COG-RESEARCH: MITIGATING GROUNDING DRIFT IN LONG-HORIZON MULTIMODAL DEEP RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal deep research agents can acquire reliable evidence yet fail to preserve it through long-horizon reasoning and report synthesis. We characterize this failure as Grounding Drift: previously established visual observations, claim–source bindings, or task constraints become omitted, corrupted, or misbound at later stages. We introduce Cog-Research, a post-training framework that converts intermediate research evidence into supervision for grounded report synthesis through source-linked, task-specific rubrics constructed from research traces, workspace artifacts, and human reports. We build a process-grounded dataset of 500 Plan-and-Act research tasks and perform trajectory SFT followed by rubric-guided GRPO with rewards for evidence-informed criterion coverage and report quality. Built on Qwen3-VL-8B-Instruct, Cog-Research improves the base agent by 4.85/3.54/4.91 points on MMDeepResearch-Bench, DeepResearch-Bench II, and MiroEval-Multimodal, respectively, and reduces the task-level incidence of primary Grounding Drift errors from 30.7% to 17.1% on MMDeepResearch-Bench.

1 INTRODUCTION

Deep research agents combine planning, retrieval, tool use, and report synthesis to answer complex research requests (Huang et al., 2025; Li et al., 2025b; Tongyi DeepResearch Team et al., 2025). In multimodal research, they must integrate images, charts, tables, and documents with external textual evidence. Success requires preserving each observation together with its association with the correct source, entity, and task requirement throughout a long research trajectory.

These associations can break even after perception and retrieval succeed. An agent may correctly extract facts, save them in a workspace, and later attach them to the wrong image or claim. The resulting report can be fluent and well cited while contradicting its own research record.

To understand where failures arise, we trace visual observations, retrieved evidence, workspace artifacts, and their later use rather than inspecting only final reports. On the base Qwen3-VL-8B-Instruct agent over MMDeepResearch-Bench (140 tasks), 43 of 111 attributable primary errors show a recurring *correct-before/incorrect-after* pattern: valid grounding is established upstream but later omitted, corrupted, or reassigned. Figure 1 makes this temporal mechanism explicit. In Q28, negative VC values are transferred to $LiPF_6$ stocks; in Q101, Kiki's visual attributes are reassigned to Chihiro; and in Q133, distinct image sources collapse into one purported milestone. We call this failure **Grounding Drift**: previously established, still-relevant grounding is lost, corrupted, or misbound in a later research state.

This diagnosis exposes a training mismatch. Existing systems preserve research state through shared memory, verified facts, or citation-aware reasoning (MindDR Team & Li Auto Inc., 2026; Zhang et al., 2025a; 2026), but storing evidence does not ensure that its source and entity bindings survive repeated downstream reuse. Likewise, supervision centered on tool success or final report quality need not connect intermediate grounding to later synthesis. This motivates a simple design principle: make recoverable intermediate evidence and provenance part of the training signal, and reward whether downstream reasoning and reporting preserve task-local bindings.

Cog-Research operationalizes this principle with trajectory SFT and **process-grounded, rubric-guided GRPO**. Its training resource links research tasks, Planner/Actor traces, workspace artifacts, and structured rubrics. The constructed source-linked criteria connect evidence to downstream re-

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.

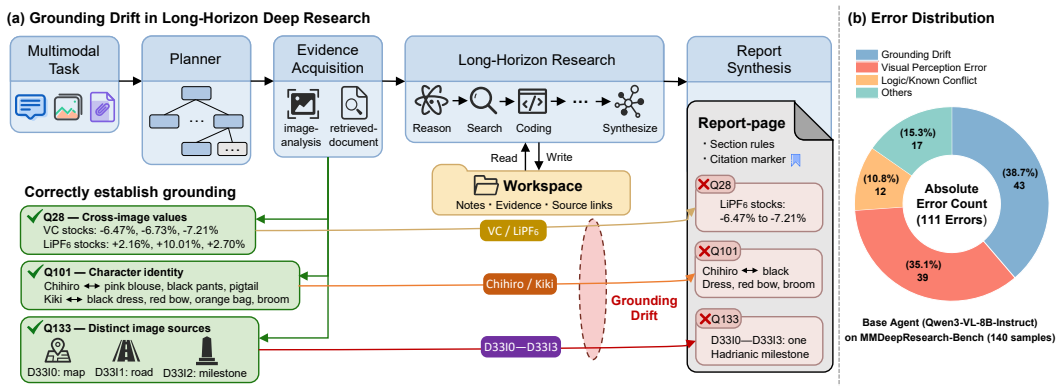


Figure 1: Grounding Drift in long-horizon multimodal research. (a) Examples of cross-image value transfer (Q28), character-identity misbinding (Q101), and collapse of distinct image sources (Q133). (b) Base-agent diagnostic on MMDeepResearch-Bench (140 tasks): 43 of 111 attributable primary errors are Grounding Drift.

quirements. SFT initializes planning, reasoning, tool use, and workspace operations; RL then scores process execution, evidence use, criterion coverage, and report quality over complete trajectories.

Our contributions are threefold:

1. **Grounding Drift definition and analysis.** We define a trajectory-level evidence-preservation failure and provide an annotation protocol and multimodal case analysis separating Grounding Drift from first-occurrence errors and hallucination.
2. **Plan-and-Act research data.** We construct 500 research tasks with linked Planner/Actor traces, SFT samples, structured rubrics, and workspace artifacts.
3. **Process-grounded rubric-guided GRPO.** We connect intermediate evidence and workspace artifacts to source-linked, task-specific criteria and optimize the policy with process- and outcome-guided GRPO to reduce Grounding Drift during evidence reuse.

2 Related Work

2.1 Deep Research Agents

Deep research combines information acquisition, planning, tool use, and report writing (Huang et al., 2025). WebThinker interleaves search with reasoning and drafting; WebDancer and DeepResearcher train tool-using policies through trajectory construction and reinforcement learning (Li et al., 2025b; Wu et al., 2025; Zheng et al., 2025). Tongyi DeepResearch, Step-DeepResearch, MiroThinker, and Marco DeepResearch further develop context management, research skills, and verification (Tongyi DeepResearch Team et al., 2025; Hu et al., 2025; MiroMind Team et al., 2026; Zhu et al., 2026).

Memory and verification are especially relevant to grounding preservation. Existing systems often use persistent workspaces. Mind DeepResearch uses shared memory, Co-Sight maintains structured verified facts, and MS-Agent manages context through files (MindDR Team & Li Auto Inc., 2026; Zhang et al., 2025a; ModelScope Team, 2026). Chaining the Evidence introduces citation-aware rubrics and evidence-chain rewards (Zhang et al., 2026). Citation Drift studies changing references, while Context Drift considers divergence during extended interaction (Ram, 2025; Dongre et al., 2025). Other works address workspace context bloat through dynamic workspace compaction or iterative report consolidation. WebResearcher consolidates findings into evolving reports and focused workspaces, while IterResearch periodically reconstructs research state (Qiao et al., 2025; Chen et al., 2026a). DR-DCI expands a local corpus workspace and supports workspace-preserving context resets (Lu et al., 2026). Their primary objective is context scaling and interaction efficiency. Cog-Research studies a complementary failure: whether observations that remain available are still bound to the correct image, entity, claim, and task constraint when reused later. We treat tool ob-

servations and workspace artifacts as recoverable provenance for training and diagnosis, rather than proposing a new compression or reset policy.

Our reward design builds on rubric-based outcome evaluation and process supervision (Lightman et al., 2024; Wang et al., 2024; Yuan et al., 2026). Research traces and workspaces provide task-specific criteria and evidence, while deterministic rules evaluate coverage and observable failures. This connects process-grounded data to final synthesis without introducing an LLM reward judge. We use GRPO (Shao et al., 2024; Guo et al., 2025) to optimize the shared policy and use trajectory analysis to measure grounding preservation.

2.2 Multimodal Deep Research Agents

Relative to text-only research, multimodal research must retain both textual evidence and visual observations, including their cross-modal provenance. A retrieved statement about a similar object cannot automatically serve as evidence about the user's reference image. Maintaining these distinctions throughout long-horizon research introduces additional failure modes beyond factual acquisition.

The progression from VQA to multimodal search and then report-oriented research clarifies the task requirements. Classic VQA evaluates an answer to an image–question pair (Antol et al., 2015); MMSearch adds web-based multimodal retrieval and answer synthesis (Jiang et al., 2024). Their answer-oriented endpoints do not require a full analytical report. MMDeepResearch-Bench and MiroEval evaluate longer research outputs and their evidence use (Huang et al., 2026a; Ye et al., 2026). However, QA and tool-centric evaluation do not fully capture aggregation across heterogeneous sources, cross-modal synthesis, and discourse organization (Huang et al., 2025). These abilities matter when reports combine prose, tables, charts, and citations.

Vision-DeepResearch, WebWatcher, and MMSearch-R1 train visual/textual information seeking (Huang et al., 2026b; Geng et al., 2026; Wu et al., 2026). Multimodal DeepResearcher targets text–chart reports, while R1-VL studies step-wise multimodal reinforcement learning (Yang et al., 2026; Zhang et al., 2025b). MTA-Agent provides verified multi-hop multimodal QA trajectories, while MM-DeepResearch constructs search-intensive QA data and tool-use trajectories (Peng et al., 2026; Yao et al., 2026). Cog-Research links role-specific traces and intermediate evidence artifacts to report-oriented rubrics, providing supervision across planning, execution, and synthesis.

3 Grounding Drift in Multimodal Deep Research

We represent a research trajectory as

$$\tau = (x, I) \to P \to (R_1, A_1, O_1) \to \cdots \to (R_T, A_T, O_T) \to y,$$

where x and I are the query and multimodal inputs, P is the plan, and R_t, A_t, and O_t denote reasoning, action, and observation at step t; y is the final report. For DAG execution, this sequence is a dependency-consistent serialization of reasoning/action/observation events and workspace updates (e.g. a topological sort or timestamp-ordered serialization of actions and workspace updates). Each action conditions on the information actually visible to its Actor.

A grounding instance $g = (v, b, s)$ consists of a validated observation or requirement v, its binding b to a claim, entity, image, source, or task constraint, and recoverable support s. We define **Grounding Drift** as a trajectory-level failure where (1) g is correctly established upstream, (2) it remains relevant to the task, and (3) a later research state or report omits, corrupts, replaces, or misbinds it without supporting evidence for the change. Initial perception errors, evidence-supported revisions, and irrelevant intermediate facts are not Grounding Drift. Detailed omission and relevance criteria are given in Appendix A.1.

Relation to hallucination and perception errors. These concepts describe different aspects of failure: Visual Perception Error concerns whether information is acquired correctly, hallucination concerns whether generated content is supported, and Grounding Drift concerns whether correctly established grounding is preserved downstream. We use *visual hallucination* narrowly for visual content asserted without image support; it is therefore a subset of Visual Perception Error. Ground-

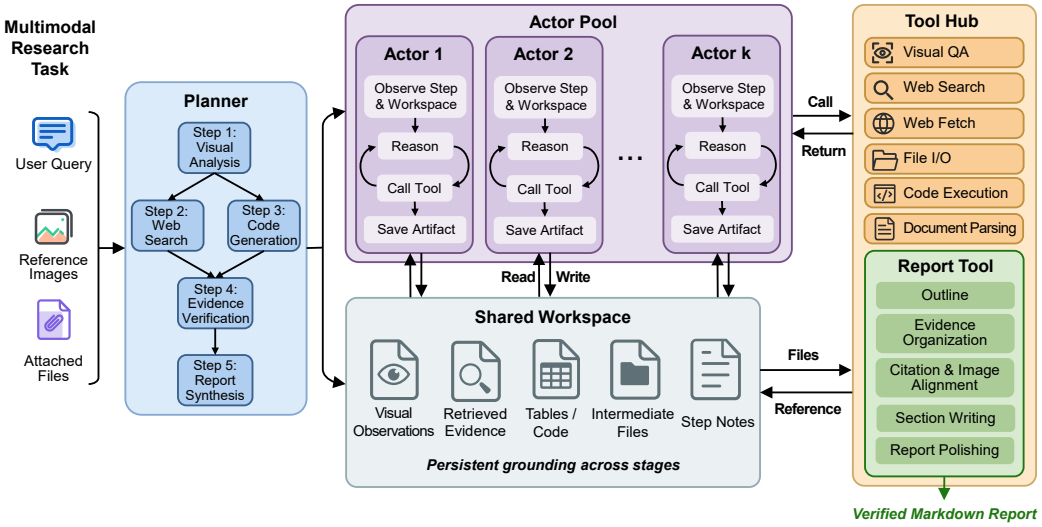


Figure 2: Cog-Research execution architecture. The shared workspace persists research artifacts across Plan-and-Act stages.

ing Drift instead requires a temporal correct-before/incorrect-after witness and can occur through omission or rebinding without hallucination.

