INFORMATION-GAIN CALIBRATED ON-POLICY DISTILLATION FOR DEEP SEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) has recently emerged as a promising approach for transferring the capabilities of large agents to compact students. In deep search, where agents iteratively retrieve external evidence, a teacher's preference among plausible queries before retrieval may not predict which queries will return useful evidence. Standard OPD nevertheless supervises each student query from the preceding context alone, inducing a failure mode we term ***evidence-blind supervision (EBS)***. To understand this, we regroup the OPD objective by interaction turns and reveal two implicit priors: a *budget prior* that allocates supervision by token volume, and a *direction prior* that favors tokens preferred by the teacher, neither of which accounts for the evidence returned by the current query. Our empirical analysis shows that both priors contribute to EBS in practice: the first search turn accounts for 51.1% of total absolute information gain but receives only 18.0% of the supervision budget, while 43.7% of search queries have teacher rewards that disagree in sign with post-retrieval information utility. To mitigate this, we propose `IG-OPD`, which uses post-retrieval information gain (IG) to calibrate both priors: reallocating supervision toward information-critical turns while calibrating query preferences using the utility of retrieved evidence. Experiments on six multimodal search benchmarks show that `IG-OPD` improves average accuracy by 3.4 and 1.7 percentage points for 2B and 4B students over standard OPD, enabling the 4B student to surpass its 8B teacher on MMSearch and BrowseComp-VL.

1 INTRODUCTION

Real-world queries often require knowledge beyond a model's pretrained parametric memory, such as time-sensitive facts and long-tailed entities (Lewis et al., 2020; Mallen et al., 2023; Jiang et al., 2024). To address this, search agents equip Large Multimodal Models (LMMs) with external search tools to iteratively reason over retrieved evidence. However, acquiring these capabilities typically requires large-scale agentic reinforcement learning (RL) on heavyweight models (e.g., MMSearch-R1 (Wu et al., 2025), OpenSearch-VL (Chen et al., 2026), and HyperEyes (Li et al., 2026b)), making the agents expensive to train and deploy. Thus, on-policy distillation (OPD) offers a natural route to transfer such RL-acquired capabilities to compact students by supervising the student's own rollouts with dense teacher signals derived from a reverse-KL objective. Previous work on OPD focused mainly on single-turn settings such as mathematics and coding, emphasizing stabilizing training under high-variance rewards (Hao et al., 2026; Xie et al., 2026; Jin et al., 2026) or combining OPD with outcome-based RL (Tan et al., 2026; Xu et al., 2025). Recent work has extended OPD to multi-turn settings, such as embodied reasoning and web navigation, focusing primarily on mitigating error accumulation across interaction turns (Wang et al., 2026b; Li et al., 2026a; Zhou et al., 2026a).

Despite their success, these approaches typically derive their supervision of each student action using the preceding context alone, *without* accounting for the subsequent environmental feedback. This consequence-agnostic supervision becomes problematic in deep search, where we observe a recurring failure mode: a search question may admit several plausible queries that differ in wording, language, or retrieval path, but their usefulness still depends on the evidence they actually retrieve. Even a query preferred by the teacher might yield little useful evidence, while a less favored one could retrieve the evidence needed to answer the question. Yet standard OPD still scores the student's query using teacher–student token likelihoods conditioned only on the pre-retrieval context, so the resulting score may not reflect the query's realized utility. It can therefore penalize a useful

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.

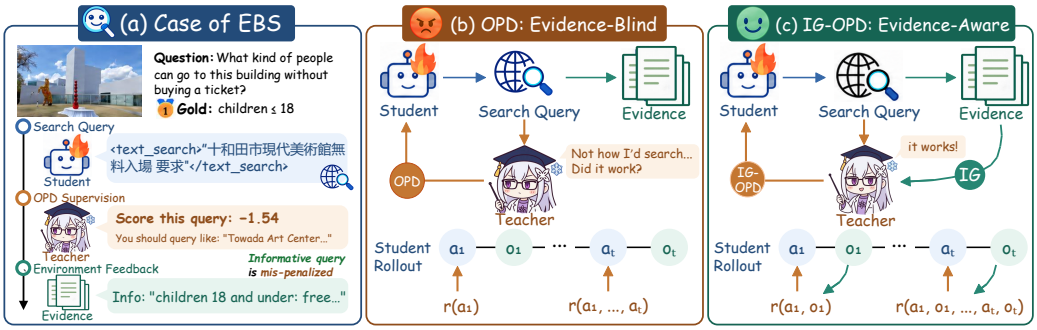


Figure 1: Illustration of **evidence-blind supervision** in OPD for search agents. **(a) EBS example** shows that a useful search query is mis-penalized even though it subsequently retrieves the gold evidence. **(b) Standard OPD** supervises the agent's actions without considering the returned evidence. **(c) IG-OPD** enables evidence-aware supervision by incorporating post-retrieval information gain.

query simply because the teacher preferred a different one. Figure 1 illustrates a case: the student correctly identifies the "Towada Art Center" and issues a Japanese query about free admission. Standard OPD still scores it at -1.54, favoring the teacher's wording even though the teacher's query retrieves no useful evidence. The student's query, by contrast, retrieves the gold evidence, yet the supervision ignores it. We refer to this failure mode as **evidence-blind supervision (EBS)**.

