backbones, including GiGPO Feng et al. (2025), which assigns step-level credit by grouping repeated states. The gain on GiGPO shows that TACT is also effective when applied to a fine-grained credit assignment method.

Method	Single-Hop QA			Multi-Hop QA				AVG
	NQ	TriviaQA	PopQA	HotpotQA	2WikiMultiHopQA	MuSiQue	Bamboogle	
GRPO	47.05	63.72	46.98	43.97	44.65	16.15	42.40	43.56
+TACT	47.98	65.00	47.81	44.05	44.02	17.38	46.40	44.66
GSPO	48.70	64.50	47.42	45.01	42.75	18.26	44.00	44.38
+TACT	50.27	65.23	50.82	49.28	50.27	19.08	50.40	47.91
SAPO	47.44	63.05	47.53	41.15	40.99	16.44	42.40	42.71
+TACT	47.50	65.31	48.19	47.88	47.99	20.65	43.20	45.82
GiGPO	47.20	64.13	47.95	41.89	43.29	16.26	43.20	43.42
+TACT	47.95	65.42	48.20	44.47	45.38	18.04	46.40	45.12

Table 4: Accuracy (%) of Qwen3-4B on single-hop and multi-hop text-only QA benchmarks. AVG is the unweighted mean accuracy across the seven benchmarks.

Robustness to the transfer coefficient β. We sweep $\beta \in \{0.1, 0.25, 0.4, 0.6, 0.8, 1.0\}$ to test TACT's sensitivity. Results are shown in Figure 5 and Table 3; detailed dynamics for β=0.25 are deferred to Appendix D.3. At small β (e.g., 0.1), transfer is weak and performance is close to GRPO (+2.82% AVG). As β increases, TACT improves and peaks at $\beta = 0.8$. TACT outperforms GRPO across all tested values, indicating robustness to β.

We further ablate the two key components of TACT at $\beta = 1.0$: removing the support factor (*w/o sup.*) and removing the conservative clamp (*w/o clip.*). Removing support degrades AVG from 59.71% to 55.00% (-4.71%), confirming that the support factor is essential for filtering unreliable signals from small reference groups. Removing the clamp causes notable entropy divergence during training (Figure 5, right), as some tool-use steps in failing trajectories receive positive advantages, destabilizing the policy update; the final AVG (59.07%) also falls below the full model, confirming that the upper-bound constraint is necessary for both training stability and final performance.

4.4 Transfer Precision

A natural concern is that the transfer gates select steps by parameter similarity alone, so transferred credit might reward coincidental parameter matches rather than genuinely informative tool calls. To rule this out, we directly measure *transfer precision*: the fraction of transfer-triggered tool calls whose responses actually contain the information necessary to answer the question, evaluated with a criterion fully decoupled from parameter similarity. Specifically, the question-specific necessary information is first extracted from the reasoning paths of successful trajectories on the same question, and an independent LLM judge then assesses each tool response against this reference.

As shown in Table 5, 79.4% of the 2,378 transfer-triggered steps contain question-necessary information, compared with only 27.5% for 2,375 failing steps that do not trigger transfer—the two-level

	Human (subset)	LLM judge (subset)	LLM judge (full)
Transfer-triggered	83.0% (265)	76.2% (265)	**79.4%** (2,378)
Non-triggered	—	—	**27.5%** (2,375)

Table 5: Transfer precision: fraction of tool responses containing question-necessary information, evaluated independently of parameter similarity. Parentheses indicate the number of tool-use steps; the LLM judge agrees with human labels at F1 = 0.87.

gates select primarily on informational value rather than surface parameter overlap. Human annotation on a 265-step subset yields 83.0%, with the LLM judge agreeing with human labels at F1 = 0.87, confirming that the large-scale estimate is reliable, if slightly conservative.

4.5 Training Dynamics Analysis

Advantage compensation for credit misassignment not only improves final benchmark performance but also yields observable behavioral differences throughout training (see Appendix D.2 for an analysis of why credit misassignment persists during training). Figure 6 compares the training dynamics of GRPO with and without TACT along four dimensions.

Training accuracy and entropy. As shown in Figure 6(a), TACT achieves higher training accuracy than GRPO throughout training. Figures 6(b) and 5(b) show a steady decline in GRPO's policy entropy, reflecting diminished exploration. By compensating misassigned negative advantages on informative tool-use steps, TACT alleviates excessive penalization, supports beneficial exploration, and helps prevent premature convergence to low-diversity behavior. *Response length and tool usage.* Figures 6(c) and (d) show that response length and the number of tool calls per rollout decline throughout GRPO training. TACT counteracts this trend, with both measures exhibiting more varied dynamics. Together, these observations indicate that credit misassignment persistently suppresses tool use during training and that TACT mitigates this effect through advantage compensation.

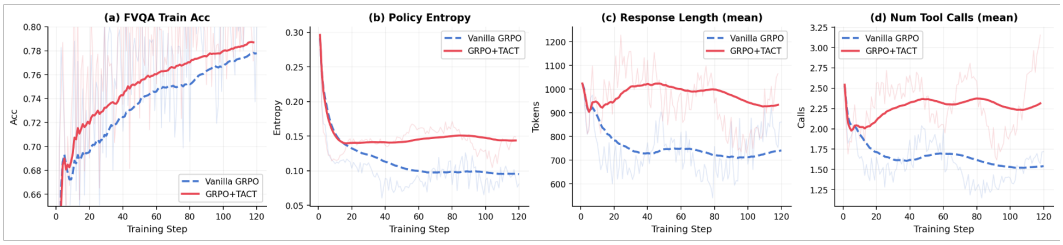

Figure 6: Training dynamics of Vanilla GRPO vs. GRPO+TACT (β=0.25) on Qwen3-VL-4B.