This distinction is reflected empirically. On the 140-task MMDeepResearch-Bench Base audit, 43 tasks contain a primary Grounding Drift error. We identify 55 hallucination cases, including 15 visual and 40 non-visual cases. Only 5 of the 43 Grounding Drift cases overlap with the non-visual hallucination set (11.6%), corresponding to 12.5% of non-visual hallucinations. Thus, Grounding Drift is a trajectory-level diagnostic dimension rather than a synonym or mutually exclusive alternative to hallucination. Detailed overlap analyses with hallucination, context loss, citation/source errors, and entity confusion are reported in Appendix A.1.

Diagnostic analysis. We analyze the Base, SFT, and SFT-RL checkpoints on the 140-task MMDeepResearch-Bench. Gemini-2.5-Pro proposes candidate errors from benchmark evaluator explanations, followed by human verification against the raw traces, artifacts, images, and reports. Each trajectory is assigned at most one attributable primary error—Grounding Drift, Visual Perception Error, Logic/Known Conflict, or Others—with drift requiring validated before/after evidence. Appendix A.1 details the annotation protocol and Top-2 co-occurrence analysis; Appendix A.2 provides case studies, and Appendix A.9 reports broader agent error distributions.

4 METHOD

4.1 OVERVIEW

Cog-Research uses the Plan-and-Act architecture (Erdogan et al., 2025) in Figure 2. A Planner constructs a dependency-aware DAG; Actors execute ready steps with visual QA, search, code, document, and file tools; and a shared workspace persists evidence and intermediate conclusions for synthesis. Trajectory SFT and rubric-guided GRPO jointly train the shared Qwen3-VL-8B-Instruct checkpoint across Planner and Actor.

4.2 PROCESS-GROUNDED DATA AND RUBRIC CONSTRUCTION

Existing deep-research training predominantly uses serialized ReAct traces (Tongyi DeepResearch Team et al., 2025; Li et al., 2025b; Huang et al., 2026b). Answer-oriented resources such as MMSearch-Plus, MM-BrowseComp, and DeepMMSearchVQA emphasize browsing or VQA (Tao et al., 2026; Li et al., 2025a; Narayan et al., 2025). Plan-and-Act research instead requires super-

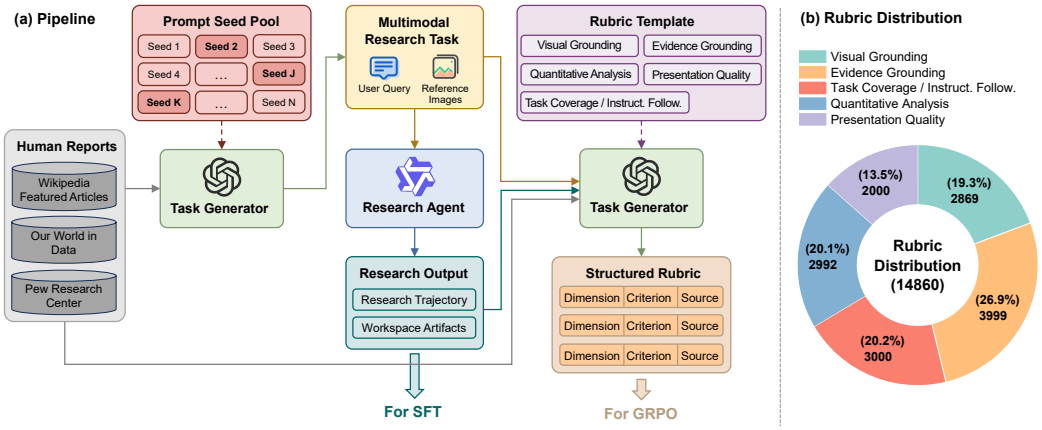


Figure 3: Process-grounded data and rubric construction. Human-authored reports seed multimodal research tasks. A teacher agent produces trajectories and workspace artifacts; source-linked rubrics are constructed from the task, images, trace/workspace evidence, and human-authored source report. The right panel shows the rubric distribution across five dimensions.

vision across explicit planning, execution, workspace updates, and report synthesis. Our corpus retains these linked components (Figure 3).

Task synthesis. Tasks are derived from human-authored reports and associated images from Wikipedia Featured Articles, Our World in Data, and Pew Research Center (Wikipedia, 2026; Our World in Data, 2026; Pew Research Center, 2026). Directly asking a language model to convert each report into a question often yields tasks that are narrow in scope and repetitive. To address this, we constructed a pool of prompt seeds to control attributes such as research objectives, reasoning requirements, desired coverage, multimodal dependencies, and expected report structure. GPT-5.5 (OpenAI, 2026) combines sampled seeds with source material to generate tasks involving visual interpretation, cross-source research, comparison, quantitative reasoning, and structured synthesis.

Trajectory collection. Qwen3.7-Max (Alibaba Cloud, 2026) executes the tasks under Plan-and-Act. Each task links its query, images, Planner/Actor records, workspace artifacts, and report. The corpus contains 500 Planner records and 3,368 Actor-step records: **3,868 trace records**. The 500 RL/rubric tasks use the same task IDs as SFT. Appendices A.3–A.4 detail data composition, filtering, and overlap checks.

Structured rubrics. GPT-5.5 receives the task, images, teacher trace, workspace, the original human-authored source report, and a five-dimension template, producing

$$\mathcal{R}_i = \{(d_j, c_j, s_j, w_j)\}_{j=1}^{M_i}.$$

Each tuple specifies a **Dimension** d_j, atomic **Criterion** c_j, **Source** metadata s_j, and rubric weight w_j. The five dimensions are Visual Grounding, Evidence Grounding, Task Coverage / Instruction Following, Quantitative Analysis, and Presentation Quality. They cover visual interpretation, evidence use, required content, numerical reasoning, and coherent presentation. Source metadata records provenance and supporting URLs. Criteria retain task-specific entity–attribute–value statements together with their sources, rather than only isolated keywords. The construction instructions require correct figure–object–claim correspondence, source alignment with the task entity and time scope, and numerical consistency in units, periods, and comparison baselines. The original human-authored source report is included as a factual source for rubric construction. There are **14,860** criteria for the 500 tasks. Deterministic evaluators score criterion coverage, citation-local co-occurrence, and source consistency; rubric weights also define a diagnostic aggregate. Appendix A.5 distinguishes these checks from full relation-level verification.

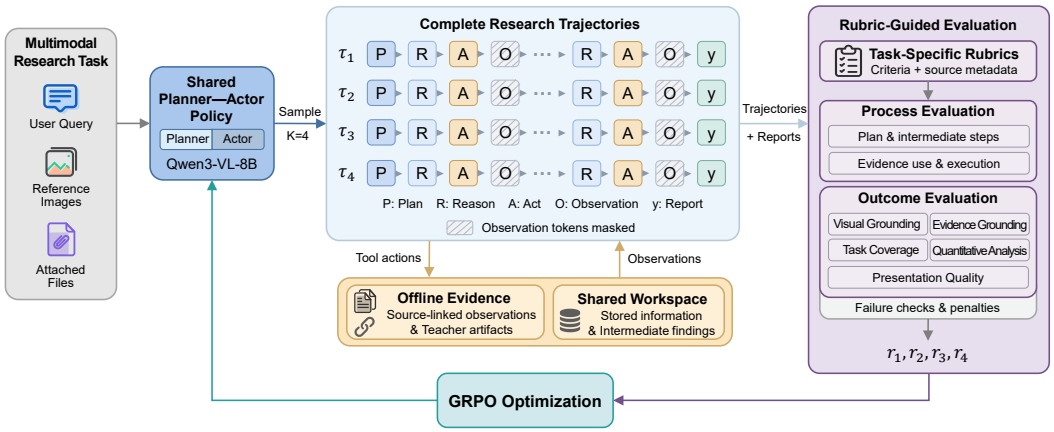


Figure 4: Process-grounded training workflow. Teacher traces and workspace artifacts support trajectory SFT, an offline evidence environment, and hidden evaluation rubrics. A shared Planner/Actor policy generates complete research trajectories and final reports, which are scored by deterministic process and report evaluators to produce a trajectory reward for GRPO.

4.3 Trajectory SFT

Planner and Actor examples train the same Qwen3-VL-8B-Instruct checkpoint with $\mathcal{L}_{\text{SFT}} = -\sum_{t \in \mathcal{A}} \log \pi_\theta(a_t \mid H_t)$, where $\mathcal{A}$ contains assistant-token positions and H_t is the available prefix. System, user, and tool-observation tokens are masked from the loss. Turn-level samples teach local tool-result-to-action behavior, while workflow-level samples expose longer execution patterns. This initializes reasoning, tool use, workspace operations, and synthesis before RL.

4.4 Rubric-Guided GRPO with Process-Grounded Rewards

Online policy learning in an offline environment. Our RL implementation adapts JADE's shared-parameter planning/execution framework (Chen et al., 2026b) with GRPO. Each rollout generates a plan, reasoning, tool calls, intermediate summaries, and a report. The action space includes search, browsing, vision, code, file, and report tools. Retrieval accesses fixed task-local evidence, including historical tool outputs. All generated Planner/Actor tokens receive policy loss; tool observations are masked. The policy is trained on-policy through fresh trajectories in an offline retrieval environment (Figure 4).

4.4.1 Process-Grounded Evaluation

Process supervision uses observations, intermediate findings, and task requirements recorded during research. Structured rubrics associate each criterion with an evaluation dimension and its source; the task-local offline evidence environment exposes the corresponding source-preserving observations during rollout. Training rewards are computed by deterministic Python evaluators rather than online LLM/VLM judges. Binding information enters the reward through task-specific factual criteria, coverage within citation-bearing sentences, and checks against the task-local evidence URLs. These provide partial binding supervision; they do not constitute a complete verifier of entity–attribute–value–source relations.

The process branch independently scores plan quality, matched execution steps, and completion:

$$R_{\text{step,raw}} = 0.15A_{\text{plan}} + 0.75\overline{S}_{\text{step}} + 0.10\rho_{\text{completion}} - P_{\text{dependency}},$$

where $P_{\text{dependency}} = \min(0.20, 0.04N_{\text{invalid dependency}})$. For a matched teacher/execution step k,

$$S_k = 0.70S_{\text{coverage}} + 0.12S_{\text{evidence}} + 0.08S_{\text{artifact}} + 0.10S_{\text{dependency}}.$$

Here the final report contributes only a binary nonempty-submission indicator to the average step score, not GEN/EVI-style report quality. Missing a plan, missing a report, or leaving most steps unfinished caps the process score at 0.40, 0.60, and 0.55, respectively; the resulting trajectory-level score is denoted R_{step}.

4.4.2 OUTCOME SCORING AND GRPO OPTIMIZATION

For each report dimension d, rubric criteria are macro-averaged into $C_d \in [0, 1]$ and combined with a deterministic heuristic H_d:

$$\widetilde{R}_d = 0.70H_d + 0.30C_d.$$

Data preparation requires at least two criteria in every dimension. If any required dimension is absent, the parser rejects the whole structured rubric and falls back to the base heuristics.

The report/evidence branch then computes

$$R_{\text{raw}} = 0.33\widetilde{R}_{\text{visual}} + 0.27\widetilde{R}_{\text{evidence}} + 0.16\widetilde{R}_{\text{task}} + 0.13\widetilde{R}_{\text{quant}} + 0.11\widetilde{R}_{\text{report}} - P_{\text{invalid}} - P_{\text{repeat}}.$$

The score is first clipped to $[0, 1]$, then bounded by the smallest active hard-error cap. Thus $R_{\text{report}} = \min(\text{clip}(R_{\text{raw}}, 0, 1), C_{\min})$ when a hard gate fires, and equals the clipped raw score otherwise. Hard-error caps cover empty/crashed output, detected source/data failures, severe repetition, and serious format/instruction failures; source/data checks are limited to the available structural checks and supplied failure metadata. Appendix A.5 specifies criterion scoring, heuristics and penalties.