To understand this issue, we regroup the OPD objective by interaction turns, revealing two implicit priors that govern supervision allocation and update direction (see Appendix A.1 for the derivation):

$$J_{\text{OPD}} = \mathbb{E}_{\tau\sim\pi_S} \sum_{t=1}^{T} \underbrace{\frac{n_t}{N}}_{\text{budget prior}} \cdot \frac{1}{n_t} \sum_{i\in\text{turn}_t} \underbrace{\log \frac{\pi_T}{\pi_S}}_{\text{direction prior}} \tag{1}$$

where n_t is the token count of turn t, $N = \sum_t n_t$ is the trajectory length, and the log-ratio between teacher π_T and student π_S is evaluated at each student-sampled token. Under this formulation, (i) the *budget prior* n_t/N allocates each turn's share of supervision by raw token volume, so a verbose synthesis turn can outweigh a concise search turn even when the latter contributes more to finding the answer. Our empirical analysis in Section 4.1 confirms this effect: the first search turn accounts for 51.1% of the total information criticality, yet receives only 18.0% of OPD's token budget. Meanwhile, (ii) the *direction prior* $\log \frac{\pi_T}{\pi_S}$ favors queries closer to the teacher's preference, regardless of whether they retrieve useful evidence. Empirically, we find that roughly 43.7% of query supervision is inconsistent with post-retrieval information utility. Together, these limitations naturally motivate a key question: *How can we seamlessly integrate information available from post-retrieval evidence into the OPD process without distorting or destabilizing training dynamics?*

Building upon these, we introduce information gain (IG) as a lightweight post-retrieval signal to mitigate the two priors in EBS. Therefore, we propose **I**nformation-**G**ain Calibrated **O**n-**P**olicy **D**istillation (`IG-OPD`), a novel OPD framework for evidence-aware supervision. Specifically, `IG-OPD` employs Information-Guided Turn-Credit Calibration (ITC) to adjust the *budget prior* by information criticality, and Information-Guided Query-Preference Calibration (IQC) to reweight the *direction prior* by information utility. Extensive experiments in a live web environment demonstrate that `IG-OPD` improves average accuracy by +3.4 and +1.7 for 2B and 4B students, respectively. It also boosts the gold-evidence hit rate by 14.7% relatively, with the 4B student surpassing its 8B teacher on MMSearch and BrowseComp-VL. To summarize, our main contributions are as follows:

- We identify and formalize **EBS**, a systematic failure mode of OPD in deep search agents where supervision is assigned without considering retrieved evidence, showing how supervision assigned before retrieval can under-credit information-critical turns and penalize useful queries.
- We propose `IG-OPD`, an evidence-aware OPD framework that uses post-retrieval information to calibrate OPD's two priors through complementary mechanisms: **ITC** reallocates cross-turn budgets by information criticality, while **IQC** calibrates query preferences with evidence utility.
- Across six multimodal search benchmarks, `IG-OPD` improves average accuracy over OPD by 3.4 and 1.7 percentage points for the 2B and 4B students, respectively. The 4B student also surpasses its 8B teacher on MMSearch and BrowseComp-VL, with zero inference overhead.

2 Related Work

Multimodal search agents. Multimodal search agents extend question answering beyond parametric memory by retrieving related web evidence. Recent systems mainly use reinforcement learning (RL) to optimize multi-step retrieval and reasoning (Zhang et al., 2025; Lin et al., 2025; Hong et al., 2025). For example, MMSearch-R1 and WebWatcher train search policies with GRPO (Wu et al., 2025; Geng et al., 2026), while OpenSearch-VL combines multi-hop training data with safeguards against cascading tool errors (Chen et al., 2026). However, sparse outcome rewards complicate credit assignment and make training costly, motivating more sample-efficient supervision.

On-policy distillation. On-policy distillation (OPD) provides dense teacher feedback on student-generated trajectories, offering a sample-efficient approach to capability transfer. Existing studies mainly address three vulnerabilities in OPD. **Training instability** arises from high-variance distillation rewards, motivating reward normalization (Hao et al., 2026), reward clipping (Ko et al., 2026), and proximal teacher distributions (Xie et al., 2026). **Indiscriminate imitation** overlooks differences in uncertainty and supervision quality across tokens and trajectories. To solve this, related methods introduce entropy-aware objectives (Jin et al., 2026), token selection (Huang et al., 2025; Xu et al., 2026b), or selective filtering and reweighting (Li et al., 2026c; Zheng et al., 2026). In multi-turn agents, **compounding errors and distribution shift** further undermine teacher guidance as early mistakes alter subsequent interactions (Zhou et al., 2026b; Liao et al., 2026). TCOD progressively extends the distillation horizon (Wang et al., 2026b), while teacher intervention steers trajectories toward more reliable guidance (Li et al., 2026a). SOD and SAGE instead adapt supervision across steps or turns to mitigate unreliable signals on diverging histories (Zhong et al., 2026; Zhou et al., 2026a). Besides, other approaches **combine distillation with RL objectives** (Xu et al., 2025; 2026a; Zhang et al., 2026), with ATOD gradually shifting from OPD to RL (Tan et al., 2026).

While these strategies refine supervision delivery from various angles, they fundamentally suffer from evidence blindness in multi-hop search: supervising student actions without considering the post-retrieval utility of generated queries, ultimately giving rise to EBS. To overcome this, `IG-OPD` leverages information gain as a post-retrieval signal to calibrate both turn budgets and update directions, effectively rectifying the misalignment between OPD priors and realized evidence utility.

3 Preliminaries

3.1 Problem Formulation: Multi-Turn Search Agents

Following Wang et al. (2026a), we consider a dataset $\mathcal{D} = \{(x, y^\star)\}$, where each query $x = (q, v)$ pairs a textual question q with an optional image v, and $y^\star$ is the gold answer. Starting from x, the agent iteratively alternates between internal deliberation and tool invocation (e.g., `text_search`, `crop`) until concluding with a final response, yielding a multi-turn rollout $\tau = (\tau_1, \ldots, \tau_T)$. Each intermediate turn $\tau_t = (h_t, a_t, o_t)$ $(t < T)$ encapsulates a chain of thought h_t in `<think>`, a tool call a_t in `<tool_call>`, and the resulting evidence o_t. For the final turn $\tau_T = (h_T, \hat{y})$, the search agent will organize and reason over collected evidence to produce a final synthesis response $\hat{y}$ wrapped in `<response>`. Because observations are externally produced by the environment, training losses are computed exclusively on tokens generated by the student policy. To better formalize `IG-OPD` in the next section, we denote H_t as the cumulative interaction history after turn t, initialized with $H_0 = x$. Moreover, we use $\mathcal{I}_t$ to index the generated tokens in turn t, with $n_t = |\mathcal{I}_t|$ and total trajectory length $N = \sum_{t=1}^{T} n_t$. For each token $i \in \mathcal{I}_t$, $u_{t,i}$ denotes the sampled token and $z_{t,i}$ its complete conditioning prefix, which incorporates all preceding context and observations.