5 Conclusion

We identify and quantify credit misassignment as a systematic failure mode of GRPO in search settings, and propose TACT as a lightweight remedy. By exploiting the intent consistency of parameter-matched calls to information-acquisition tools, TACT compensates misassigned advantages through confidence-gated credit transfer, requiring no additional annotation, models, or sampling while introducing only negligible computational overhead. Across multiple multimodal search benchmarks, TACT consistently improves over strong baselines, remains robust to β, is algorithm-agnostic, and effectively mitigates entropy collapse and tool-use degradation during training. TACT is orthogonal to existing multimodal search agent designs and can be seamlessly integrated into any group-based RL training pipeline equipped with information-acquisition tools whose parameters specify the information sought. More broadly, intent consistency of parameter-matched calls may extend beyond the three validated tools to other tool-augmented settings where call parameters specify information-acquisition actions.

AI USE STATEMENT

Generative AI tools were used to assist with English-language editing and polishing of the manuscript. LLMs were also used as part of the research methodology: Qwen3-VL-32B-Instruct judged semantic correctness for non-multiple-choice training rewards and open-ended evaluation (Appendix C.2 and C.4), while an independent GLM-5.2 judge assessed whether tool responses contained information needed to answer the question (Appendix D.1). To assess the latter judge, the authors annotated a subset of 265 tool-use steps; the judge achieved an F1 score of 0.87 against these annotations. The authors take responsibility for the manuscript's content, experimental design, analysis, and conclusions.

ETHICS STATEMENT

This work studies credit assignment for agents that retrieve external information. The 265-step subset used to assess the transfer-precision judge was annotated by the authors, without recruiting external annotators. The experiments use existing search benchmarks. Retrieved content may be inaccurate or sensitive, and LLM-based judgments may introduce evaluation bias. We describe the evaluation protocol and compare the transfer-precision judge with author annotations to make this source of uncertainty more transparent (Appendix D.1).

REPRODUCIBILITY STATEMENT

To support reproducibility, Appendix C documents the training datasets and model initialization; optimization, rollout, and reward settings; and tool-parameter similarity measures and transfer thresholds. Appendix C.4 describes benchmark selection and evaluation protocols, with multimodal benchmark results reported as mean pass@1 accuracy over three independent runs. Appendix D.1 details the transfer-precision assessment and its comparison with author annotations. As stated in the abstract, the code and models are planned for public release upon acceptance.

REFERENCES

Ilze Amanda Auzina, Joschka Strüber, Sergio Hernández-Gutiérrez, Shashwat Goel, Ameya Prabhu, and Matthias Bethge. Intrinsic credit assignment for long horizon interaction, 2026. URL https://arxiv.org/abs/2602.12342.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Jiajun Chai, Guojun Yin, Zekun Xu, Chuhuai Yue, Yi Jia, Siyu Xia, Xiaohan Wang, Jiwen Jiang, Xiaoguang Li, Chengqi Dong, et al. Rlfactory: A plug-and-play reinforcement learning post-training framework for llm multi-turn tool-use. *arXiv preprint arXiv:2509.06980*, 2025.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14948–14968, 2023.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, et al. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4637–4646, 2025.

Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, et al. Sensenova-mars: Empowering multimodal agentic reasoning and search via reinforcement learning. *arXiv preprint arXiv:2512.24330*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, 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.

Google DeepMind. Gemini-3-Pro. https://deepmind.google/models/gemini/pro/, 2025.

Chengqi Dong, Chuhuai Yue, Hang He, Rongge Mao, Fenghe Tang, S Kevin Zhou, Zekun Xu, Xiaohan Wang, Jiajun Chai, and Guojun Yin. Training multi-image vision agents via end2end reinforcement learning. *arXiv preprint arXiv:2512.08980*, 2025.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training, 2025. URL https://arxiv.org/abs/2505.10978.

Mingyang Fu, Yuyang Peng, Benlin Liu, Yao Wan, and Dongping Chen. Livevqa: Live visual knowledge seeking. *arXiv e-prints*, pp. arXiv–2504, 2025.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, et al. Webwatcher: Breaking new frontier of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*, 2025.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Google. Gemini-3-Flash. https://blog.google/products/gemini/gemini-3-flash/, 2025.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, et al. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*, 2024.

Xue Jiang, Yihong Dong, Mengyang Liu, Hongyi Deng, Tian Wang, Yongding Tao, Rongyu Cao, Binhua Li, Zhi Jin, Wenpin Jiao, et al. Coderl+: Improving code generation via reinforcement with execution semantics alignment. *arXiv preprint arXiv:2510.18471*, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual agentic reinforcement fine-tuning. *arXiv preprint arXiv:2505.14246*, 2025.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

Kartik Narayan, Yang Xu, Tian Cao, Kavya Nerella, Vishal M Patel, Navid Shiee, Peter Grasch, Chao Jia, Yinfei Yang, and Zhe Gan. Deepmmsearch-r1: Empowering multimodal llms in multimodal web search. *arXiv preprint arXiv:2510.12801*, 2025.

OpenAI. GPT-5. `https://openai.com/gpt-5`, 2025.