Joint RL combines the two trajectory-level branches as

$$r_k = (1 - \lambda)R_{\text{report},k} + \lambda R_{\text{step},k}, \qquad \lambda = 0.5.$$

For each group of $K = 4$ rollouts,

$$\widehat{A}_k = \frac{r_k - \mu_r}{\sigma_r + \epsilon}.$$

GRPO applies this advantage to all generated Planner/Actor tokens under their actual local contexts (Shao et al., 2024; Guo et al., 2025). Tool errors are counted within each sampled trajectory, allowing P_{invalid} to vary across the group. Appendix A.5 gives the full optimization and rollout configuration, and Section 5.4 compares supervision variants.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Benchmarks and metrics. We evaluate MMDeepResearch-Bench (Huang et al., 2026a) (140 image–text tasks; GEN, EVI, VEF, MM), DeepResearch-Bench II (Li et al., 2026) (132 text-oriented tasks; InfoRecall, Analysis, Presentation), and MiroEval (Ye et al., 2026) (70 text-only and 30 multimodal tasks; Coverage, Insight, Instruction Following, Clarity), using each benchmark's prescribed aggregation.

Models and evaluation. Cog-Research uses Qwen3-VL-8B-Instruct (Bai et al., 2025); we evaluate the Base, SFT, and SFT-RL checkpoints against WebThinker-R1-7B, Vision-DeepResearch-8B, MiroThinker-v1-8B, MS-Agent, and Co-Sight, with Tongyi-DeepResearch-30B scores from the respective benchmark papers. For rerun systems, we report means over three runs. Gemini-2.5-Pro evaluates MMDeepResearch-Bench and DeepResearch-Bench II, while GPT-5.1 evaluates MiroEval (Google DeepMind, 2025; OpenAI, 2025); Grounding Drift is separately diagnosed by the evidence-review protocol in Section 3. Appendix A.6 gives rerun and judge settings, and Appendix A.12 reports checkpoint dispersion; the appendix also details compute costs.

5.2 OVERALL DEEP-RESEARCH PERFORMANCE

Table 1, with checkpoint dispersion in Appendix A.12, shows that Cog-Research improves Overall from 23.22 to 28.07 on MMDeepResearch-Bench, 26.93 to 30.47 on DeepResearch-Bench II, 48.36 to 53.27 on MiroEval-Multimodal, and 50.38 to 54.13 on MiroEval-Text. The final checkpoint achieves the highest scores among the 7–8B systems listed in the table on all four aggregate metrics. On the primary MMDeepResearch-Bench, Evidence Quality increases by 7.76 points and Visual Evidence Fidelity by 5.69 over Base, consistent with the two grounding dimensions targeted by training.

Table 1: Main results (3-run rerun means; Tongyi scores are from benchmark papers). Higher is better. Yellow cells mark the best 7–8B result among the systems in this table.

DeepResearch Agent	MMDeepResearch-Bench					DeepResearch-Bench II			
	Overall	GEN	EVI	VEF	MM	**Overall**	InfoRecall	Analysis	Presentation
WebThinker-R1-7B	**13.22**	34.63	5.91	5.41	11.45	**11.07**	6.97	15.75	41.69
MS-Agent-8B	**16.56**	34.04	17.65	4.00	12.18	**14.09**	10.38	15.28	47.42
Vision-DeepResearch-8B	**18.80**	44.95	8.33	18.75	11.88	**25.66**	20.33	21.28	84.69
Co-Sight-8B	**22.73**	40.79	16.44	23.81	16.26	**25.92**	20.41	29.74	71.54
MiroThinker-v1-8B	**22.46**	48.40	15.80	7.63	21.73	**27.37**	21.16	26.34	88.57
Cog-Research-8B	**23.22**	41.06	16.97	24.22	16.90	**26.93**	21.10	30.32	76.42
Cog-Research-8B-SFT	**25.61**	41.96	23.34	25.00	17.39	**28.57**	23.97	31.63	66.90
Cog-Research-8B-SFT-RL	**28.07**	46.14	24.73	29.91	18.12	**30.47**	25.55	35.25	68.72
Tongyi-DeepResearch-30B	**29.02**	52.34	27.04	12.86	26.23	**29.89**	22.95	35.89	86.13

DeepResearch Agent	MiroEval-Multimodal					MiroEval-Text				
	Overall	Coverage	Insight	Instruct. Follow.	Clarity	**Overall**	Coverage	Insight	Instruct. Follow.	Clarity
WebThinker-R1-7B	**27.59**	32.93	23.57	35.83	32.64	**22.26**	24.91	21.53	30.48	26.06
MS-Agent-8B	**24.81**	25.67	25.25	31.93	32.91	**25.89**	28.80	24.02	36.05	32.41
Vision-DeepResearch-8B	**38.21**	40.18	34.60	56.11	50.20	**33.39**	37.20	28.20	49.94	41.52
Co-Sight-8B	**48.70**	56.79	45.62	64.13	51.51	**49.12**	55.06	50.95	62.42	47.28
MiroThinker-v1-8B	**47.28**	51.23	43.16	60.38	59.36	**47.51**	56.27	40.60	66.38	54.49
Cog-Research-8B	**48.36**	57.03	45.19	65.94	53.32	**50.38**	56.52	52.46	63.91	49.67
Cog-Research-8B-SFT	**52.44**	60.46	54.97	62.28	55.79	**54.01**	62.42	53.11	68.07	54.28
Cog-Research-8B-SFT-RL	**53.27**	63.29	53.88	60.66	55.29	**54.13**	61.47	55.07	67.66	53.85
Tongyi-DeepResearch-30B	**-**	-	-	-	-	**-**	-	-	-	-

SFT improves all four Overall scores over Base. RL further improves both DeepResearch-Bench II and MiroEval. Overall, the gains concentrate in grounding, coverage, and analysis, which are most directly aligned with our training objective. We note that not all metrics improved. DeepResearch-Bench II Presentation drops from 76.42 to 68.72. One possible explanation is the lack of explicit presentation-specific supervision during trajectory SFT and the relatively small presentation weight during RL. The decline in MiroEval-Multimodal instruction following may reflect a distribution shift between our predominantly image-based training tasks and MiroEval's more diverse multimodal document inputs. We therefore retain MiroEval as an out-of-distribution transfer test.

5.3 Grounding Drift Reduction

Table 2: Primary-error diagnostic on MMDeepResearch-Bench.

Variant	Grounding Drift	Visual Perception Error	Logic/Known Conflict	Others	Error tasks
Base	43	39	12	17	111/140
SFT	36	37	13	14	100/140
w/o SFT	30	38	8	11	87/140
Outcome-only	31	35	14	14	94/140
Process-only	23	36	7	15	81/140
SFT-RL	24	35	10	12	81/140

Table 2 shows that, categorized by primary error, Grounding Drift incidence decreases from **30.7%** (43/140) to **17.1%** (24/140) on MMDeepResearch-Bench, while the total number of error-bearing tasks falls from 111 to 81. In contrast, the absolute number of Visual Perception Errors changes only from 39 to 35. This pattern suggests that the improvement is primarily associated with preserving already established grounding during downstream reasoning and synthesis, rather than reducing first-occurrence visual perception errors. The criteria for primary errors, inter-annotator reliability, and Top-2 co-occurrence analyses are reported in Appendix A.1.

5.4 Contributions of SFT and Process-Grounded RL

Table 3 summarizes the training-strategy ablations. The *w/o SFT* variant removes trajectory-SFT initialization. The two RL-branch ablations retain the SFT checkpoint and the same interactive Plan-and-Act rollout environment: *Outcome-only* optimizes R_{report} ($\lambda = 0$), whereas *Process-only* optimizes R_{step} ($\lambda = 1$). Detailed ablation protocols are provided in Appendix A.14.

Table 3: Training-strategy ablation. Outcome-only removes RL process/step scoring; process-only removes report-outcome scoring. w/o SFT alone removes supervised initialization.

DeepResearch Agent	MMDeepResearch-Bench					MiroEval-Multimodal				
	Overall	GEN	EVI	VEF	MM	**Overall**	Coverage	Insight	Instruct. Follow.	Clarity
Full (SFT + joint RL)	**28.07**	46.14	24.73	29.91	18.12	**53.27**	63.29	53.88	60.66	55.29
w/o SFT	**26.92**	45.85	26.22	23.24	17.44	**48.38**	57.28	42.66	66.62	51.03
Outcome-only (w/o process reward)	**26.18**	44.61	22.23	26.55	17.61	**51.54**	57.37	53.90	63.39	51.90
Process-only (w/o outcome reward)	**27.98**	45.13	25.09	29.91	18.17	**49.53**	56.88	51.65	60.99	53.49

Table 4: Ablation of dimension-associated reward components.

DeepResearch Agent	MMDeepResearch-Bench					MiroEval-Multimodal				
	Overall	GEN	EVI	VEF	MM	**Overall**	Coverage	Insight	Instruct. Follow.	Clarity
Cog-Research-8B-SFT-RL	**28.07**	46.14	24.73	29.91	18.12	**53.27**	63.29	53.88	60.66	55.29
w/o Visual Grounding component	**26.75**	45.40	25.11	21.55	16.09	**52.87**	61.67	53.22	61.01	54.90
w/o Evidence Grounding component	**25.94**	45.94	20.88	26.50	17.29	**52.77**	61.61	54.54	60.10	55.18
w/o Task Following component	**26.91**	44.33	24.06	28.16	17.30	**51.25**	59.10	54.07	58.31	54.21
w/o Quantitative Analysis component	**27.16**	45.60	24.42	27.39	17.45	**52.21**	62.70	51.84	60.18	55.32
w/o Presentation Quality & Hard Gates	**21.57**	38.41	16.94	21.07	15.30	**47.26**	56.23	45.35	55.97	49.78

Removing SFT reduces MiroEval-Multimodal Overall from 53.27 to 48.38 and MMDeepResearch Visual Evidence Fidelity from 29.91 to 23.24. This suggests that trajectory SFT provides an important initialization for visual-evidence use and transferable execution behavior. The two reinforcement learning (RL) reward branches serve distinct roles. Outcome-only lowers Overall from 28.07 to 26.18 on MMDeepResearch-Bench and from 53.27 to 51.54 on MiroEval-Multimodal, showing that process rewards contribute to overall performance. Process-only remains close to Full on MMDeepResearch-Bench (27.98 vs. 28.07) and yields 23 Grounding Drift errors versus 24 for Full; however, its MiroEval-Multimodal score falls from 53.27 to 49.53, suggesting that the joint objective improves cross-benchmark transfer.

5.5 Reward Component Analysis

Removing Visual Grounding lowers VEF from 29.91 to 21.55, while removing Evidence Grounding lowers EVI from 24.73 to 20.88. Task Following and Quantitative Analysis ablations similarly reduce their aligned MiroEval metrics. This pattern suggests that the dimension-associated rewards provide targeted learning signals rather than acting only as a generic regularizer.

The largest degradation follows removing the Presentation Quality and Hard Gates bundle, including the associated invalid-tool and repetition penalties, lowering Overall by 6.50 points on MMDeepResearch-Bench and 6.01 on MiroEval-Multimodal. This coupled ablation measures the contribution of structure and failure controls together. Table 12 in Appendix A.8 provides complementary repetition diagnostics: severe repetition falls from 19.91% after SFT to 5.93% after joint RL, and rises to 25.89% without presentation/repetition controls. Output stabilization is therefore a substantial part of the observed benefit, rather than evidence of binding preservation alone.

6 Conclusion

We identify *Grounding Drift* as a trajectory-level failure pattern in long-horizon multimodal research, where previously established evidence, provenance, or task constraints are lost, distorted, or misbound during later reasoning and report synthesis. To address this, we introduce Cog-Research, which combines trajectory SFT with process-grounded, rubric-guided GRPO over Plan-and-Act traces, workspace evidence, and task-specific criteria. Cog-Research improves overall report scores across three research benchmarks and reduces primary Grounding Drift incidence from 30.7% to 17.1% on MMDeepResearch-Bench. Ablations show that combining outcome and process rewards yields stronger cross-benchmark performance than either objective alone. Integrating grounding-preserving supervision with dynamic workspace reconstruction and adaptive search is a promising direction for long-horizon research.

AI USE STATEMENT

Generative AI supported both the research pipeline and manuscript preparation. GPT-5.5 generated synthetic tasks and rubrics; Qwen3.7-Max produced teacher trajectories and reports; and benchmark-specific models evaluated research outputs. AI tools also assisted literature discovery, brainstorming, methodology refinement, coding, debugging, and writing. The authors reviewed AI-assisted outputs and take responsibility for the study and manuscript.