3.2 On-Policy Distillation

On-policy distillation (OPD) transfers knowledge from a frozen teacher π_T to a student π_S with parameters θ on the student's own rollouts $\tau \sim \pi_S(\cdot \mid x)$. Here, the teacher supervises on the student's state distribution rather than on fixed demonstrations, thereby eliminating the exposure gap of offline distillation. The objective of OPD is the reverse KL divergence $D_{\mathrm{KL}}\big(\pi_S(\cdot \mid z_{t,i}) \,\|\, \pi_T(\cdot \mid z_{t,i})\big)$ averaged over student-generated tokens. Following standard practice in large-scale agent distillation (Wang et al., 2026b), we instantiate the objective using the sampled token log-likelihood ratio:

$$\widehat{\mathcal{L}}_{\mathrm{OPD}}(\theta) = -\,\mathbb{E}_{x\sim\mathcal{D},\,\tau\sim\pi_S}\Big[\frac{1}{N}\sum_{t=1}^{T}\sum_{i\in\mathcal{I}_t} D_{t,i}^{\mathrm{OPD}}\Big], \qquad D_{t,i}^{\mathrm{OPD}} \equiv \log\frac{\pi_T(u_{t,i} \mid z_{t,i})}{\pi_S(u_{t,i} \mid z_{t,i})}. \tag{2}$$

Regrouping Eq. 2 by interaction turns recovers the objective decomposition in Eq. 1, revealing two implicit priors. Crucially, $D_{t,i}^{\mathrm{OPD}}$ depends solely on the prefix $z_{t,i}$ and never consults the observation o_t returned by turn t, giving rise to the failure mode of EBS. To mitigate this, we show how to leverage realized post-retrieval evidence to dynamically calibrate this OPD objective in Section 4.

4 Information-Gain Calibrated On-Policy Distillation

Eq. 1 shows that vanilla OPD determines both *how much* each turn is supervised (n_t/N) and *in which direction* tokens are pushed ($\log \frac{\pi_T}{\pi_S}$) via quantities fixed before search execution. `IG-OPD` preserves the dense teacher signal while calibrating both priors with a posterior signal available upon observation return. We first define this post-retrieval information gain g_t and elaborate on the practical effects of EBS (Sec. 4.1), then recalibrate turn supervision budgets (ITC, Sec. 4.2) and adjust query update directions (IQC, Sec. 4.3). Finally, Sec. 4.4 synthesizes the overall objective.

4.1 Diagnosing Evidence-Blind Supervision with Information Gain

Post-retrieval information gain. We quantify the value of a search turn by how much its evidence changes belief about the gold answer, following IGPO (Wang et al., 2026a). Given the interaction history H_t after the t-th observation, we compute length-normalized log-likelihood of the reference answer $y^\star = (y_1^\star, \ldots, y_L^\star)$ under the teacher and convert it into a stabilized confidence potential:

$$\bar{\ell}_t = \frac{1}{L}\sum_{j=1}^{L} \log \pi_T\big(y_j^\star \mid H_t, y_{<j}^\star\big), \qquad \Phi_t = \log\big(\exp(\bar{\ell}_t) + \epsilon\big), \tag{3}$$

where $\epsilon = 10^{-3}$ prevents numerical instability in the logarithm as token probabilities approach zero. The information gain of turn t is defined as the potential change induced by its observation: $g_t = \Phi_t - \Phi_{t-1}$ for $t < T$, with Φ_0 evaluated on the initial query $H_0 = x$. Intuitively, a higher g_t indicates that the evidence conveys more valuable information. Moreover, these turn-level gains telescope to the net confidence change, anchoring the calibration against collapse (Ng et al., 1999).

Empirical Diagnosis of EBS. Standard OPD assigns its supervision before evidence arrives, naturally prompting us to ask: *How does EBS manifest in practice, and how severely do the two priors mismatch the posterior information utility?* To examine this, we roll out a student agent (Qwen3-VL-2B-Instruct (Bai et al., 2025)) on the MMSearch benchmark (Jiang et al., 2024), collecting 300 multi-turn trajectories spanning roughly 1,300 interaction turns. For each turn, we compute both the dense reverse-KL OPD reward scored by the teacher (OpenSearch-VL-8B (Chen et al., 2026)) and its post-retrieval information gain g_t. Notably, we centralize the OPD rewards by subtracting their mean across candidate rollouts to offset the negative bias inherent to sampled RKL.

To assess the *budget prior*, we treat each turn's magnitude $|g_t|$ as its information criticality and compare its normalized share against the default budget ratio n_t/N, thus revealing a **Turn-level EBS**. As shown in Figure 2 (Top), the first search turn accounts for 51.1% of the total $|g_t|$ on average, yet receives only 18.0% of the supervision budget, yielding just 0.35 units of supervision per unit of information. In contrast, the final response turn, which primarily synthesizes and organizes previously collected evidence, absorbs 47.6% of the budget despite contributing zero information gain. These observations imply that informative turns can suffer from supervision starvation, while verbose turns dominate update gradient.

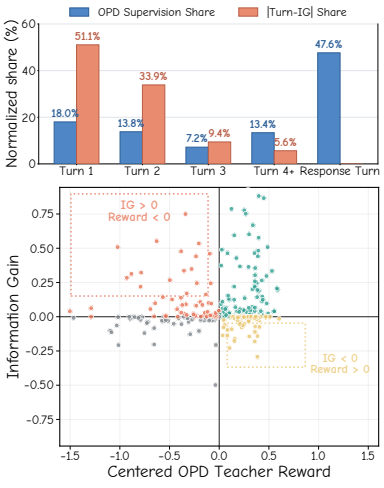


Figure 2: Empirical analysis of EBS under standard OPD's budget prior (Top) and direction prior (Bottom).