Xiangyu Peng, Can Qin, An Yan, Xinyi Yang, Zeyuan Chen, Ran Xu, and Chien-Sheng Wu. Mtagent: An open recipe for multimodal deep search agents, 2026. URL `https://arxiv.org/abs/2604.06376`.

Tuoshi Qi, Shenshen Bu, Yingfei Xiang, and Zhiming Dai. Eviagent: Evidence-driven agent for radiology report generation. *arXiv preprint arXiv:2603.13956*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Hui-Ze Tan, Xiao-Wen Yang, Hao Chen, Jie-Jing Shao, Yi Wen, Yuteng Shen, Weihong Luo, Xiku Du, Lan-Zhe Guo, and Yu-Feng Li. Hindsight credit assignment for long-horizon llm agents. *arXiv preprint arXiv:2603.08754*, 2026.

Qwen Team. Qwen3-vl, 2025. URL `https://qwen.ai/blog?id=qwen3-vl`.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, et al. Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7920–7939, 2025.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing lmms to search. *arXiv preprint arXiv:2506.20670*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Chuhuai Yue, Chengqi Dong, Yinan Gao, Hang He, Jiajun Chai, Wei Lin, and Guojun Yin. Promoting efficient reasoning with verifiable stepwise reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 41, pp. 34530–34538, 2026.

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, 2026. URL https://arxiv.org/abs/2602.02185.

Chenchen Zhang. From reasoning to agentic: Credit assignment in reinforcement learning for large language models. *arXiv preprint arXiv:2604.09459*, 2026.

Yifan Zhang, Liang Hu, Haofeng Sun, Peiyu Wang, Yichen Wei, Shukang Yin, Jiangbo Pei, Wei Shen, Peng Xia, Yi Peng, Tianyidan Xie, Eric Li, Yang Liu, Xuchen Song, and Yahui Zhou. Skywork-r1v4: Toward agentic multimodal intelligence through interleaved thinking with images and deepresearch, 2025. URL https://arxiv.org/abs/2512.02395.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

A Appendix

B Multimodal Search Agents

Reinforcement learning has emerged as the dominant paradigm for training LLM-based agents, achieving significant advances across diverse domains including mathematical reasoning Shao et al. (2024); Yue et al. (2026), code generation Jiang et al. (2025), and web navigation Wei et al. (2025). In tool-augmented settings, agents follow a think–act–observe loop Yao et al. (2022), progressively collecting information through multi-step interactions with external environments and achieving consistent performance gains across diverse application scenarios Dong et al. (2025); Qi et al. (2026); Wei et al. (2025). With outcome correctness as the reward signal, Agentic RL paradigm enables efficient end-to-end training, supported by dedicated agentic RL frameworks Sheng et al. (2024); Chai et al. (2025). Multimodal search is a representative scenario of tool-augmented reasoning, requiring models to answer open-domain questions over mixed image–text inputs. In this setting, agents built on multimodal foundation models such as Qwen2.5-VL Bai et al. (2025) invoke multiple rounds of image search and text retrieval in coordination, progressively integrating multimodal tool outputs into the reasoning chain to produce the final answer, achieving substantial performance gains across multiple open-domain visual question answering benchmarks Hong et al. (2025); Geng et al. (2025); Chng et al. (2025). In this work, we instantiate TACT within the multimodal search setting using three tool types: image search , text search, and region zoom-in, serving visual entity grounding, background knowledge retrieval, and fine-grained visual understanding, respectively.

C Implementation Details

C.1 Training Data

RL training uses the same training set as SenseNova-MARS Chng et al. (2025), comprising approximately 3,859 samples across four subsets:

Dataset	#Samples	Reward
FVQA-train	1,894	LLM-judge + format
DeepEyes-4K (non-MCQ)	425	LLM-judge + format
Visual-Probe-train	912	LLM-judge + format
DeepEyes-4K (MCQ)	628	Exact match + format

Table 6: Composition of the RL training set.

All subsets are used exclusively for RL training. No supervised fine-tuning (SFT) warm-up stage is applied; we train directly from the instruction-tuned Qwen3-VL-8B-Instruct checkpoint. Training uses the VeRL framework Sheng et al. (2024) with a global batch size of 128, a learning rate of 1×10^{-6}, and $G = 8$ rollout trajectories per question. Each trajectory allows up to 10 interaction rounds, with a per-round token budget of 8,192 and a cumulative trajectory limit of 32,768 tokens. During training, text search is implemented via a local Wikipedia-based RAG system to reduce computational cost.

C.2 Reward Function

The outcome reward combines two components:

Accuracy reward. For non-MCQ samples, we use Qwen3-VL-32B-Instruct as an LLM-as-a-Judge (greedy decoding, temperature = 0): if the model's final answer is judged semantically correct, $r_{\text{acc}} = 1.0$; otherwise $r_{\text{acc}} = 0.0$. For MCQ samples, rule-based exact match is applied, assigning $r_{\text{acc}} = 1.0$ for a correct choice and 0.0 otherwise.

Format reward. At each non-terminal turn, the response must contain a `<thinking>` block followed by a single valid `<tool_call>` in JSON Schema format. At the terminal turn, the response must contain a `<thinking>` block followed by an `<answer>` tag. Full compliance yields $r_{\text{format}} = 0.5$; any violation yields 0.0.

The final outcome reward is $R = r_{\text{acc}} + r_{\text{format}} \in \{0,\ 0.5,\ 1.0,\ 1.5\}$.

C.3 Parameter Similarity Computation

TACT defines parameter similarity separately for each tool type. All similarity functions are rule-based and require no embedding models or external LLMs.