ETHICS STATEMENT

This work studies reliability failures in multimodal deep-research agents and does not aim to automate high-stakes decision making. The constructed corpus is derived from publicly accessible human-authored materials and model-generated tasks, trajectories, and rubrics. Potential concerns include biases or factual errors inherited from source materials and generative models, as well as provenance and licensing considerations when redistributing derived data and images. Any public release should preserve source attribution and follow the applicable licenses and terms of the original sources. Cog-Research is designed to improve evidence preservation but does not eliminate hallucination or guarantee factual correctness; outputs should therefore be independently verified in high-stakes settings.

REPRODUCIBILITY STATEMENT

The anonymous repository and appendix document task construction, trajectory SFT, the offline evidence environment, deterministic reward evaluation, GRPO settings, and benchmark evaluation. The appendix includes dataset statistics, overlap checks, annotation rules, diagnostic definitions, run-to-run variability, and compute costs. Our Code and datasets are available at `https://anonymous.4open.science/r/Cog-Research-66F5`

REFERENCES

Alibaba Cloud. Qwen3.7-Max, 2026. URL `https://www.alibabacloud.com/help/en/model-studio/qwen3-7-max`. Accessed: 2026-06-01.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2015.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Guoxin Chen, Zile Qiao, Xuanzhong Chen, Donglei Yu, Haotian Xu, Xin Zhao, Ruihua Song, Wenbiao Yin, Huifeng Yin, Liwen Zhang, et al. Iterresearch: Rethinking long-horizon agents with interaction scaling. In *International Conference on Learning Representations*, 2026a.

Yiqun Chen, Erhan Zhang, Tianyi Hu, Shijie Wang, Zixuan Yang, Meizhi Zhong, Xiaochi Wei, Yan Gao, Yi Wu, Yao Hu, et al. Jade: Bridging the strategic-operational gap in dynamic agentic rag. In *Forty-third International Conference on Machine Learning*, 2026b.

Vardhan Dongre, Ryan A Rossi, Viet Dac Lai, David Seunghyun Yoon, Dilek Hakkani-Tür, and Trung Bui. Drift no more? context equilibria in multi-turn llm interactions. *arXiv preprint arXiv:2510.07777*, 2025.

Lutfi Eren Erdogan, Nicholas Lee, Sehoon Kim, Suhong Moon, Hiroki Furuta, Gopala Anumanchipalli, Kurt Keutzer, and Amir Gholami. Plan-and-act: Improving planning of agents for long-horizon tasks. *arXiv preprint arXiv:2503.09572*, 2025.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Kuan Li, Yida Zhao, et al. Webwatcher: Breaking new frontiers of vision-language deep research agent. In *International Conference on Learning Representations*, 2026.

Google DeepMind. Gemini 2.5 Pro, 2025. URL https://blog.google/innovation-and-ai/models-and-research/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed: 2026-06-01.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in llms through reinforcement learning. *Nature*, (8081), 2025.

Chen Hu, Haikuo Du, Heng Wang, Lin Lin, Mingrui Chen, Peng Liu, Ruihang Miao, Tianchi Yue, Wang You, Wei Ji, et al. Step-deepresearch technical report. *arXiv preprint arXiv:2512.20491*, 2025.

Peizhou Huang, Zixuan Zhong, Zhongwei Wan, Donghao Zhou, Samiul Alam, Xin Wang, Zexin Li, Zhihao Dou, Li Zhu, Jing Xiong, et al. Mmdeepresearch-bench: A benchmark for multimodal deep research agents. *arXiv preprint arXiv:2601.12346*, 2026a.

Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, Zehui Chen, Yao Hu, Philip Torr, Feng Zhao, and Wanli Ouyang. Vision-deepresearch: Incentivizing deepresearch capability in multimodal large language models. In *Forty-third International Conference on Machine Learning*, 2026b.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Huichi Zhou, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, et al. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*, 2025.

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.

Ruizhe Li, Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench ii: Diagnosing deep research agents via rubrics from expert report. *arXiv preprint arXiv:2601.08536*, 2026.

Shilong Li, Xingyuan Bu, Wenjie Wang, Jiaheng Liu, Jun Dong, Haoyang He, Hao Lu, Haozhe Zhang, Chenchen Jing, Zhen Li, et al. Mm-browsecomp: A comprehensive benchmark for multimodal browsing agents. *arXiv preprint arXiv:2508.13186*, 2025a.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. *Advances in Neural Information Processing Systems*, 2025b.

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 *International Conference on Learning Representations*, 2024.

Yi Lu, Zhuofeng Li, Ping Nie, Haoxiang Zhang, Yuyu Zhang, Kai Zou, Wenhu Chen, Jimmy Lin, Dongfu Jiang, and Yu Zhang. Dr-dci: Scaling direct corpus interaction via dynamic workspace expansion. *arXiv preprint arXiv:2606.14885*, 2026.

MindDR Team and Li Auto Inc. Mind DeepResearch technical report. *arXiv preprint arXiv:2604.14518*, 2026.

MiroMind Team, S. Bai, L. Bing, L. Lei, R. Li, X. Li, X. Lin, E. Min, L. Su, B. Wang, et al. MiroThinker-1.7 & H1: Towards heavy-duty research agents via verification. *arXiv preprint arXiv:2603.15726*, 2026.

ModelScope Team. Agentic Insight v2: A lightweight and extensible multimodal deep research framework. MS-Agent GitHub repository, 2026. URL https://github.com/modelscope/ms-agent/tree/main/projects/deep_research/v2. Accessed: 2026-03-01.

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-5.1, 2025. URL https://openai.com/index/gpt-5-1/. Accessed: 2026-06-01.

OpenAI. GPT-5.5, 2026. URL https://openai.com/index/introducing-gpt-5-5/. Accessed: 2026-06-01.

Our World in Data. OWID, 2026. URL https://ourworldindata.org/. Accessed: 2026-06-01.

Xiangyu Peng, Can Qin, An Yan, Xinyi Yang, Zeyuan Chen, Ran Xu, and Chien-Sheng Wu. MTA-Agent: An open recipe for multimodal deep search agents. *arXiv preprint arXiv:2604.06376*, 2026.

Pew Research Center. Pew, 2026. URL https://www.pewresearch.org/. Accessed: 2026-06-01.

Zile Qiao, Guoxin Chen, Xuanzhong Chen, Donglei Yu, Wenbiao Yin, Xinyu Wang, Zhen Zhang, Baixuan Li, Huifeng Yin, Kuan Li, et al. Webresearcher: Unleashing unbounded reasoning capability in long-horizon agents. *arXiv preprint arXiv:2509.13309*, 2025.

Gokul Srinath Seetha Ram. Citation drift: Measuring reference stability in multi-turn llm conversations. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, 2025.

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.

Xijia Tao, Yihua Teng, Xinxing Su, Xinyu Fu, Jihao Wu, Chaofan Tao, Ziru Liu, Haoli Bai, Rui Liu, and Lingpeng Kong. Mmsearch-plus: A simple yet challenging benchmark for multimodal browsing agents. In *International Conference on Learning Representations*, 2026.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, et al. Tongyi DeepResearch technical report. *arXiv preprint arXiv:2510.24701*, 2025.

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)*, 2024.

Wikipedia. Wikipedia, 2026. URL https://www.wikipedia.org/. Accessed: 2026-06-01.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhenglin Wang, Zhengwei Tao, Ding-Chu Zhang, Zekun Xi, Robert Tang, et al. Webdancer: Towards autonomous information seeking agency. *Advances in Neural Information Processing Systems*, 2025.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing lmms to search. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, July 2026. doi: 10.18653/v1/2026.acl-long.114.

Zhaorui Yang, Bo Pan, Han Wang, Yiyao Wang, Xingyu Liu, Luoxuan Weng, Yingchaojie Feng, Haozhe Feng, Minfeng Zhu, Bo Zhang, et al. Multimodal deepresearcher: Generating text-chart interleaved reports from scratch with agentic framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Huanjin Yao, Qixiang Yin, Min Yang, Ziwang Zhao, Yibo Wang, Haotian Luo, Jingyi Zhang, and Jiaxing Huang. Mm-deepresearch: A simple and effective multimodal agentic search baseline. *arXiv preprint arXiv:2603.01050*, 2026.

Fangda Ye, Yuxin Hu, Pengxiang Zhu, Yibo Li, Ziqi Jin, Yao Xiao, Yibo Wang, Lei Wang, Zhen Zhang, Lu Wang, et al. MiroEval: Benchmarking multimodal deep research agents in process and outcome. *arXiv preprint arXiv:2603.28407*, 2026.

Huining Yuan, Zelai Xu, Huaijie Wang, Xiangmin Yi, Jiaxuan Gao, Xiao-Ping Zhang, Yu Wang, Chao Yu, and Yi Wu. Verifiable process rewards for agentic reasoning. *arXiv preprint arXiv:2605.10325*, 2026.

Hongwei Zhang, Ji Lu, Shiqing Jiang, Chenxiang Zhu, Li Xie, Chen Zhong, Haoran Chen, Yurui Zhu, Yongsheng Du, Yanqin Gao, et al. Co-sight: Enhancing llm-based agents via conflict-aware meta-verification and trustworthy reasoning with structured facts. *arXiv preprint arXiv:2510.21557*, 2025a.

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 *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-VL: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Bin Zhu, Qianghuai Jia, Tian Lan, Junyang Ren, Feng Gu, Feihu Jiang, Longyue Wang, Zhao Xu, and Weihua Luo. Marco deepresearch: Unlocking efficient deep research agents via verification-centric design. *arXiv preprint arXiv:2603.28376*, 2026.

A Appendix

A.1 Grounding Drift Annotation Protocol

The unit is one MMDeepResearch-Bench trajectory. The benchmark evaluator's explanation supplies candidate errors; the raw trace, workspace, images, and report supply adjudication evidence. For a reliability audit, 140 trajectories are stratified-randomly sampled across the Base, SFT, and SFT-RL checkpoints: 47, 46, and 47 trajectories, respectively. The annotation label set includes Correct when no attributable error is identified; the reliability sample is not restricted to the four error labels. Two annotators label them independently while checkpoint identity is hidden. For each example they record a primary label and at most one secondary label, the upstream "before" evidence, the downstream error, the source binding, and whether the temporal condition for Grounding Drift is satisfied.

Primary labels and omission boundaries. The primary label records the adjudicated benchmark-relevant failure, not every error in a trajectory. Uninterpretable/truncated/severely repetitive outputs are assigned *Other/Generation failure*; a correct upstream grounding followed by task-relevant omission, corruption, or misbinding is *Grounding Drift*; an already-wrong first visual/OCR extraction is *Initial visual/OCR bottleneck*; and an independent calculation, logic, or domain-knowledge error with otherwise correct visual/source binding is *Isolated knowledge/logic/calculation*. Remaining cases are *Other*. If two independent errors both affect the core answer, the evaluator's explicitly critical/primary error is selected; otherwise the error affecting more required subtasks is primary, with unresolved ties sent to a third annotator. The second independent error is retained as a secondary label.

For omission, annotators apply the same operational rule as Section 3. They first identify an anchor in the user request or a still-required Planner-DAG subgoal; the upstream fact must materially support that anchor. They then check whether the final state/report contains the fact, a semantically equivalent summary, or an evidence-supported replacement. Omission drift is assigned only when none is present and the anchored requirement is consequently unsupported, incomplete, or misbound. Side observations without such an anchor, facts subsumed by a supported summary, and evidence-supported corrections are permissible compression rather than drift.

Boundary audit against hallucination and adjacent concepts. We distinguish the dimensions being measured rather than treating the labels as mutually exclusive. Visual Perception Error concerns the information-acquisition stage; Grounding Drift concerns later preservation/reuse; hallucination concerns output support. We define *visual hallucination* narrowly as asserting visual content not supported by the image. It is therefore a subset of Visual Perception Error: missed detections, OCR/reading mistakes, fine-grained recognition errors, and spatial/relation errors are perception failures but are not automatically hallucinations. *Non-visual hallucination* denotes unsupported textual, knowledge, numerical, or source generation. Grounding Drift requires all three temporal conditions in Section 3: correct upstream grounding, continued task relevance, and a concrete downstream corruption/omission/misbinding witness.

On the Base checkpoint, the audit contains 55 hallucination cases: 15 visual and 40 non-visual. Let H denote the non-visual hallucination set for the overlap analysis; first-occurrence visual hallucinations are already accounted for under Visual Perception Error. Five cases belong to both H and Grounding Drift (GD), yielding

$$P(H \mid GD) = \frac{5}{43} = 11.6\%, \qquad P(GD \mid H) = \frac{5}{40} = 12.5\%,$$

$$J(H, GD) = \frac{5}{43 + 40 - 5} = 6.4\%.$$

Thus, most hallucinations do not satisfy the temporal witness required by Grounding Drift, and most Grounding Drift cases are not hallucinations.