To assess the *direction prior*, we treat each query's realized signed g_t as an **empirical** post-retrieval information utility and contrast it with the centered OPD reward, uncovering a severe **Query-level EBS**. As illustrated in Figure 2 (Bottom), roughly 43.7% of the search turns fall into the second and fourth quadrants, where the prior OPD reward and post-retrieval g_t disagree in sign. Specifically, 19.8% of search queries retrieve useful evidence ($g_t > 0$) yet are penalized, while 23.9% retrieve misleading evidence ($g_t < 0$) yet are rewarded. These results show that **in-**

herent teacher preference prior to retrieval is not a faithful proxy for actual evidence utility, causing beneficial search actions to be penalized while misleading ones are falsely reinforced.

These empirical findings naturally motivate us to use IG as a post-retrieval signal and propose `IG-OPD` to address each symptom in turn, using $|g_t|$ to recalibrate the budget (ITC) and the sign of g_t to recalibrate the direction (IQC). We will also discuss the reliability of IG g_t in Section 5.5.

4.2 Information-Guided Turn-Credit Calibration

The budget prior n_t/N in Eq. 1 implicitly assumes the longest turn to dominate the distillation. In deep search, the final response turn, which merely consolidates collected evidence, is typically far longer than the search turns that acquired it. Consequently, the critical search behavior receives the least supervision, dictated by token volume rather than information value. ITC replaces this length-induced weight with an information-induced one by first normalizing each turn to remove token-length bias, and then modulating it using the turn's posterior information magnitude:

$$q_t = \frac{1 + \lambda \tanh(|g_t|/\sigma_g)}{\sum_{j=1}^{T} [1 + \lambda \tanh(|g_j|/\sigma_g)]}, \qquad \sum_{t=1}^{T} q_t = 1, \tag{4}$$

where σ_g is the batch root-mean-square and λ controls how strongly IG magnitude reshapes the budget. The $\tanh$ prevents a single extreme-gain turn from monopolizing the budget, while the additive constant guarantees that every turn, including the response, retains a baseline share of teacher supervision, so answer synthesis remains distilled rather than discarded. Setting $\lambda = 0$ recovers a uniform per-turn budget $q_t = 1/T$, which we use as an equal-turn ablation; our default is $\lambda = 0.1$.

To implement q_t without changing the trainer's existing token-level loss reduction, we rescale each sampled token's OPD reward by the ratio between its calibrated and original turn-level shares:

$$A^{\mathrm{ITC}}_{t,i} = \frac{N q_t}{n_t} D^{\mathrm{OPD}}_{t,i}, \qquad \text{so that} \qquad \frac{1}{N}\sum_{t=1}^{T}\sum_{i\in\mathcal{I}_t} A^{\mathrm{ITC}}_{t,i} = \sum_{t=1}^{T} q_t \Big(\frac{1}{n_t}\sum_{i\in\mathcal{I}_t} D^{\mathrm{OPD}}_{t,i}\Big). \tag{5}$$

Compared to Eq. 1, ITC leaves the direction prior untouched and replaces only the budget prior n_t/N with q_t. ITC deliberately uses the *absolute* gain: whether helpful or harmful, a large $|g_t|$ means the turn mattered and merits more supervision, while its sign is handled separately by IQC.

4.3 Information-Guided Query-Preference Calibration

Within a search turn τ_t, the direction prior ($\log \frac{\pi_T}{\pi_S}$) is computed before evidence arrives, pushing every token toward the teacher's wording. Therefore, a student query that brings decisive evidence can still be suppressed. IQC mitigates this by using the signed gain g_t as a post-retrieval calibration signal. The correction applies only to action-critical tokens (e.g., tool names, arguments, and the query string), excluding formatting tokens like `<tool_call>`. We denote this set by $\mathcal{M}_t \subseteq \mathcal{I}_t$ with $m_t = |\mathcal{M}_t|$. Since g_t is a property of the whole tool call rather than individual tokens, we treat a_t as one macro-action with realized utility g_t and split the correction evenly over $\mathcal{M}_t$:

$$A^{\mathrm{IQC}}_{t,i} = \mathbf{1}[i \in \mathcal{M}_t]\,\frac{\alpha_k\, g_t}{m_t}, \qquad \sum_{i\in\mathcal{M}_t} A^{\mathrm{IQC}}_{t,i} = \alpha_k\, g_t. \tag{6}$$

The second identity ensures that the total correction assigned to a search decision equals its scaled realized gain, regardless of the exact number of tokens used to express the action. This uniform allocation operates directly at the macro-action level, ensuring robust credit assignment across the entire tool invocation while preventing length-induced token biases. Accordingly, a positive g_t increases the OPD signal on the decision tokens, whereas a negative g_t decreases it, correcting the teacher's pre-retrieval preferences when they disagree with the utility observed after retrieval.

Cosine-ramped evidence coefficient. Early in training, as the student explores search space, raw posterior signals exhibit high natural variance. Rather than over-fitting to early noisy rollouts, we introduce a cosine schedule (further analyzed in Section 5.5) to smoothly transition the student from initial exploration to evidence-calibrated refinement over steps $k = 0, \ldots, K$:

$$\alpha_k = \frac{\alpha_{\max}}{2}\Big[1 - \cos\Big(\pi\,\frac{k}{K}\Big)\Big], \tag{7}$$

Power-posterior interpretation. The additive form of Eq. 6 is not an arbitrary reward bonus. Consider an idealized sequence-level OPD objective, where the unmodified per-trajectory reward is $\log p_T(\tau \mid x) - \log p_S(\tau \mid x)$, and write $\Psi_t = \exp(\Phi_t)$. By telescoping, adding $\alpha_k \sum_t g_t$ yields:

$$\log p_T(\tau \mid x) - \log p_S(\tau \mid x) + \alpha_k \sum_t g_t = \log \frac{p_T(\tau \mid x)\, \Psi_{T-1}(\tau)^{\alpha_k}}{p_S(\tau \mid x)} + \text{const}, \tag{8}$$

i.e., the student is distilled toward a generalized power posterior $\tilde{p}_{T,k}(\tau \mid x, y^\star) \propto p_T(\tau \mid x)\, \Psi_{T-1}(\tau)^{\alpha_k}$ that tilts the teacher's prior over trajectories by a confidence pseudo-likelihood of the gold answer, with α_k acting as a likelihood power. IQC can therefore be read as distilling from a teacher whose preferences have been updated by the evidence the student actually retrieved. Our practical objective departs from this idealized form in three ways: the budget is calibrated by ITC, Ψ_t is a length-normalized pseudo-likelihood rather than a full sequence likelihood, and the correction is assigned to $\mathcal{M}_t$ at the macro-action level. The full derivation is deferred to Appendix A.3.