Image search (`image_search_tool`). This tool takes no text parameters; all image search calls for a given question share the same input image. Parameter equivalence is therefore trivial: within each training batch, all image search calls for the same question are merged into a single reference group, and parameter similarity is identically 1.0. Since image search has no free parameters, the confidence score α is determined solely by the support factor. We set $\alpha_{\min,\text{isearch}} = 1.0$, meaning advantage transfer is triggered only when the reference group covers all successful trajectories in the batch (i.e., support factor $= 1.0$).

Region zoom-in (`image_zoom_in_tool`). Parameter equivalence is measured by bounding-box IoU. During reference library construction (Phase 3), zoom-in calls with IoU $\geq \theta_{\text{IoU}} = 0.7$ are merged into the same reference group. During advantage transfer (Phase 4), the minimum confidence threshold is $\alpha_{\min,\text{zoom}} = 0.5$.

Text search (`text_search_tool`). Parameter equivalence is measured by query string similarity, computed as a weighted combination of three lexical metrics:

$$\text{sim}(q_1, q_2) = 0.3 \cdot J + 0.5 \cdot \text{OC} + 0.2 \cdot \text{BG} \tag{6}$$

where J is the Jaccard coefficient over stopword-filtered tokens, OC is the Overlap Coefficient (robust to asymmetric query lengths), and BG is character-bigram Jaccard similarity (capturing morphological variants). Reference library construction uses aggregation threshold $\theta_{\text{text}} = 0.8$; Phase 4 matching uses $\alpha_{\min,\text{text}} = 0.5$.

C.4 Evaluation Benchmarks

We evaluate on seven publicly available multimodal search benchmarks that collectively span a broad range of retrieval difficulty, temporal freshness, and image resolution requirements.

MMSearch Jiang et al. (2024) contains 300 image-accompanied questions drawn from news events and domain knowledge published after August 2024, testing the model's ability to ground visual

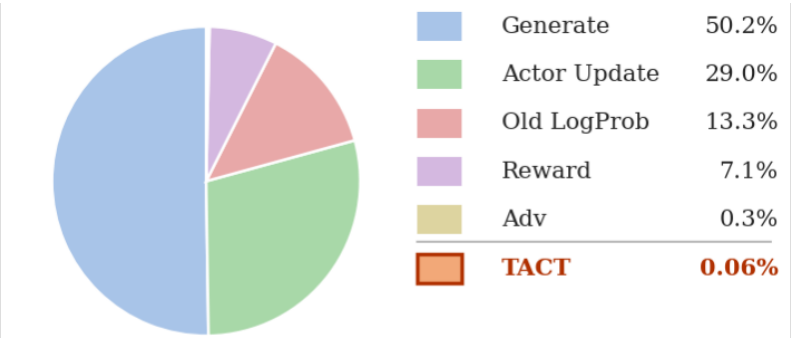

Figure 7: Wall-clock time breakdown per training step. TACT (Phase 3+4) accounts for only **0.06%**, negligible compared to the dominant components of token generation (50.2%) and actor update (29.0%).

observations in up-to-date factual knowledge. We follow the standard split of 171 questions used for system-level evaluation.

HR-MMSearch Chng et al. (2025) is a knowledge-intensive benchmark comprising 305 questions paired with high-resolution (4K) images sourced from international news agencies (2025 events), spanning eight topical domains. Approximately 60% of questions are classified as hard, requiring three or more tool invocations to answer correctly.

FVQA-test Wu et al. (2025) is a curated 1,800-sample test set consisting of three equal-sized subsets: FVQA-auto-vc (600 samples with automatic validation), InfoSeek Human Split (600 samples with manually corrected annotations), and an expert-annotated subset (600 samples).

InfoSeek Chen et al. (2023) is a visual entity retrieval benchmark grounded in Wikidata, covering fine-grained factual attributes that are unlikely to be memorized during pre-training. We sample 2,000 instances from the official test split.

SimpleVQA Cheng et al. (2025) focuses on factual accuracy, combining post-2023 VQA instances with expert-curated image-question pairs across 1,013 English samples.

LiveVQA Fu et al. (2025) aggregates 3,602 image-question pairs from major news outlets (CNN, BBC, etc.), spanning 14 thematic categories, and emphasizes temporal reasoning over recently published content.

MAT-Search Liu et al. (2025) is a manually crafted and human-verified benchmark of 150 examples designed to evaluate agentic multimodal reasoning, requiring models to handle composite multi-hop queries, retrieve external information, and invoke tools effectively across varying reasoning depths.

All benchmarks are evaluated with pass@1 accuracy; web search is powered by the Serper API. For open-ended questions, Qwen3-VL-32B-Instruct serves as an LLM-as-a-Judge with greedy decoding (temperature $= 0$). For all benchmarks, we report the mean pass@1 accuracy over three independent runs.

C.5 Computational Overhead of TACT

TACT's Phase 3 (reference library construction) and Phase 4 (advantage transfer) run entirely on CPU using vectorized NumPy operations and lightweight string matching—no neural network forward passes are required. Under our training configuration (global batch size 128, $G = 8$ rollouts per question), these two phases together take under 1 second per training step, accounting for only **0.06%** of total wall-clock training time, as illustrated in Figure 7. TACT therefore introduces negligible computational overhead and can be seamlessly integrated into any group-based advantage estimation algorithm.