Table 5: Overlap between the 43 Base-checkpoint Grounding Drift cases and adjacent error concepts. These descriptors may overlap.

Adjacent concept	Overlap with GD	Share of GD
Context loss / context corruption	15/43	34.9%
Citation / source-reference error	12/43	27.9%
Entity confusion / entity misbinding	10/43	23.3%

These overlaps clarify the novelty boundary: context loss, citation error, and entity misbinding describe recurring forms or consequences of failure, whereas Grounding Drift groups only cases with a validated correct-before/incorrect-after trajectory witness. It is therefore a trajectory-level diagnostic lens rather than a claim that the underlying error primitives are entirely new.

Table 6 gives representative strict-audit witnesses illustrating that these labels capture different axes of the same trajectory. "Task H" marks whether the task contains a non-visual hallucination; "GD H" marks whether the drift event itself is hallucinated rather than a pure omission or rebinding.

Table 6: Representative Grounding Drift boundary-audit cases. Y indicates that the descriptor applies; – indicates that it does not apply in the audited case.

QID	Main temporal change	Task H	GD H	Context loss	Source error	Entity confusion
Q15	U.S. BLS data later bound to a Japan claim	Y	–	–	Y	–
Q16	Established ∼8.5% later replaced by 2.5%	Y	Y	Y	–	–
Q28	Negative returns from image A bound to image-B companies	–	–	–	Y	Y
Q29	Current price reused as historical-event-window price	Y	Y	Y	–	–
Q42	Near-full modulo remapping later changed to $1/N$	–	–	Y	–	–
Q47	Correct $1.54\times$ later rewritten as $2.105\times$	–	–	Y	–	–
Q85	Natural gas rise later changed to all-fuels decline	–	–	Y	–	–
Q101	Kiki/Chihiro attributes transferred after correct separation	–	–	–	–	Y
Q119	School-level policy reframed as national DfE guidance	Y	Y	–	Y	Y
Q133	Distinct map/infographic/milestones collapsed to one source	Y	Y	–	Y	Y

Inter-annotator reliability. Before adjudication, the two annotators agree on the Top-1 label for $115/140 = 82.1\%$ of the audit examples. Their marginal counts over {Grounding Drift, Visual/OCR, Logic/Knowledge, Other, Correct} are $(35, 37, 13, 14, 41)$ and $(34, 39, 11, 15, 41)$, giving $P_e = 0.238$ and Cohen's $\kappa = 0.766$. For the set-valued Top-2 labels, sample-level Jaccard

agreement is 0.857:

$$J = \frac{1}{N}\sum_{i=1}^{N}\frac{|A_i \cap B_i|}{|A_i \cup B_i|} = 0.857.$$

Counts and denominators. Table 7 reports all four primary categories, including the training-strategy ablations. Drift share uses the attributable-error count in each row; task-level incidence uses all 140 tasks. Primary-error and Top-2 counts report the first audit pass for each checkpoint, rather than an average over the three evaluation runs used for report-performance scores. The 140-trajectory inter-annotator reliability sample described above is a separate audit subset.

Table 7: Primary-error counts on the same 140-task benchmark. Visual denotes initial perception/OCR/entity errors; Logic denotes isolated knowledge, logic, or calculation errors.

Variant	Grounding Drift	Visual Perception Error	Logic/Known Conflict	Others	Error tasks
Base	43	39	12	17	111/140
SFT	36	37	13	14	100/140
w/o SFT	30	38	8	11	87/140
Outcome-only	31	35	14	14	94/140
Process-only	23	36	7	15	81/140
SFT-RL	24	35	10	12	81/140

Supplementary Top-2 co-occurrence audit. A secondary label is added when the recorded reason identifies another independent error that affects report quality. Downstream consequences of the same initial error are counted once. Both analyses use the attributable-error denominators Base (111), SFT (100), w/o SFT (87), outcome-only (94), process-only (81), and SFT-RL (81). Categories can overlap, so their counts can sum to more than the number of error-bearing tasks. Across all benchmark tasks, Top-2 drift count changes from 48 to 34.

Table 8: Top-2 error analysis. Cells report counts. Each task contributes at most two independent labels.

Variant	Grounding Drift	Visual Perception Error	Logic/Known Conflict	Others	Error tasks
Base	48	50	15	28	111/140
SFT	47	51	13	30	100/140
w/o SFT	40	48	10	21	87/140
Outcome-only	46	45	15	24	94/140
Process-only	35	49	12	15	81/140
SFT-RL	34	45	14	14	81/140

A.2 Case Studies with Raw Evidence

The following cases trace how image identities, numerical observations, source provenance, and task constraints change across research stages.

Q28: cross-image value transfer. The upstream record correctly separates two images: `Q28I0.jpg` concerns vinylene carbonate (VC), while `Q28I1.jpg` concerns $LiPF_6$. Their stock-price changes are recorded as follows (company names are romanized):

Image / topic	Company	Change
`Q28I0.jpg` / VC	Furi Gufen	-6.47%
	Huasheng Lidian	-6.73%
	Yongtai Keji	-7.21%
`Q28I1.jpg` / $LiPF_6$	Tianji Gufen	$+2.16\%$
	Shida Shenghua	$+10.01\%$
	Tianci Cailiao	$+2.70\%$

The final report transfers the negative changes from `Q28I0.jpg` to the three companies in `Q28I1.jpg`, stating that all three $LiPF_6$ stocks declined by 6.47–7.21%. The numbers were observed, but their image/company bindings were corrupted.

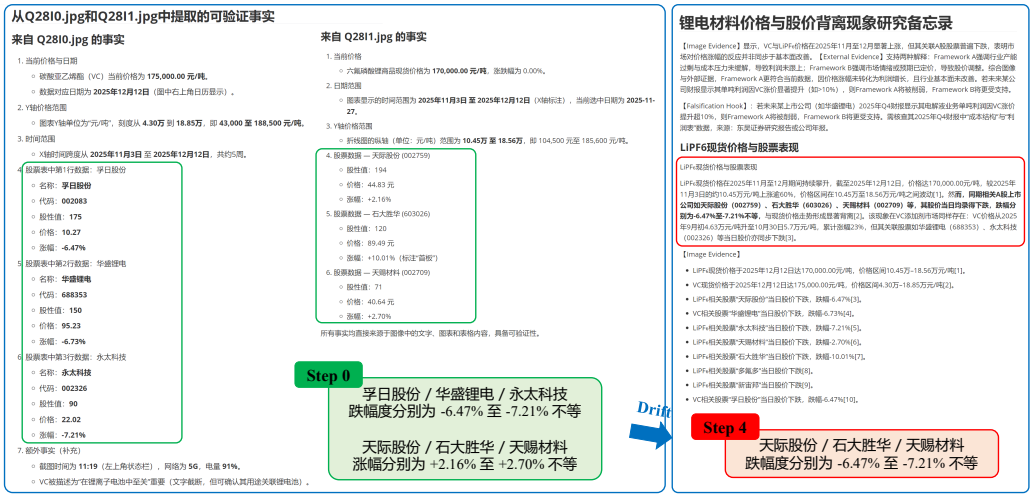


Figure 5: Q28: values correctly extracted from the VC image are reassigned to companies in the $LiPF_6$ image during report synthesis.

Q101: cross-section identity misbinding. The upstream record establishes `D1I0.png` as Chihiro from *Spirited Away* and `D1I1.png` as Kiki from *Kiki's Delivery Service*. In a later section on the dynamic color palette of *Spirited Away*, the report assigns Kiki's black dress, red bow, broom, and backpack from `D1I1.png` to Chihiro. The extracted attributes persist, but their character and source identities change across sections.

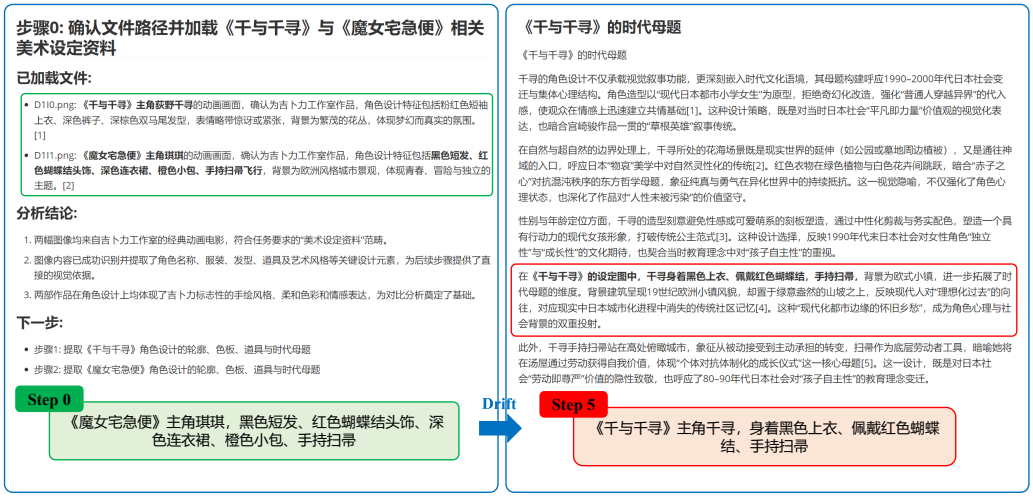


Figure 6: Q101: visual attributes bound to Kiki in the upstream record are reassigned to Chihiro in a later report section.

Q133: multi-image source collapse. The upstream analysis distinguishes `D33I0` as a historical map, `D33I1` as a road-teaching infographic, and `D33I2` as a milestone from Trajan's reign. The final report instead describes `D33I0`–`D33I3` as different rubbings or photographs of one Hadrianic

milestone and attributes the same inscription to all of them. Distinct image–fact bindings collapse into a single purported source.

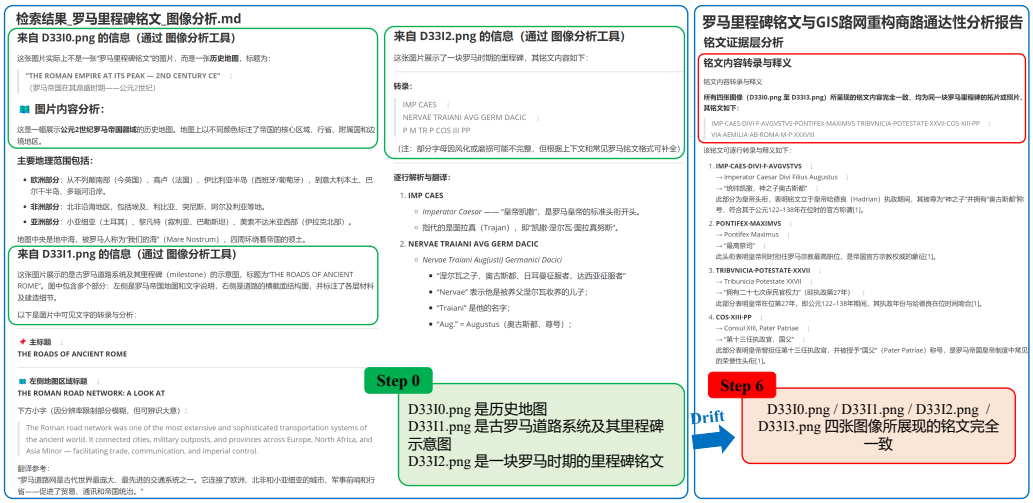


Figure 7: Q133: a map, an instructional infographic, and a Trajanic milestone lose their distinct identities and are described as views of one Hadrianic milestone.

Additional cases.

- **Q15—cross-country evidence misbinding.** Retrieved evidence explicitly identifies the U.S. Bureau of Labor Statistics and U.S. energy-price/core-CPI data. The final report uses these data to support claims about Japan's inflation and policy-transmission mechanisms.
- **Q16—numerical corruption.** Image analysis records a 2001 peak near 8.5%; final synthesis rewrites it as approximately 2.5%.
- **Q27—image-to-workspace misbinding.** Image identifiers and contents become mismatched during the early analysis and storage of three reference images, before final report synthesis.
- **Q29—date and task-constraint drift.** The requested event occurred on July 26, 2021, but the model treats it as occurring on July 26, 2023. It also substitutes explicitly acknowledged hypothetical/simulated values for the required historical data.
- **Q85—contradictory synthesis.** Visual analysis and evidence checking establish continued growth in natural gas. The report first states that coal, oil, and natural gas all decline, then later acknowledges growth in natural gas.
- **Q119—authority-level misattribution.** The document is initially identified as a Victoria Road Primary School policy, but the report presents it as the U.K. DfE's 2024 national guidance.