4.4 Overall Objective

Combining the two calibrations gives the per-token advantage used to train `IG-OPD`:

$$A_{t,i} = \underbrace{\frac{N q_t}{n_t} D_{t,i}^{\mathrm{OPD}}}_{\text{ITC: IG-calibrated turn credit}} + \underbrace{\mathbf{1}[i \in \mathcal{M}_t] \frac{\alpha_k g_t}{m_t}}_{\text{IQC: posterior evidence correction}} \tag{9}$$

During each update, we treat these calibrated advantages as fixed supervision signals for sampled tokens. We then optimize the student using the following token-level policy-gradient surrogate:

$$\widehat{\mathcal{L}}_{\texttt{IG-OPD}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D},\, \tau\sim\pi_{S,\theta_k}} \left[\frac{1}{N} \sum_t \sum_{i\in\mathcal{I}_t} \mathrm{sg}(A_{t,i}) \log \pi_{S,\theta}(u_{t,i} \mid z_{t,i}) \right], \tag{10}$$

where θ_k denotes the policy used to sample the rollout and sg stops gradients through the advantage. Thus, ITC reallocates the normalized teacher-supervision budget across turns, while IQC reinforces or suppresses sampled query tokens according to the signed post-retrieval gain. The IQC scores sum to $\alpha_k g_t$ within each search turn, independently of q_t, and the gains telescope across turns. Computing g_t and q_t requires only one additional lightweight teacher scoring pass of the reference answer per turn but no additional rollouts. The training procedure is given in Algorithm 1 in Appendix A.4.

5 Experiment

5.1 Experiment Setup

Benchmarks & Metrics. We evaluate on six multimodal web-search benchmarks spanning single-hop factual lookup, multi-hop retrieval, and long-horizon deep search: SimpleVQA (Cheng et al., 2025), LiveVQA (Fu et al., 2025), MMSearch (Jiang et al., 2024), MMSearch+ (Tao et al., 2026), FVQA test split (Wang et al., 2017), and BrowseComp-VL (Geng et al., 2026). Following Chen et al. (2026), we report accuracy judged against reference answers by GPT-4o.

Baselines. To evaluate the effectiveness of our proposed `IG-OPD`, we utilize several methods spanning from RL to OPD for comparison, including vanilla decoding, *SFT*, *GRPO* (Shao et al., 2024), *IGPO* (Wang et al., 2026a), *OPD* (Agarwal et al., 2024), and *TCOD* (Wang et al., 2026b).

Implementation Details. We adopt OpenSearch-VL-8B (Chen et al., 2026) as the teacher agent. To optimize efficiency and supervision quality, we optimize its tool set and further fine-tune it with GRPO to elevate the distillation ceiling (Appendix B.1). Our students are Qwen3-VL-2B/4B-Instruct (Bai et al., 2025), which share the teacher's vocabulary. Before distillation, we cold-start the student agent via SFT on 10K trajectories sampled from SearchVL-SFT-36k to equip the student with basic agentic search capability. We then perform OPD training using 4K difficulty-stratified samples sourced from Search-VL-RL-8K (Chen et al., 2026), VDR-Bench-2K (Zeng et al., 2026), and FVQA train split (Wang et al., 2017). Notably, both training and evaluation are conducted in a live web environment, with search actions served by the Serper and Jina APIs. All training runs are executed on $8 \times$ NVIDIA H100 GPUs. Unless stated otherwise, `IG-OPD` uses $\lambda = 0.1$, $\alpha_{\max} = 10$, and a $K = 60$-step cosine schedule. Appendix B.3 provides full training and implementation details.

Table 1: Performance on six multimodal search benchmarks. All benchmarks are out-of-distribution with respect to our training data. **Bold** and underline mark the best and second-best score.

Method	Multimodal Search QA						Avg.
	SimpleVQA	LiveVQA	MMSearch	MMSearch+	FVQA	BrowseComp-VL	
Direct Reasoning							
GPT-5	61.6	44.4	35.1	19.1	54.4	48.6	43.9
Gemini-2.5-Pro	63.0	60.3	39.8	14.5	60.7	43.1	46.9
Claude-4-Sonnet	50.9	38.5	18.7	4.0	35.3	29.3	29.5
Agentic Workflow — OpenSearch-VL-8B (Teacher)							
GRPO Enhanced	71.6	66.0	64.3	19.5	70.7	44.7	56.1
Agentic Workflow — Qwen3-VL-2B-Instruct (Student)							
Zero-Shot	21.0	8.0	14.7	1.8	8.0	1.3	9.1
SFT	57.0	52.7	51.0	12.2	57.0	29.0	43.2
GRPO	61.7	52.0	53.7	10.8	60.3	31.7	45.0
IGPO	60.3	53.7	55.3	12.2	61.0	33.3	46.0
OPD	62.3	52.0	55.7	11.0	62.0	33.0	46.0
TCOD	**63.0**	53.3	57.7	10.4	61.7	31.3	46.2
IG-OPD (Ours)	62.7	**56.3**	**62.0**	**14.4**	**63.7**	**37.3**	**49.4**
Agentic Workflow — Qwen3-VL-4B-Instruct (Student)							
Zero-Shot	58.0	51.3	49.3	9.0	57.0	24.0	41.4
SFT	65.0	61.3	54.0	16.7	64.7	38.0	50.0
GRPO	66.7	62.3	60.3	15.3	65.0	41.7	51.9
IGPO	66.0	63.7	61.0	17.2	66.7	43.3	53.0
OPD	66.7	63.7	62.0	16.2	**67.7**	43.7	53.3
TCOD	63.7	63.3	59.7	16.2	67.3	42.7	52.2
IG-OPD (Ours)	**67.7**	**65.0**	**66.7**	**17.6**	67.3	**45.7**	**55.0**

5.2 CAN IG-OPD TRANSFER DEEP-SEARCH CAPABILITY TO COMPACT STUDENTS?

Table 1 reports results on all six benchmarks for both student scales.