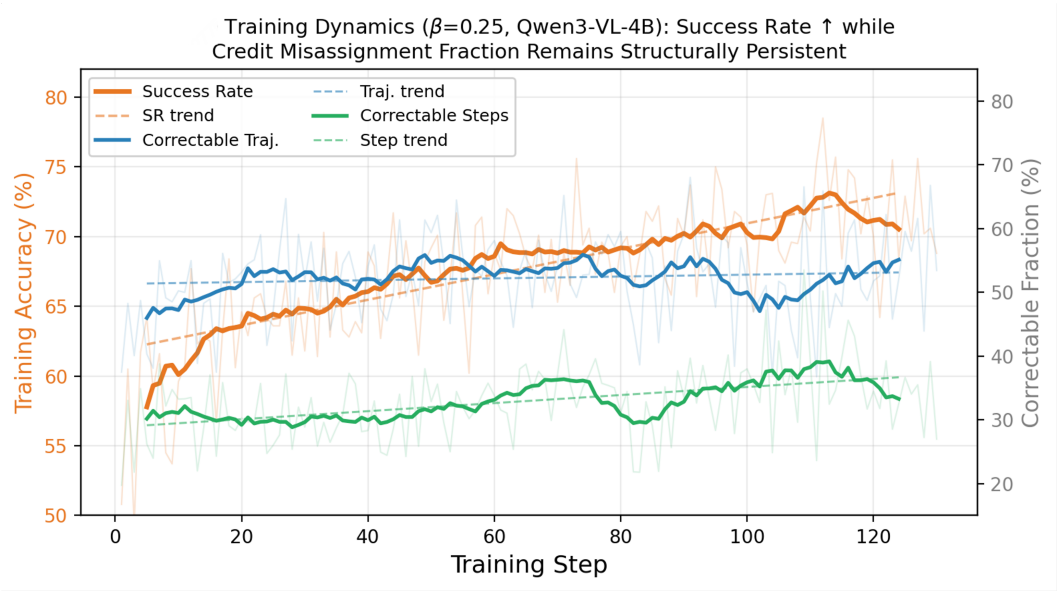

Figure 8: Training dynamics of GRPO+TACT (β=0.25, Qwen3-VL-4B). Training accuracy increases steadily throughout training, while the fraction of correctable trajectories (Correctable Traj.) and correctable steps (Correctable Steps) remain structurally stable and exhibit a slow upward trend, indicating that credit misassignment persists across the entire training horizon.

D Detailed Analysis of TACT

D.1 Transfer Precision Evaluation Protocol

For each question, we extract the set of facts required to reach the correct answer (entity names, attributes, and relations) from the reasoning paths of successful trajectories on the same question, yielding a necessary-information reference that is independent of the parameter-similarity criterion used by the transfer gates. An independent LLM judge, GLM-5.2 GLM-5-Team et al. (2026) (not involved in training), then receives the question, this reference, and one tool response, and labels whether the response contains the necessary information in full, in part, or not at all; a step is counted as precision-positive if the information is present in full or in part, a lenient criterion consistent with the semantics of credit transfer. The judge never sees parameter similarities, confidence scores, or transfer decisions, and for image search and region zoom-in responses it additionally examines the retrieved image's caption and metadata. The evaluation covers a human-annotated subset of 265 transfer-triggered steps and, at full scale, 2,378 transfer-triggered steps plus a control cohort of 2,375 steps from failing trajectories that do not trigger transfer. On the human subset, the LLM judge agrees with human labels at F1 $= 0.87$, indicating that the large-scale estimate is reliable.

D.2 Persistence of Credit Misassignment During Training

As shown in Figure 8, TACT training leads to a continuous improvement in training accuracy, yet the fractions of correctable trajectories and correctable steps do not decline in tandem; instead, they remain structurally stable throughout training. This indicates that in multi-tool search tasks, the phenomenon of high-quality tool-use steps appearing within failing trajectories remains prevalent as training progresses. As the model learns more complex multi-step reasoning strategies, such "locally effective but globally failing" trajectories do not disappear naturally with improving model capability. Consequently, TACT's advantage compensation mechanism consistently faces meaningful correction targets throughout the entire training cycle, rather than being effective only in the early stages.



Figure 9: Six-dimensional training dynamics of GRPO+TACT (β=0.25, Qwen3-VL-4B). **(a)** Training accuracy; **(b)** Correctable trajectory fraction; **(c)** Correctable step fraction; **(d)** Mean credit transfer magnitude Δ; **(e)** Absolute advantage correction; **(f)** Relative advantage correction. Dashed lines indicate linear trend fits.

D.3 Detailed Training Dynamics of TACT

Figure 9 presents the complete training dynamics of GRPO+TACT at β=0.25. Training accuracy (panel a) exhibits a sustained upward trend; meanwhile, the correctable trajectory fraction (panel b) and correctable step fraction (panel c) also grow slowly, consistent with the analysis in Section D.2 and further confirming that the demand for compensation does not diminish during training.

The mean credit transfer magnitude Δ (panel d) remains broadly stable throughout training without systematic decay, indicating that the advantage correction applied at each compensation step is consistently sized. Both the absolute advantage correction (panel e) and the relative advantage correction (panel f) exhibit minimal drift, demonstrating that TACT's compensation strength adapts to maintain a stable level throughout training—neither growing excessively as training proceeds nor degrading as the success rate rises.

D.4 Per-Tool-Type Credit Transfer Analysis

Different tool types exhibit a pronounced and stable stratification pattern under TACT correction, which can be understood systematically along three dimensions.