A.3 Training Data Statistics

Table 9 separates independent tasks, role-specific trace records, and derived SFT samples. The SFT and RL/rubric task sets share IDs 0–499.

Table 10 summarizes the composition of the 500 structured-rubric tasks, including source categories, language, difficulty, and image counts.

A.4 Data Filtering and Overlap Checks

The pipeline filters empty contexts, short assistant targets, selected samples without tool behavior, and length violations, then deduplicates rendered conversations. Task-level filters check execution

Table 9: Training-data statistics.

Data	Count
Independent tasks	500
Planner trace records	500
Actor-step trace records	3,368
Total teacher trace records	3,868
Turn-level SFT samples	27,559
Workflow-level SFT samples	7,690
SFT samples (pre-filter)	35,249
Overlength samples removed	360
SFT training samples	34,889
Tasks with structured rubrics	500
Total rubric criteria	14,860

Table 10: Distribution of the 500 structured-rubric tasks.

Category	**Item**	**Count**
Report source	Wikipedia Featured Articles	257
	Pew Research Center	116
	Our World in Data	127
Language	English	310
	Chinese	131
	German	25
	French	24
	Finnish	10
Difficulty	Easy	137
	Medium	248
	Hard	115
Image count	0 images	2
	1 image	235
	2 images	128
	3 images	135

success, final-report artifacts, blocked tasks, and trajectory correctness. Evidence deduplication is task-local and source-sensitive; rubric clauses are compared using token similarity.

The audited data contain 500 unique training prompts, with zero exact prompt overlap and zero image-SHA256 overlap against the locally available MMDeepResearch-Bench and MiroEval data.

A.5 Reward and Optimization Configuration

Deterministic evaluation and rubric aggregation. The reward configuration is `reward_model.style="rule"`; report and trajectory evaluation use deterministic Python code rather than online LLM/VLM reward judges. Optional `judge_reason` and `failure_tags` are externally precomputed metadata, not online judges. For dimension d, each valid criterion receives $c_{d,k} \in [0, 1]$ and the dimension rubric score is an unweighted macro-average,

$$C_d = \frac{1}{|K_d|} \sum_{k \in K_d} c_{d,k}.$$

Non-object criteria are skipped, missing criterion IDs are generated, missing `evaluation` defaults to `content`, and a dimension with no valid criterion receives $C_d = 0$. Normal data preparation requires at least two criteria per dimension; if any of the five required dimensions is absent, the complete structured rubric is invalidated and scoring falls back to H_d. Rubric dimension weights are used only for the diagnostic aggregate

$$R_{\text{rubric}} = \frac{\sum_d w_d C_d}{\sum_d w_d},$$

and are not added again to the policy reward, which uses fixed coefficients $(0.33, 0.27, 0.16, 0.13, 0.11)$.

For a `content` criterion, token coverage is

$$c = \min\left(1, \frac{|T_{\text{criterion}} \cap T_{\text{report}}|}{0.55|T_{\text{criterion}}|}\right),$$

and when the criterion contains numbers this score is multiplied by $0.55 + 0.45 \times$ number coverage. English tokenization follows `[a-z][a-z0-9_-]{2,}|\d+(?:\.\d+)?%?`; Chinese text uses overlapping bigrams of contiguous Han-character strings. A `claim_with_citation` criterion uses

$$c = 0.20 c_{\text{whole report}} + 0.80 c_{\text{resolved cited sentences}},$$

where both terms use the token and numerical coverage rules above, but the second restricts the evaluated text to sentences containing resolvable citations. This encourages criterion content to co-occur with citations rather than merely appear somewhere in the report. It does not by itself establish that the specific URL next to a claim entails that claim.

Binding information and source consistency. The rubric builder (`scripts/structure_quiz_train_rubrics.py`) draws task-specific statements from existing rubric entries, Planner/Actor traces, and intermediate workspace artifacts. Its instructions preserve figure–object–claim correspondence and entity, attribute, numerical value, unit, time scope, and source associations. Newly introduced workspace facts require support from other sources, and agent-generated final reports are excluded as factual inputs to rubric construction. Criterion records retain `provenance` and `supporting_urls` alongside the statement.

The evaluator (`cosight_rl/rewards/reward_evaluator.py`) checks whether citation identifiers map uniquely to URLs, whether mappings are resolvable, whether orphan mappings remain, and how many claim-like sentences contain citations. It also compares report URLs with the current task's local evidence pack, using report URL precision and evidence URL coverage in H_{local}. These checks reward traceable use of task-local sources and discourage unsupported URL generation; URL membership alone does not establish semantic support for an individual claim.

Scope of binding verification. The implementation combines binding information retained in rubric statements, citation-local co-occurrence, and evidence-pool source checks. It does not implement a complete relation-level verifier for $(\text{Entity}, \text{Attribute}, \text{Value}, \text{Source})$. For example, a

report that exchanges values between entities while retaining all entities, numbers, and URLs can still obtain substantial coverage scores. Optional judge-supplied failure labels can inform hard gates, but should not be conflated with verification performed by the lexical and URL checks themselves. Grounding Drift is therefore measured separately using the trajectory-level evidence audit in Appendix A.1.

Evidence and rollout scope. The Evidence Corpus and hidden evaluator rubrics are distinct objects. Appendix A.13 gives corpus limits and the tool-access protocol. The policy generates the current plan, evidence queries, workspace updates, and report; $N_{\text{invalid tool}}$ is the count of invalid tool actions recorded for the sampled trajectory, not a fixed count copied from a teacher prefix. Invalid calls and invalid evidence-ID references can therefore affect within-group rewards. Historical code/browser outputs enter the rollout as retrieved observations.

Report/outcome reward. For each dimension, $\widetilde{R}_d = 0.70H_d + 0.30C_d$, and the complete raw report score is

$$R_{\text{raw}} = 0.33\widetilde{R}_{\text{visual}} + 0.27\widetilde{R}_{\text{evidence}} + 0.16\widetilde{R}_{\text{task}} + 0.13\widetilde{R}_{\text{quant}} + 0.11\widetilde{R}_{\text{report}} - P_{\text{invalid}} - P_{\text{repeat}}.$$

It is first clipped to $[0, 1]$ and then bounded by the smallest active catastrophic-error cap,

$$R_{\text{report}} \equiv R_{\text{final}} = \min\left(\text{clip}(R_{\text{raw}}, 0, 1), \min_{h \in \mathcal{H}} \text{cap}(h)\right),$$

with the second argument omitted when no hard gate fires.

Process reward: reference steps and criterion coverage. The step evaluator (`cosight_rl/rewards/step_reward_evaluator.py`) uses deterministic criterion coverage, not embeddings, natural-language inference, or an LLM judge. The implementation's `semantic` field denotes this coverage score. Let $\mathcal{T}_{\text{nf}} = \{t_k\}_{k=1}^{K_{\text{nf}}}$ be the non-final reference steps and $\mathcal{G} = \{g_j\}_{j=1}^{M}$ the completed non-final steps saved by the agent. Each reference step specifies criteria $\{c_{k,l}\}_{l=1}^{L_k}$, expected tools, and expected artifacts. The text scored for g_j concatenates its `objective` and `summary`.

Let $\mathcal{V}(x)$ be the token set of text x. English words have at least three characters; numbers and percentages are retained as complete tokens. Contiguous Chinese text is split into overlapping character bigrams, and common stopwords are removed. For a generated text x and criterion c with a nonempty token set, define

$$o(x, c) = \frac{|\mathcal{V}(x) \cap \mathcal{V}(c)|}{|\mathcal{V}(c)|}, \qquad b(x, c) = \min\left(1, \frac{o(x, c)}{0.45}\right).$$

With $\mathcal{N}(c) \subseteq \mathcal{V}(c)$ denoting numerical tokens, the criterion score is

$$C(x, c) = \begin{cases} b(x, c)\left(0.65 + 0.35\dfrac{|\mathcal{N}(c) \cap \mathcal{V}(x)|}{|\mathcal{N}(c)|}\right), & |\mathcal{N}(c)| > 0, \\ b(x, c), & |\mathcal{N}(c)| = 0. \end{cases}$$

An empty criterion scores zero. The step-pair coverage is

$$S^{k,j}_{\text{coverage}} = \frac{1}{L_k} \sum_{l=1}^{L_k} C(g_j, c_{k,l}).$$

These step-criterion thresholds and numerical weights differ from the report-criterion rules above.

Step-pair scoring and one-to-one matching. Each non-final reference/generated step pair receives

$$S_{k,j} = 0.70S^{k,j}_{\text{coverage}} + 0.12S^{k,j}_{\text{evidence}} + 0.08S^{k,j}_{\text{artifact}} + 0.10S^{k,j}_{\text{dependency}}.$$

The evidence component is one if g_j contains an evidence ID or t_k does not require external retrieval, and zero otherwise. External retrieval is required when the reference teacher tool names

include search, web fetching, or document extraction. The artifact component is one if g_j saves an artifact or t_k does not require one, and zero otherwise. The dependency component is the indicator of g_j's `dependency_valid` flag. Evidence-ID access is additionally constrained by the rollout environment (Appendix A.13).

All $S_{k,j}$ pairs are sorted in descending score order. The evaluator greedily accepts each positive-score pair whose reference and generated steps are both unused. Thus each generated step can satisfy at most one reference step, and each reference step can be matched at most once. Unmatched reference steps score zero; unmatched or extra generated steps do not enter the average step score. Reusing a single saved conclusion cannot satisfy several reference steps. This procedure corresponds to `_match_teacher_steps` in the step evaluator.

Plan alignment. Plan matching includes all T reference steps, including the final-report step. In reference-step order, the evaluator selects the unused generated plan-step description with the highest mean criterion coverage. Let a_k be the selected coverage, or zero for an unmatched reference step, and let G be the number of generated plan steps. Then

$$r_{\text{count}} = \frac{\min(G,T)}{\max(G,T)}, \qquad A_{\text{plan}} = \left(\frac{1}{T}\sum_{k=1}^{T} a_k\right)(0.70 + 0.30 r_{\text{count}}).$$

If either plan is empty, $A_{\text{plan}} = 0$. Missing steps reduce coverage through $a_k = 0$, while extra steps reduce the count-alignment factor. This is implemented by `_greedy_plan_alignment`.

Step aggregation, completion, and caps. The final-report step is excluded from execution-step pair matching. It contributes only

$$S_{\text{report}} = \mathbf{1}[\text{a nonempty report is submitted}].$$

Let $\pi(k)$ denote the execution-step match for reference step k, with unmatched terms set to zero. The denominator includes every non-final reference step and one final-report step:

$$\overline{S}_{\text{step}} = \frac{\sum_{k=1}^{K_{\text{nf}}} S_{k,\pi(k)} + S_{\text{report}}}{K_{\text{nf}} + 1}.$$

Completion instead counts all M completed steps saved in the workspace, not just matched steps:

$$\rho_{\text{completion}} = \min\left(1, \frac{M}{\max(1, K_{\text{nf}})}\right).$$

The raw process reward and invalid-dependency penalty are

$$R_{\text{step,raw}} = 0.15 A_{\text{plan}} + 0.75 \overline{S}_{\text{step}} + 0.10 \rho_{\text{completion}} - P_{\text{dependency}},$$
$$P_{\text{dependency}} = \min(0.20, 0.04 N_{\text{invalid dependency}}).$$

The caps are 0.40 for a missing plan and 0.60 for a missing submitted report. A cap of 0.55 applies when $M < \max(1, \lfloor K_{\text{nf}}/2 \rfloor)$. Let C_{step} be the smallest active cap, or one if no cap is active. The final process reward is

$$R_{\text{step}} = \min(\text{clip}(R_{\text{step,raw}}, 0, 1), C_{\text{step}}).$$

GEN/EVI-style report-quality scores are not used in this branch. Joint RL combines R_{step} with R_{report} using $\lambda = 0.5$ as in Section 4.4.

Active penalties and caps. The tool penalty is

$$P_{\text{invalid}} = \min(0.20, 0.05 N_{\text{invalid tool}}).$$

The exact soft repetition penalty is given in Appendix A.8. The raw report score subtracts these penalties, is clipped to $[0, 1]$, and is then bounded by the lowest triggered hard cap. Caps are bounds, not additional subtractions.

Detected failure condition	Reward cap
Empty or crashed output	0.05
Detected source/data failure	0.20
Severe repetition	0.25
Serious format/instruction violation	0.35

Source checks detect malformed URLs and orphan citation mappings; local-evidence matching additionally measures whether report URLs occur in the task evidence pool. These structural and provenance checks do not independently verify factual entailment or numerical relations.