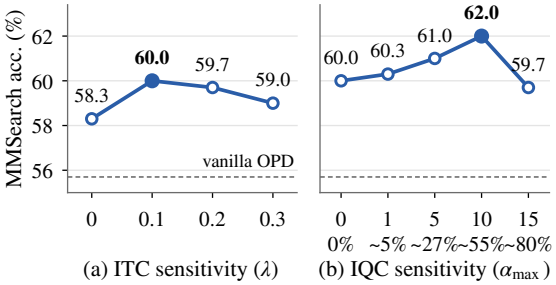
IG-OPD achieves the best average at both scales and lets a 4B student rival its 8B teacher. Under the same cold-start model and training sources, IG-OPD improves average accuracy over vanilla OPD by 3.4 points at 2B (49.4 vs. 46.0) and by 1.7 points at 4B (55.0 vs. 53.3), ranking first on five of the six benchmarks in both student groups. Notably, the 4B student surpasses the 8B teacher on MMSearch (66.7 vs. 64.3) and BrowseComp-VL (45.7 vs. 44.7), effectively matching the teacher's capability overall with half the scale. Although the teacher provides dense supervision, IQC calibrates query-level signals using evidence actually retrieved by the student rather than relying solely on the teacher's prior preference for query wording. The student effectively breaks the OPD ceiling, allowing the student to outperform its teachers. Both IG-OPD students also outperform all direct-reasoning frontier models (e.g., 49.4 for our 2B student vs. 46.9 for Gemini-2.5-Pro), confirming that these benchmarks genuinely require active retrieval rather than parametric recall.

Gains concentrate on search-intensive benchmarks where EBS is most consequential. Compared with OPD, IG-OPD achieves the largest gains on search-intensive benchmarks: MMSearch (+6.3 at 2B / +4.7 at 4B), BrowseComp-VL (+4.3 / +2.0), LiveVQA (+4.3 / +1.3), and MMSearch+ (+3.4 / +1.4). In contrast, on shallow-search benchmarks with limited retrieval steps, IG-OPD maintains consistently competitive performance (e.g., SimpleVQA and FVQA). This directly confirms our intuition: with fewer search turns, there is minimal supervision budget to misallocate (ITC) and few query directions to correct (IQC). As trajectories involve more search turns, vanilla OPD accumulates severe supervision misalignment (Figure 2), making our calibration increasingly crucial.

Calibrating supervision helps where a turn curriculum and outcome rewards do not. TCOD's turn scheduling offers marginal gains at 2B (46.2 vs. 46.0) and underperforms OPD at 4B (52.2 vs. 53.3), suggesting that delaying supervision alone is insufficient without calibrating credit magnitude or direction. Outcome-based GRPO also trails OPD at both scales (45.0/51.9), consistent with the difficulty of assigning credit through sparse trajectory-level rewards. Crucially, IGPO (incorporating IG into policy optimization) reaches only 46.0 at 2B and 53.0 at 4B, trailing IG-OPD by 3.4 and

Table 2: Component ablation on MMSearch (2B student). The IQC-only variant isolates query calibration, while the remaining rows progressively build IG-OPD from OPD.

Variant	Acc.	Δ
OPD (budget n_t/N)	55.7	–
+ IQC only (no ITC)	56.7	+1.0
+ turn normalization (λ=0)	58.3	+2.6
+ ITC with signed g_t	59.3	+3.6
+ ITC with $\lvert g_t \rvert$ (default)	60.0	+4.3
+ IQC (= IG-OPD)	**62.0**	**+6.3**



(a) ITC sensitivity (λ) (b) IQC sensitivity ($\alpha_{\max}$)

Figure 3: Sensitivity on MMSearch (2B student). (a) ITC strength λ with IQC disabled. (b) IQC strength $\alpha_{\max}$ with ITC at its default, with secondary x-ticks indicating relative correction magnitude to OPD.

2.0 points. Although its turn-level feedback improves over outcome-based GRPO, it only matches OPD at 2B and falls slightly below it at 4B. These results suggest that the benefit of information gain depends on how it enters the learning signal: using it to calibrate dense teacher supervision preserves token-level guidance while adjusting supervision according to retrieved evidence. In our setting, this combination of ITC and IQC is more effective than using IG as a turn-level policy reward.

5.3 Which Calibration Matters?

Both priors are harmful, and their corrections are complementary. Table 2 isolates and progressively combines the components of IG-OPD. Replacing the length-induced budget n_t/N with a uniform per-turn budget improves OPD by 2.6 points (58.3), surpassing TCOD (57.7); the token-volume prior is harmful even before information gain is introduced. Modulating this budget by the realized information magnitude adds 1.7 points (60.0), with absolute gain outperforming signed gain (59.3). This supports using magnitude to determine *how much* a turn is supervised, while IQC handles direction. IQC alone yields a 1.0-point gain (56.7), whereas adding it after ITC contributes 2.0 points (62.0). The two calibrations thus address distinct but coupled failure modes: ITC reallocates supervision toward informative turns but cannot correct a teacher-disfavored query, while IQC corrects its direction but is less effective when the original budget under-credits search turns.

IG-OPD is robust to λ and benefits from a moderate evidence coefficient. Figure 3(a) shows that any $\lambda \in [0.1, 0.3]$ outperforms the uniform budget, with the best result at $\lambda = 0.1$ and a gentle decline thereafter (59.7, 59.0); since $\tanh(\cdot) \leq 1$, even $\lambda = 0.3$ bounds the ratio between the most and least credited turns by 1.3, so ITC reshapes rather than overturns the teacher's supervision. Figure 3(b) shows that accuracy increases monotonically with $\alpha_{\max}$ up to 10 (60.0→62.0), at which point the IQC correction amounts to roughly half of the OPD signal, and drops to 59.7 at $\alpha_{\max} = 15$ as excessively large corrections disrupt the underlying dense teacher prior, confirming that posterior calibration acts best as a complementary guidance rather than a complete replacement. These observations suggest that the posterior evidence should remain a moderate correction.