Support factor stratification. As shown in the top row of Figure 10, image search achieves a median support factor close to 1.0, with the P25–P75 band highly concentrated near the upper bound. This property stems from the consistency of image search call parameters: for a given question, all image search steps share identical inputs, so a reference group covering all successful trajectories can almost always be found within the same training batch, maximally satisfying the support condition. By contrast, region zoom-in and text search exhibit much wider P25–P75 bands and lower mean values, reflecting higher diversity in call parameters across samples and correspondingly limited reference group coverage.

Stable stratification of confidence score α. The middle row shows that image search achieves a median $\alpha \approx 0.94$, concentrated in the high-value range, while region zoom-in ($\alpha \approx 0.05$–0.15) and text search ($\alpha \approx 0.15$–0.35) are substantially lower with wider spread. This three-tier stratification remains stable across all 130 training steps, demonstrating that TACT's confidence estimation mechanism reliably discriminates between tool types and does not drift with training state.

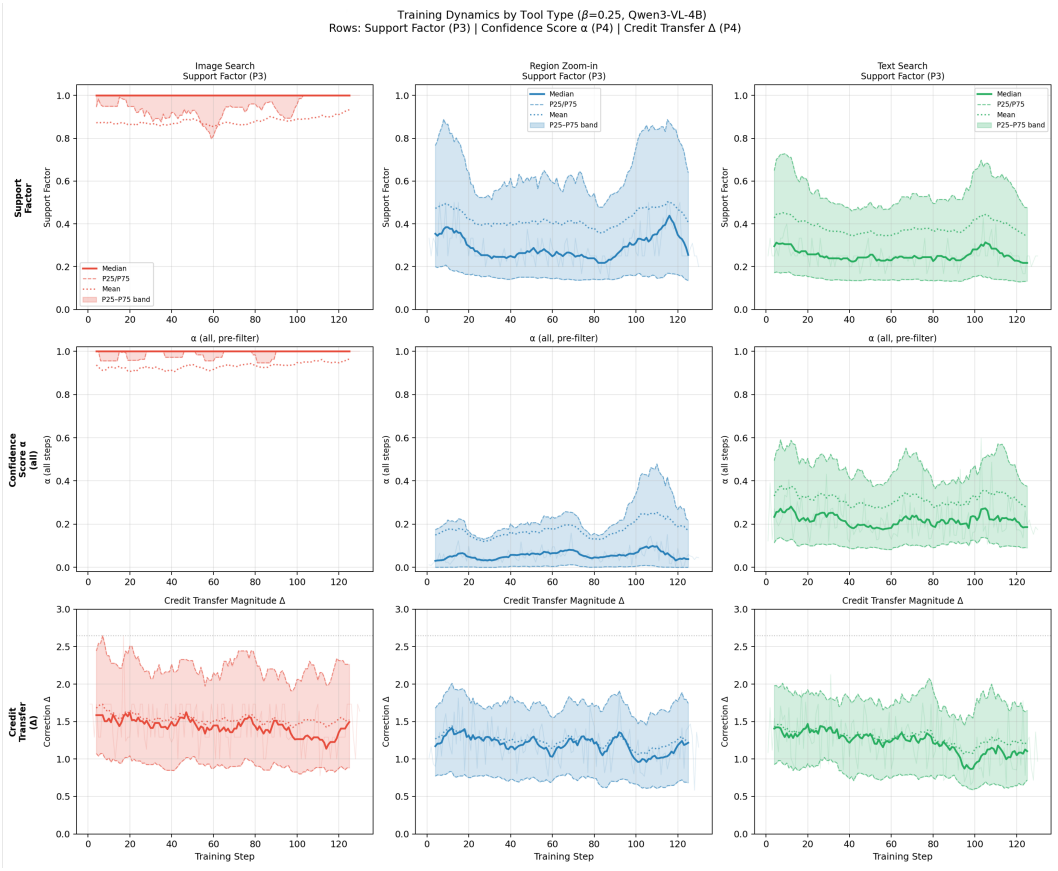

Figure 10: Per-tool-type training dynamics of GRPO+TACT (β=0.25, Qwen3-VL-4B) across 130 training steps. Columns correspond to three tool types (Image Search / Region Zoom-in / Text Search). Rows show the support factor (top), confidence score α (middle), and credit transfer magnitude Δ (bottom), each plotted as median, P25/P75 band, and mean.

Consistency between Δ and confidence stratification. The bottom row reveals that image search has the highest median $\Delta \approx 1.5$, while region zoom-in and text search are somewhat lower (approximately 1.0–1.2), with all three remaining stable throughout training. The inter-tool ranking of Δ is entirely consistent with the stratification of α, indicating that TACT's correction magnitude adaptively reflects signal quality rather than applying a uniform fixed correction to all tool steps: stronger corrections are applied to high-confidence image search steps, while more conservative compensation is maintained for lower-confidence text search steps with lower query similarity.

The consistency across all three metrics jointly validates the effectiveness of TACT's confidence-gated design: through the confidence gating mechanism, TACT achieves differentiated advantage compensation across tool types, actively correcting high-quality signals while avoiding over-intervention on less reliable ones.

E Case Study: Reasoning Patterns of TACT-Trained Model

To qualitatively illustrate the search strategies learned by TACT, we present four representative cases from the evaluation set, covering different reasoning patterns exhibited by the trained model. Each case shows the input question, the ground-truth answer, the sequence of tool calls invoked, and the model's reasoning trajectory.