Base heuristics. Let $\mathrm{clip}(z) = \min(1, \max(0, z))$.

For visual grounding, c_{img} is the fraction of expected image identifiers (filenames/stems or generic identifiers) mentioned in the report, n_v counts visual-description terms, n_u counts uncertainty terms, and N_{img} is the expected image count:

$$H_{\mathrm{visual}} = \mathrm{clip}\left(0.65c_{\mathrm{img}} + 0.25\min\left(1, \frac{n_v}{\max(4, 4N_{\mathrm{img}})}\right) + 0.10u\right),$$

where $u = \min(1, n_u/3)$ if $n_u > 0$, and $u = 0.25$ otherwise. With no expected image the default is 0.55.

For evidence grounding, the citation score is $H_{\mathrm{cit}} = 0.25H_{\mathrm{legacy}} + 0.75H_{\mathrm{aligned}}$. When mappings are resolvable,

$$H_{\mathrm{aligned}} = 0.05 + 0.40m + 0.20q + 0.15u + 0.10d + 0.10r - 0.25o,$$

where m measures citation-to-reference mapping integrity, q claim citation coverage, u URL quality, d source diversity, r reference-section presence, and o the orphan-mapping ratio. If local evidence is available,

$$H_{\mathrm{evidence}} = 0.55H_{\mathrm{cit}} + 0.45H_{\mathrm{local}};$$

otherwise it equals H_{cit}. The local score itself blends legacy host/title/URL overlap (0.25) with alignment features (0.75): report URL precision, evidence URL coverage, mapping integrity, claim coverage, and orphan mappings.

Task following starts from 0.45, adds capped heading/bullet/table contributions, and subtracts penalties for violated requested Markdown structure, table/list/short-answer requirements, explicit two-/three-item constraints, and language mismatch.

For quantitative analysis, n_{num} and n_{formula} count numeric and formula matches. If the prompt contains a quantitative cue,

$$H_{\mathrm{quant}} = \mathrm{clip}(0.30 + \min(0.35, 0.015n_{\mathrm{num}}) + \min(0.20, 0.03n_{\mathrm{formula}}) + 0.05\mathbf{1}_{\mathrm{uncertain}});$$

otherwise $H_{\mathrm{quant}} = \mathrm{clip}(0.55 + \min(0.12, 0.004n_{\mathrm{num}}))$.

The report heuristic combines task following, structure, repetition, contradiction, uncertainty, and extreme-length checks:

$$H_{\mathrm{report}} = \mathrm{clip}(0.38 + 0.30H_{\mathrm{task}} + s_{\mathrm{structure}} - 0.75P_{\mathrm{repeat}} - p_{\mathrm{contradiction}} + b_{\mathrm{uncertainty}} - p_{\mathrm{extreme\ length}}).$$

Repetition thus affects both this component and the separate raw-score penalty. Removing presentation controls disables $\widetilde{R}_{\mathrm{report}}$ and P_{repeat} together, so that ablation tests a coupled stabilization bundle.

Optimization roles and rollout settings. Planner and Actor traces train a shared Qwen3-VL checkpoint. Only assistant tokens receive SFT loss; system, user, and tool-observation tokens are masked. The ReportToolkit tool code is not a trainable parameter. The main RL and ablation rollout configuration is:

Setting	Value
GPUs / training steps	3 / 100
Group size	4
Prompt / full-trajectory response / model length	8,192 / 16,384 / 24,576 tokens
Max assistant / user-or-tool-observation turns	20 / 20
Max parallel tool calls per turn	4
Tool-observation limit	2,400 characters, middle truncation
Rollout temperature / top-p / top-k	0.8 / 0.95 / -1
Evaluation temperature / sampling	0 / disabled
Input truncation / overlong-prompt filter	left / enabled
Learning rate	1×10^{-6}
Policy clip / KL loss coefficient	0.2 / 0.001
KL in reward	false
Train batch / mini-batch	12 / 12
PPO epochs per update	1
LoRA rank / alpha	16 / 32
LoRA targets	q/k/v/o and gate/up/down projections
Vision tower	excluded from adaptation

The 16,384-token response budget is shared by the entire generated trajectory—planning, tool calls, intermediate summaries, and final report—rather than allocated per turn. Tool outputs longer than 2,400 Python characters retain the head and tail halves. Task-local evidence pools are separately bounded to six items and 4,000 total characters (Appendix A.13).

SFT configuration. Table 11 gives the SFT setup separately from the RL settings above. The turn-level and workflow-level (workflow-chain) samples are concatenated, deduplicated by exact rendered text, and deterministically shuffled. Before length filtering, the 35,249 samples comprise 78.2% turn-level and 21.8% workflow-chain samples. Removing 360 overlength samples leaves 34,889 training samples.

A.6 Benchmark Evaluation Configuration

Rerun agents use Serper top-5 search and driving-model temperature 0, retaining native framework control flow and stopping rules. Checkpoints are evaluated three times on the same task sets with fixed configured budgets and timeout/retry rules. Live retrieval and benchmark judges contribute to run-to-run variation. Tongyi-DeepResearch-30B is not rerun: its MMDeepResearch-Bench and DeepResearch-Bench II results are taken from the respective benchmark papers (Huang et al., 2026a; Li et al., 2026).

MMDeepResearch-Bench (140 tasks) and DeepResearch-Bench II (132 tasks) use Gemini-2.5-Pro; MiroEval (70 text-only and 30 multimodal tasks) uses GPT-5.1. These benchmark judges are separate from the deterministic training evaluator. The drift diagnostic validates candidate errors against trajectory evidence. Appendix A.12 reports main-checkpoint dispersion.

A.7 Limitations

Task-specific criteria retain binding information, and citation-local coverage and evidence-pool matching provide efficient partial supervision. They cannot fully verify entity–attribute–value pairing or whether a particular cited source entails its adjacent claim. Relation-level verification and controlled entity/value/source-swap tests remain important directions for evaluating and strengthening these signals. The offline task-local corpus limits retrieval breadth and freshness. Furthermore, the training corpora exhibit a strong bias toward images (498 out of 500 tasks involve images), creating a discrepancy with the more diverse attachment types found in MiroEval (Ye et al., 2026); this gap may be a contributing factor to the observed degradation in instruction-following capabilities and the trade-off between ”process-focused” performance and generalization ability.

Table 11: Trajectory SFT configuration. The sample-type ratio is measured before length filtering.

Configuration	Setting
Base model	Qwen3-VL-8B-Instruct
Source traces	3,868 Planner and Actor trace records
SFT sample types	Turn-level tool-action and workflow-chain samples
Turn-level samples	27,559
Workflow-chain samples	7,690
Mixing strategy	Concatenation, exact-text deduplication, and deterministic shuffling; pre-filter ratio 78.2% / 21.8%
Samples before length filtering	35,249
Samples used for training	34,889; 360 overlength samples removed
Maximum sequence length	16,384 tokens
Training epochs	1
Learning rate	2×10^{-5}
LR scheduler / warmup ratio	Cosine / 0.03
Weight decay	0
Per-device batch size	1
Number of GPUs	1
Gradient accumulation	8
Effective global batch size	8
Optimization steps	4,362
Precision	BF16
Distributed training	DeepSpeed ZeRO-3
Gradient clipping	1.0
Gradient checkpointing	Enabled
Parameter-efficient tuning	LoRA, without 4-bit/8-bit quantization
LoRA rank / alpha / dropout	16 / 32 / 0.05
LoRA target modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`
Trainable parameters	43.65M / 8.81B (approximately 0.50%)
Loss mask	Supervised assistant outputs only; prompts, observations, and other context tokens are masked
Random seed	42

Table 12: Repetition evaluation. Lower is better for all four metrics.

DeepResearch Agent	Paragraph Repetition Rate	Sentence Repetition Rate	24-token n-gram Repetition Rate	Severe Repetition Sample Rate
Cog-Research-8B	3.98%	9.54%	7.88%	13.09%
Cog-Research-8B-SFT	4.23%	10.76%	9.71%	19.91%
Cog-Research-8B-SFT-RL	0.92%	4.62%	3.30%	5.93%
w/o Presentation / repetition controls	15.18%	29.46%	24.11%	25.89%

A.8 Repetition Metrics and Results

Observed repetition. Table 12 shows that severe repetition falls from 19.91% after SFT to 5.93% after joint RL, and rises to 25.89% when presentation/repetition controls are removed. Paragraph, sentence, and 24-token n-gram repetition rates show the same pattern. These results identify output stabilization as a substantial contribution of the controls; they do not isolate the effect of any one penalty or establish relation-level grounding correctness.

Normalization and metrics. Reports are NFKC-normalized. Citation IDs, URL destinations, Markdown/HTML decoration, and punctuation-only differences are removed. Nontrivial paragraphs contain at least 40 normalized characters and sentences at least 30; token n-grams use $n = 24$. For unit type $u \in \{p, s, n\}$, let $c_u(z)$ count unit z. The duplicate-extra rate is

$$r_u = \frac{\sum_z \max(0, c_u(z) - 1)}{\sum_z c_u(z)}, \qquad m_u = \max_z c_u(z),$$

with zero rates/counts for an empty unit set. A report is severe if any of the following holds:

$$\begin{aligned} & r_p > 0.25 \quad \vee \quad m_p > 5 \\ & \vee \ (r_s > 0.35 \wedge m_s > 10) \quad \vee \ (r_s > 0.18 \wedge m_s > 30) \\ & \vee \ (r_n > 0.20 \wedge m_n > 20). \end{aligned}$$

The soft penalty is

$$\begin{aligned} P_{\text{repeat}} = \min\big(0.6,\, & 0.65 r_p + 0.15 r_s + 0.10 \min(m_p/10, 1) \\ & + 0.06 \min(m_s/25, 1) + 0.04 \min(2 r_n, 1)\big). \end{aligned}$$

The severe-sample rate is the fraction of reports satisfying the Boolean rule. The normalization and thresholds correspond to `cosight_rl/rewards/repetition.py`; the severe indicator and continuous penalty are distinct quantities.

A.9 Broader Error Distributions

To test whether Grounding Drift is specific to the Cog-Research base run, we apply the same four-way error taxonomy to additional agents. The open-source set contains WebThinker-R1-7B, MS-Agent-8B, Vision-DeepResearch-8B, MiroThinker-v1-8B, Co-Sight-8B, Co-Sight-30B-A3B, and Co-Sight-235B-A22B. The closed-source set contains o3-mini, Grok-4, MiroThinker-1.7, Qwen3-VL-Plus, Gemini-2.5-Pro, and GPT-5. These figures are supplementary primary-error distributions; their denominators are separate from the corrected 111/100/81 attributable-error summaries.

A.10 Compute and Evaluation Cost

SFT uses one A100 80GB GPU with the configuration in Table 11; the main RL run takes approximately 13.32 hours on three A100 80GB GPUs (39.96 GPU-hours). Offline evidence retrieval avoids live-web latency during RL. Running Cog-Research on MMDeepResearch-Bench was reported to take 20 h 14 min on one A100 80GB, with approximately US$7 in Serper cost. Reported benchmark-judge passes cost approximately US$70 for MMDeepResearch-Bench, US$40 for DeepResearch-Bench II, and US$60 for MiroEval. API costs are aggregate estimates.

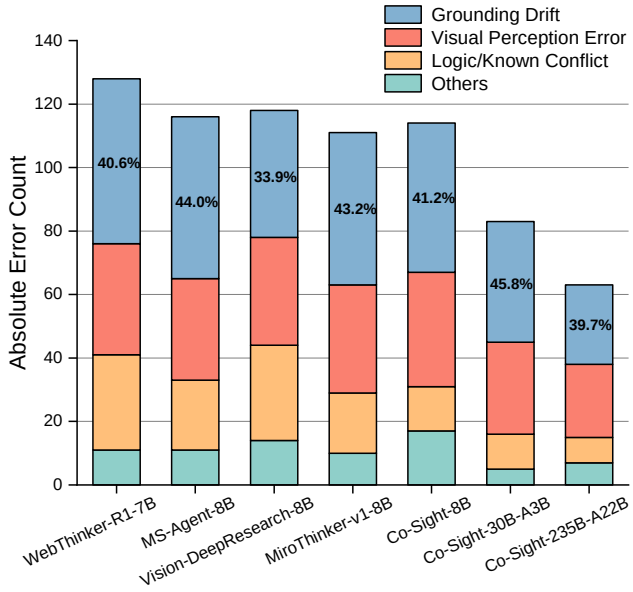


Figure 8: Primary-error distribution for additional open-source research agents.