5.4 Does IG-OPD Resolve Evidence-Blind Supervision?

We now return to the two symptoms in Figure 2 and measure them again after calibration.

ITC shifts supervision budget from token volume to information criticality. Figure 4(a) contrasts the length-induced share n_t/N with the ITC budget q_t. Under the prior, the response turn absorbs 47.6% of supervision while the third search turn receives 7.2%; under ITC, the four search groups receive 19.5%–20.8% each (80.9% total vs. 52.4%), ordered by their realized information magnitude (Search 1 highest). The response turn retains 19.1% (near uniform share $1/T$), ensuring answer synthesis is distilled rather than discarded via the additive constant in Eq. 4.

IQC improves search quality and precision. Figure 4(b) tracks how often a rollout's searches retrieve the gold evidence during training. IG-OPD raises the hit rate of the *first* search from 12.9% to 16.9% and of *any* search from 34.1% to 39.1% by step 40, with the gap widening as α_k ramps up, while the number of searches per rollout stays essentially unchanged (4.35 vs. 4.26). Calibrated

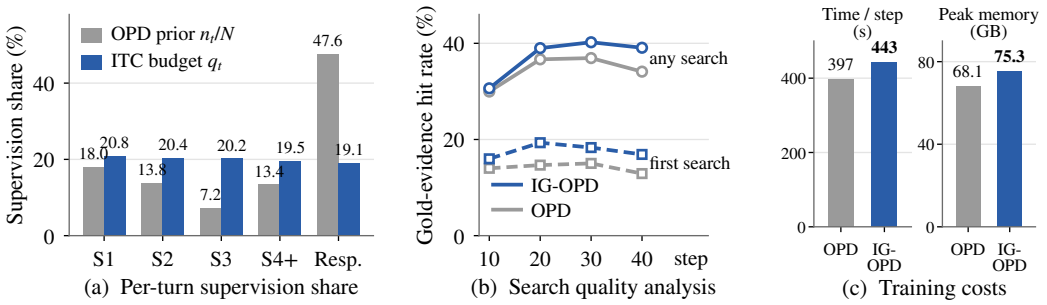


Figure 4: IG-OPD effectively mitigates evidence-blind supervision. (a) Per-turn supervision share under OPD prior n_t/N and ITC budget q_t. (b) Gold-evidence retrieval rate across training (first and any search turns). (c) Training time per step and peak GPU memory for OPD and IG-OPD.

supervision therefore improves the queries the student issues rather than their number, consistent with the 4B student's advantage over its teacher on the multi-hop benchmarks in Sec. 5.2.

Evidence calibration adds modest training overhead. Figure 4(c) compares training costs of OPD and IG-OPD on the same hardware. Time per step rises from 397 to 443 seconds (+11.6%), while peak GPU memory increases from 68.1 to 75.3 GB (+10.6%). This cost arises during training, where post-retrieval information gain is evaluated to calibrate supervision. The increase is moderate relative to the performance gains reported above. At inference, the student follows the same search workflow without computing information gain, so IG-OPD adds no inference overhead.

5.5 Is Post-Retrieval IG Reliable, and Why Ramp It?

IG reliably reflects the final outcome, but its predictive direction depends on search position. Since IG drives both calibrations, Table 3 assesses its alignment with final-answer correctness. Agreement reaches 82.2% on the first search and 70.0% on later ones. The errors exhibit clear position asymmetry: positive IG on the first search strongly predicts success (87.5%), whereas negative IG on later searches correctly predicts failure (85.7%). Focusing on the top 30% largest gains raises both rates to ∼90%. Early on, a positive jump from low potential Φ_t is decisive; later, small fluctuations are mostly verification noise, while large drops signal misleading evidence.

Table 3: Results of IG reliability.

Search position	IG>0 → correct	IG<0 → wrong
First search	87.5%	76.2%
Later searches	54.3%	85.7%

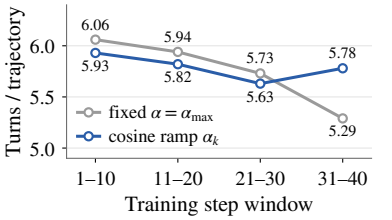


Figure 5: Results of search suppression.

The cosine ramp prevents early search suppression. Early in distillation, weak student retrievals yield predominantly negative IG. Unramped IQC ($\alpha_{\max}$ at step 0) turns this into a general search penalty, dropping average rollout length from 6.06 to 5.29 turns over the first 40 steps on training set (Figure 5). The cosine ramp keeps IQC weak initially while the student acquires teacher search behavior, ensuring IG refines rather than suppresses search once fully active (preserving 5.78 turns). The ramp thus acts as a curriculum from imitation to evidence-calibrated distillation.

6 Conclusion

We have presented IG-OPD, a novel distillation framework designed to address Evidence-Blind Supervision (EBS) in deep search agents. By diagnosing the structural mismatches in traditional On-Policy Distillation (OPD), IG-OPD introduces two synergistic, evidence-aware calibrations: Information-guided Turn-credit Calibration (ITC), which reallocates supervision based on localized information gain rather than token length, and Information-guided Query-preference Calibration (IQC), which aligns query update directions with the realized utility of retrieved evidence. Extensive evaluation across six OOD multimodal search benchmarks demonstrates that IG-OPD significantly outperforms existing distillation and reinforcement learning baselines. Notably, our 4B student effectively bridges the capacity gap, rivaling and even surpassing its 8B teacher without incurring inference-time overhead. Moving forward, extending post-retrieval evidence calibration to other environments represents a highly promising direction for further scaling autonomous web agents.

AI USE STATEMENT

We used generative AI tools to support literature discovery, code development and debugging, exploratory analysis of experimental results, figure and table preparation, and language editing. The authors independently verified the relevant literature, implementations, numerical results, and manuscript content. All methodological choices, scientific claims, and conclusions were made by the authors, who take full responsibility for the final manuscript and its accompanying artifacts.