Case 1: Entity Recognition + Knowledge Retrieval. The question asks: *"In which year did this park officially open?"* (ground truth: **1996**). The model performs a reverse image search that returns

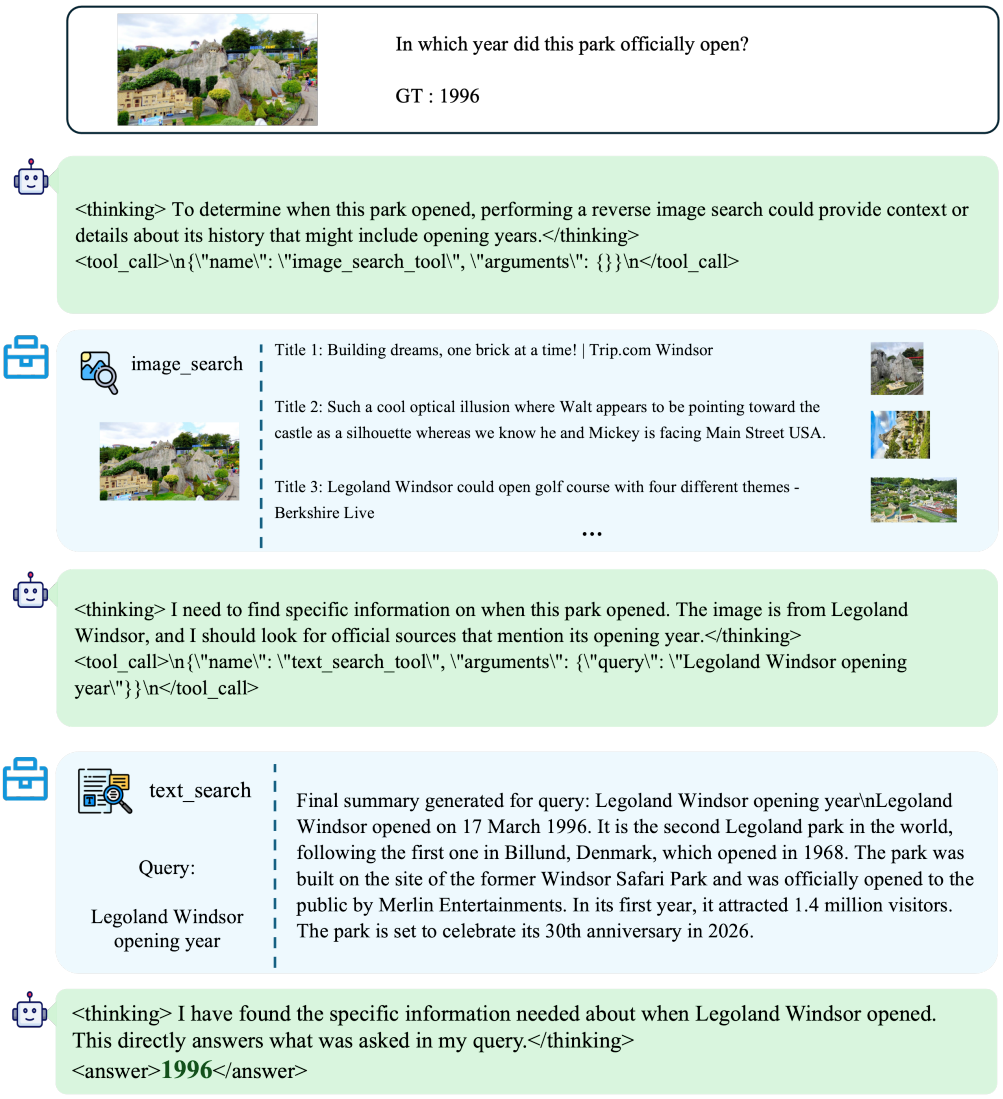

Figure 11: **Case 1: Entity Recognition + Knowledge Retrieval.** The model identifies *Legoland Windsor* via image search and retrieves its opening year with a single text query, correctly answering *1996*.

results including "Legoland Windsor could open golf course" and "LEGOLAND Windsor Resort", enabling it to identify the park. It then queries `text_search_tool` with *"Legoland Windsor opening year"*, obtaining the precise opening date of 17 March 1996. This two-step pattern—visual entity identification followed by a single factual lookup—is the most common successful strategy on knowledge-intensive benchmarks such as InfoSeek.

Case 2: Multi-Hop Iterative Query Refinement. The question asks: *"The founder of this company shown in the image was born in the same year as an Olympic Games. In which country were those Olympics held?"* (ground truth: **South Korea**). The model employs a four-tool chain spanning three reasoning hops: (1) `image_search_tool` returns CNBC headlines about *Macroscope*, identifying the company; (2) `text_search_tool`[*"Macroscope company founder"*] confirms the CEO as *Kayvon Beykpour*; (3) `text_search_tool`[*"Kayvon Beykpour birth year"*] retrieves birth year *1988*; (4) `text_search_tool`[*"1988 Summer Olympics location"*] returns *Seoul, South Korea*. Each intermediate result directly seeds the next query, bridging visual entity recognition, biographical lookup, and historical knowledge into a coherent reasoning chain.



Figure 12: **Case 2: Multi-Hop Iterative Query Refinement.** Starting from a group photo, the model chains four tool calls across three reasoning hops to arrive at *South Korea*.



Figure 13: **Case 3: Persistent Query Refinement for Numerical Retrieval.** The model refines its text query three times after receiving non-specific results, ultimately extracting the exact island width of *300 m*.