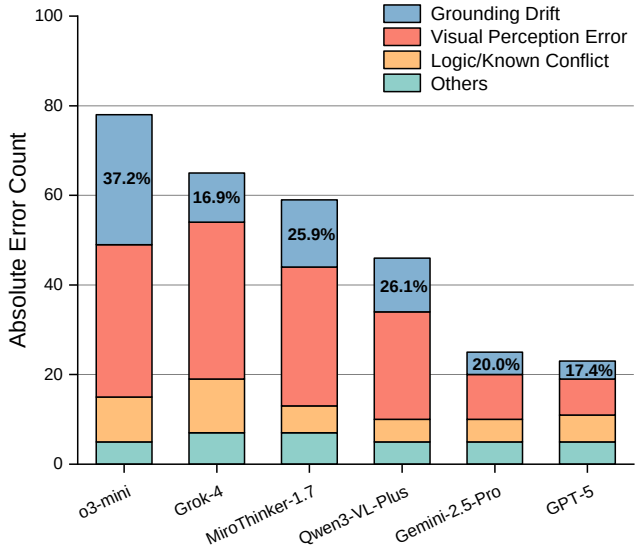


Figure 9: Primary-error distribution for additional closed-source research agents.

Table 13: Research-process diagnostics on MMDeepResearch-Bench (higher is better). These are task-level proxies.

DeepResearch Agent	Visual QA	Search Info	Plan DAG	Tool Success	Evidence Recovery
Cog-Research-8B	46.43	78.57	84.15	76.43	60.13
Cog-Research-8B-SFT	46.71	70.00	80.73	77.14	64.27
Cog-Research-8B-SFT-RL	50.00	72.86	82.70	81.43	66.50

A.11 Atomic Capability Retention Metrics

Table 13 shows that the report-grounding gains do not come from a uniform increase in every atomic capability. The final checkpoint improves Visual QA, Tool Success, and Evidence Recovery over the base model, while Search Info and Plan DAG remain slightly lower. A possible explanation is that imitation of teacher artifacts and training within small task-local evidence pools favor evidence reuse over broader exploration. Importantly, RL recovers part of the SFT-stage decline across all five measures. Cog-Research improves report grounding while largely retaining the underlying visual, search, planning, and tool-use behaviors, rather than simply becoming a stronger agent on every low-level metric.

Visual QA. The unit is one MMDeepResearch-Bench task. We evaluate whether the strongest non-final visual-analysis artifact contains sufficiently specific task-level textual visual grounding to identify the correct task-level visual truth among the 140 candidates. Candidate artifacts are selected deterministically using visual terms in filenames/content and explicit input-image IDs; search-result artifacts receive negative selection weight.

Search Info. The unit is one task. A deterministic URL/artifact validator checks whether retrieval produced substantive, traceable evidence. A task passes when it contains at least one substantive artifact with a valid external HTTP(S) URL, at least two distinct normalized URLs, and at least two distinct source hosts.

Plan DAG. The unit is an actionable requirement clause, macro-averaged within and then across tasks. Using the recorded `Planner_trace.jsonl`, requirement-to-artifact coverage is estimated with character 2–5-gram TF–IDF cosine similarity together with DAG acyclicity/dependency checks.

Tool Success. The unit is one task. The Report Tool succeeds when a timestamped final Markdown file exists, contains at least 1,000 non-markup characters, and has no crash signature. If several reports exist, the latest valid timestamped report is used downstream.

Evidence Recovery. The unit is a normalized final-report source URL, macro-aggregated by task. The union of normalized HTTP(S) URLs in non-final workspace artifacts forms the task-local source set; the metric measures whether URLs used by the final report can be recovered from that upstream evidence set. URL normalization lowercases hosts, removes `www.`, query strings/fragments, duplicate/trailing slashes, and canonicalizes HTTP/HTTPS.

A.12 Run-to-Run Variability of the Main Checkpoints

Table 14 reports mean Overall score and standard deviation over three evaluation runs of each checkpoint. Variation includes inference-backend nondeterminism, external judges, and live retrieval. These repeats measure evaluation variability for fixed checkpoints. Base-to-SFT-RL gains exceed

Table 14: Overall score (mean $\pm$ standard deviation over three evaluation runs).

Checkpoint	MMDeepResearch-Bench	DeepResearch-Bench II	MiroEval-Multimodal	MiroEval-Text
Base	23.22 ± 0.28	26.93 ± 0.15	48.36 ± 0.33	50.38 ± 0.22
SFT	25.61 ± 0.51	28.57 ± 0.31	52.44 ± 0.48	54.01 ± 0.34
SFT-RL	28.07 ± 0.46	30.47 ± 0.29	53.27 ± 0.47	54.13 ± 0.30

the observed run-to-run standard deviations on all four evaluations. The smaller SFT-to-RL gain on MiroEval-Multimodal (0.83 points) should be read alongside standard deviations of 0.48 and 0.47.

A.13 Offline Evidence Corpus Construction and Access

The Evidence Corpus stores teacher tool observations and intermediate workspace artifacts by task ID. It contains 4,597 items for 500 tasks and supports on-policy rollouts in a fixed retrieval environment. The offline environment and evidence-access settings below apply to the main joint-RL

run and the training-strategy ablations. Raw input images remain enabled during these RL runs. At inference and benchmark evaluation, the agents use their actual tools in a live web environment rather than this fixed retrieval corpus.

Workspace candidates. The builder, `scripts/build_cosight_evidence_corpus.py`, recursively reads Markdown files under task-specific workspace directories. Documents require at least 80 characters. Filenames indicating a final report, including `final_report`, `final-answer`, their Chinese equivalent, and timestamped report names, are excluded heuristically. Candidate ranking adds 5 for a URL, 2 per matched evidence/search/source/verification/visual-analysis filename cue, and 1 for content longer than 500 characters, with a -20 adjustment for suspected final reports. At most ten workspace documents are retained per task.

Trace observations. The builder retains only `role=tool` observations, with at most four per task. Supported tools are `ask_question_about_image`, `serper_search`, `fetch_website_content`, `extract_document_content`, `file_read`, and `execute_code`, `mark_step`, `file_saver`, `generate_markdown_report`. Ordinary search observations must contain a URL; visual analysis, file reading, and document extraction may be URL-free. Assistant final answers are not retained as trace observations. Planner/Actor records provide task association, while the indexed factual contents come from eligible tool results and workspace documents.

Cleaning, provenance, and deduplication. Each item has Markdown image markers and special model tokens removed, redundant whitespace collapsed, and text truncated to 2,400 characters. URLs are extracted and deduplicated. The record stores `id`, `task_id`, `source_type`, `source_ref`, `title`, `urls`, `text`, and `char_count`. The source type is principally `workspace_md` or `trace_tool`. A SHA-1-derived ID includes task, source, and content; deduplication uses the pair (task ID, evidence ID). Identical text at distinct file/trace locations may remain because source references contribute to identity. Deduplication therefore preserves distinct source locations.

Table 15: Offline Evidence Corpus manifest. Source counts are before deduplication; mean items per task is computed after deduplication (4,597/500).

Statistic	Count
Tasks with indexed evidence	500
Workspace items	2,626
Trace items	1,984
Total before deduplication	4,610
Total after deduplication	4,597
Items per covered task (min / max / mean)	4 / 14 / 9.19

The builder emits `evidence_corpus.jsonl` (flat records), `task_evidence.json` (task-indexed records), and `manifest.json` (parameters and statistics).

Task-local selection. Data preparation resolves a training row's `qid` or `id` to its task pool, then partitions items into four groups: URL-bearing trace items, URL-bearing workspace items, URL-free trace items, and URL-free workspace items. Round-robin selection balances these groups; within a group, more URLs and shorter content are preferred. The shared RL configuration uses

```
COSIGHT_MAX_EVIDENCE_ITEMS=6
COSIGHT_MAX_EVIDENCE_CHARS=4000
```

Thus each selected pool has at most six items and 4,000 total evidence characters, retaining at most five URLs *per item*. Selection is implemented in `evidence_for_record` in `scripts/prepare_cosight_rl_data.py`.

Raw images and stored visual observations. Raw input images are retained in the main RL run and the Outcome-only and Process-only ablations; the offline tool adapter does not disable image input. The vision tower is excluded from parameter adaptation, which freezes its weights rather than removing its visual inputs. Stored textual visual observations provide additional retrievable evidence. They are recognized from title/source cues such as image QA, image analysis, visual inspection, OCR, and chart analysis, and tagged `frozen_visual_observation`; here "frozen" refers to a stored observation. If no visual item is explicitly identified, a URL-free trace/workspace item is used as fallback; this fallback is not guaranteed to describe visual content. Preparation aborts if visual-observation coverage among image-bearing tasks is below 95%.

Hidden pool and action-dependent retrieval. With `COSIGHT_INCLUDE_EVIDENCE_IN_PROMPT=0`, the task-local evidence pool is kept inside the `cosight_step` environment state rather than appended to the user prompt. The policy actively accesses it with `search_evidence` and `inspect_visual`, which are training-time adapters to the offline environment. The source-tool names listed above identify historical observations indexed by the corpus, rather than live web calls during RL. For a generated query q and item e, retrieval uses

$$\mathrm{score}(q, e) = \frac{|T(q) \cap T(e)|}{\max(1, |T(q)|)},$$

where English text uses word tokens and Chinese text uses character bigrams. Search is restricted to the preselected pool for the current task, with default top-$k = 3$ and maximum 5. It uses neither BM25, embeddings, nor live web search. Results contain an evidence ID, title, URLs, text, and match score. `inspect_visual` returns items marked as frozen visual observations. The action interface is implemented in `cosight_rl/tools/trajectory_tool.py`.

Workspace provenance constraint. When `save_step` records an intermediate conclusion, it may name evidence IDs such as `E1` and `E3`. The environment checks that those IDs have already been returned to that rollout by search or visual inspection; unretrieved IDs cannot be saved as supporting evidence. `read_workspace` exposes saved work, and `submit_report` terminates report production. This enforces retrieval before citation.

Evidence versus evaluator supervision. The Evidence Corpus supplies retrievable facts and observations; dimension/step rubrics remain hidden evaluation criteria. The main joint-RL run, Outcome-only, and Process-only use the same evidence pool, raw-image input setting, and Plan-and-Act environment. Outcome-only evaluates the report with dimension rubrics; process-only evaluates plans and intermediate steps with step rubrics. Fixed evidence makes the environment deterministic conditional on actions, while newly sampled plans, queries, and summaries still make the policy rollouts on-policy.

A.14 Training-Strategy Ablation Protocol

The training-strategy ablations in Table 3 are designed to isolate supervised initialization and the two trajectory-level RL reward branches. Unless stated otherwise, the RL variants use the same Plan-and-Act environment, action and tool space, task-local Evidence Corpus, rollout budgets, sampling configuration, GRPO group size, optimizer settings, and training schedule as the full model. The shared optimization configuration is reported in Appendix A.5, and evidence construction and access are described in Appendix A.13.

Outcome-only (w/o process reward). This variant is initialized from the same trajectory-SFT checkpoint as the full model and performs the same interactive research rollouts, including planning, evidence retrieval, tool use, workspace updates, and final-report generation. Its trajectory reward is

$$r_k = R_{\mathrm{report},k},$$

which is equivalent to setting $\lambda = 0$ in the joint objective. The process branch R_{step}, including plan quality, matched-step execution, evidence/artifact preservation, and dependency-consistency terms, does not contribute to the GRPO advantage. Report-dimension rubrics, report-level penalties, and hard-error caps remain active.

Process-only (w/o outcome reward). This variant is also initialized from the trajectory-SFT checkpoint and uses the same rollout environment and evidence-access protocol. Its trajectory reward is

$$r_k = R_{\mathrm{step},k},$$

equivalent to setting $\lambda = 1$. The report-quality branch R_{report}, including the five report dimensions and associated report-quality penalties and caps, does not contribute to the GRPO advantage.

The policy nevertheless generates and submits a final report. Report completion contributes only through the binary final-report indicator in $\overline{S}_{\mathrm{step}}$ and the corresponding missing-report constraint; the semantic quality of the report is not scored by the report-outcome branch. Thus, "process only" removes report-quality supervision rather than report generation.

Controlled comparison. Consequently, the outcome-only and process-only experiments differ from the full RL configuration primarily in which reward branch contributes to the group-relative advantage. Their retrieval corpus, tool interface, rollout interaction pattern, and optimization budget are otherwise held fixed.