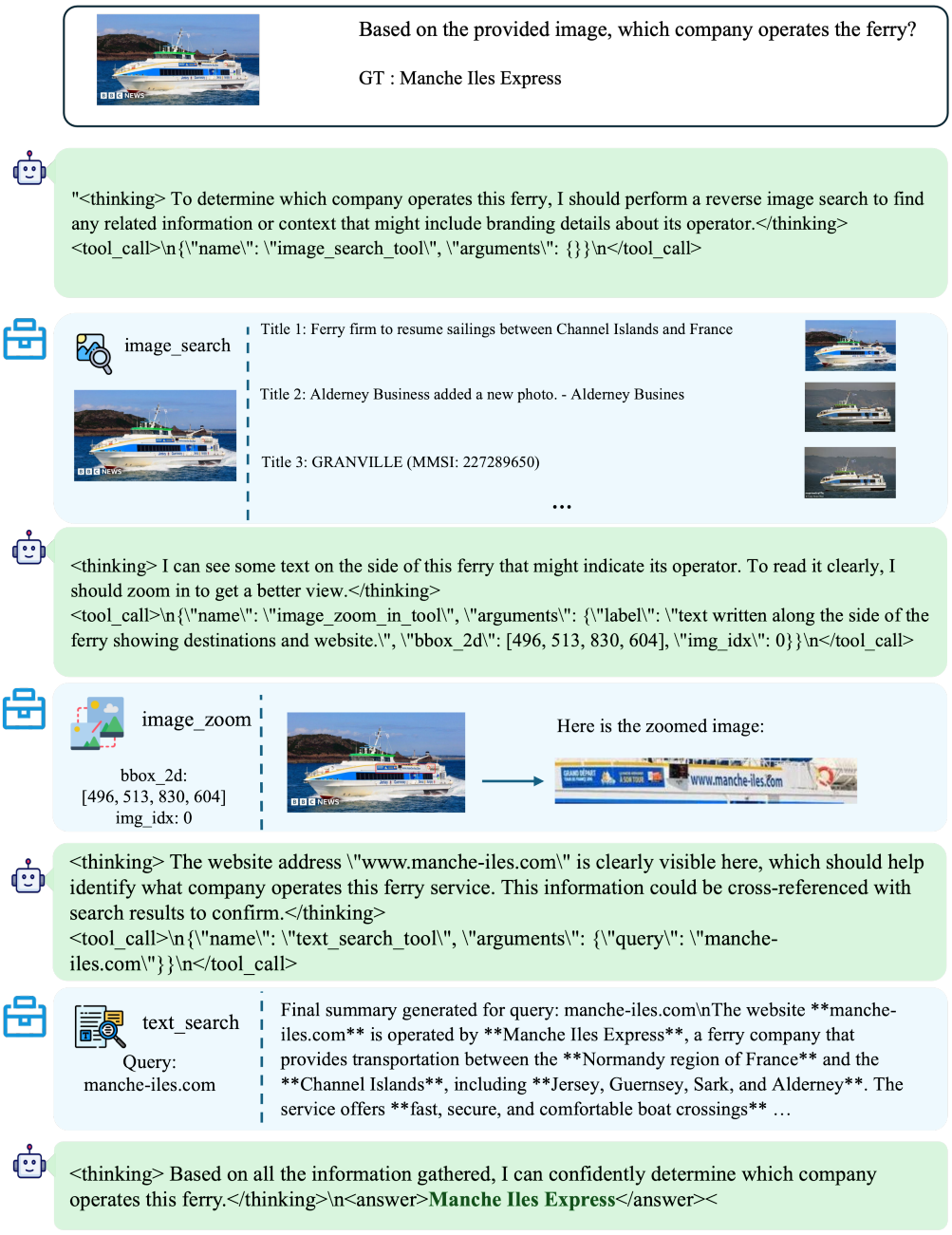

Figure 14: **Case 4: Visual Detail Extraction via Zoom + Text Confirmation.** Unable to identify the operator from a global image search, the model zooms into the hull text, reads the URL, and confirms *Manche Iles Express* via a follow-up search.

Case 3: Persistent Query Refinement for Precise Numerical Retrieval. The question asks: *"What is the width (in metre) of this island?"* (ground truth: **300**). The model invokes `image_search_tool` and identifies the island as *Frauenchiemsee* from the Wikipedia result. Three successive text searches are then issued with increasing specificity: (1) *"Frauenchiemsee island width"* returns area (15.5 ha) and population, but no explicit width; (2) *"Frauenchiemsee island dimensions"* similarly omits the metric width; (3) *"Frauenchiemsee island length and width"* finally retrieves *"approximately 620 m long and 300 m wide"*. This adaptive reformulation—persisting through partial failures until the precise numerical fact is extracted—demonstrates the model's ability to iteratively refine its information-seeking strategy rather than giving up prematurely.

Case 4: Visual Detail Extraction via Zoom + Text Confirmation. The question asks: *"Based on the provided image, which company operates the ferry?"* (ground truth: **Manche Iles Express**). The model first calls `image_search_tool`, which returns ferry-related headlines (e.g. "Ferry firm to resume sailings between Channel Islands and France") but no direct operator name. Recognising that the text painted on the hull may be legible at higher resolution, it invokes `image_zoom_in_tool` on region `[496, 513, 830, 604]`, successfully reading *"www.manche-iles.com"*. A final `text_search_tool`[*"manche-iles.com"*] confirms that this URL belongs to *Manche Iles Express*, a ferry service between Normandy and the Channel Islands. This case exemplifies a *coarse-to-fine* visual strategy: global image search for scene context, zoom for fine-grained textual evidence, and a confirmatory knowledge lookup.