REPRODUCIBILITY STATEMENT

We provide the detailed derivation and complete IG-OPD algorithm in Appendix A. The agent environment, teacher preparation, training configurations, and data construction procedures are documented in Appendix B, while the rollout and evaluation prompts are reproduced in Appendix C. The benchmark setup, evaluation metrics, baselines, and sampling protocol are described in Section 5.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

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.

Shuang Chen, Kaituo Feng, Hangting Chen, Wenxuan Huang, Dasen Dai, Quanxin Shou, Yunlong Lin, Xiangyu Yue, Shenghua Gao, and Tianyu Pang. Opensearch-vl: An open recipe for frontier multimodal search agents. *arXiv preprint arXiv:2605.05185*, 2026.

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

Mingyang Fu, Yuyang Peng, Dongping Chen, Zetong Zhou, Benlin Liu, Yao Wan, Zhou Zhao, Philip S Yu, and Ranjay Krishna. Seeking and updating with live visual knowledge. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-2513.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Kuan Li, Yida Zhao, Huifeng Yin, Yong Jiang, Pengjun Xie, Fei Huang, Huaxiu Yao, Yi R. Fung, and Jingren Zhou. Webwatcher: Breaking new frontiers of vision-language deep research agent. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 61240–61271, 2026.

Dongze Hao, Zhiwei Jin, Chen Chen, and Haonan Lu. Stabilizing on-policy distillation for mllm reasoning with global normalization. *arXiv preprint arXiv:2606.09091*, 2026.

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

Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. Selectkd: Selective token-weighted knowledge distillation for llms. *arXiv preprint arXiv:2510.24021*, 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.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Gengsheng Li, Mao Zheng, Mingyang Song, Ruiqi Liu, Tianyu Yang, Jie Sun, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Dan Zhang, et al. On-policy distillation with curriculum turn-level guidance for multi-turn agents. *arXiv preprint arXiv:2606.15912*, 2026a.

Guankai Li, Jiabin Chen, Yi Xu, Xichen Zhang, and Yuan Lu. Hypereyes: Dual-grained efficiency-aware reinforcement learning for parallel multimodal search agents. *arXiv preprint arXiv:2605.07177*, 2026b.

Yuying Li, Leqi Zheng, Yongzi Yu, Wenrui Zhou, Xuchang Zhong, Xing Hu, Jing Jin, Hangjie Yuan, and Tao Feng. Filter, then reweight: Rethinking optimization granularity in on-policy distillation. *arXiv preprint arXiv:2606.02684*, 2026c.

Baohao Liao, Hanze Dong, Christof Monz, Xinxing Xu, Li Dong, and Furu Wei. Multi-turn on-policy distillation with prefix replay. *arXiv preprint arXiv:2607.04763*, 2026.

Minhua Lin, Zongyu Wu, Zhichao Xu, Hui Liu, Xianfeng Tang, Qi He, Charu Aggarwal, Xiang Zhang, and Suhang Wang. A comprehensive survey on reinforcement learning-based agentic search: Foundations, roles, optimizations, evaluations, and applications. *arXiv preprint arXiv:2510.16724*, 2025.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning*, pp. 278–287, 1999.

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.

Qitai Tan, Zefang Zong, Yang Li, and Peng Chen. Atod: Annealed turn-aware on-policy distillation for multi-turn autonomous agents. *arXiv preprint arXiv:2606.27814*, 2026.

Xijia Tao, Teng Yihua, Xinxing Su, Xinyu Fu, Jihao Wu, Chaofan Tao, Ziru Liu, Haoli Bai, Rui Liu, and Lingpeng Kong. Mmsearch-plus: Benchmarking provenance-aware search for multimodal browsing agents. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 70385–70422, 2026.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, et al. Information gain-based policy optimization: A simple and effective approach for multi-turn search agents. In *International Conference on Learning Representations*, volume 2026, pp. 104502–104529, 2026a.

Jiaqi Wang, Wenhao Zhang, Weijie Shi, Yaliang Li, and James Cheng. Tcod: Exploring temporal curriculum in on-policy distillation for multi-turn autonomous agents. *arXiv preprint arXiv:2604.24005*, 2026b.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40 (10):2413–2427, 2017.

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

Zhengpeng Xie, Li Lyna Zhang, Zeke Xie, and Mao Yang. Trust region policy distillation. *arXiv preprint arXiv:2607.04751*, 2026.

Hongling Xu, Qi Zhu, Heyuan Deng, Jinpeng Li, Lu Hou, Yasheng Wang, Lifeng Shang, Ruifeng Xu, and Fei Mi. Kdrl: Post-training reasoning llms via unified knowledge distillation and reinforcement learning. *arXiv preprint arXiv:2506.02208*, 2025.

Shicheng Xu, Liang Pang, Yunchang Zhu, Jia Gu, Zihao Wei, Jingcheng Deng, Feiyang Pan, Huawei Shen, and Xueqi Cheng. Rlkd: Distilling llms' reasoning via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34151–34159, 2026a.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026b.

Yu Zeng, Wenxuan Huang, Zhen Fang, Shuang Chen, Yufan Shen, Yishuo Cai, Xiaoman Wang, Zhenfei Yin, Lin Chen, Zehui Chen, et al. Vision-deepresearch benchmark: Rethinking visual and textual search for multimodal large language models. *arXiv preprint arXiv:2602.02185*, 2026.

Wenlin Zhang, Xiaopeng Li, Yingyi Zhang, Pengyue Jia, Yichao Wang, Huifeng Guo, Yong Liu, and Xiangyu Zhao. Deep research: A survey of autonomous research agents. *arXiv preprint arXiv:2508.12752*, 2025.

Zhaoyang Zhang, Shuli Jiang, Yantao Shen, Yuting Zhang, Dhananjay Ram, Shuo Yang, Zhuowen Tu, Wei Xia, and Stefano Soatto. Reinforcement-aware knowledge distillation for llm reasoning. *arXiv preprint arXiv:2602.22495*, 2026.

Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. Scope: Signal-calibrated on-policy distillation enhancement with dual-path adaptive weighting. *arXiv preprint arXiv:2604.10688*, 2026.

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. Sod: Step-wise on-policy distillation for small language model agents. *arXiv preprint arXiv:2605.07725*, 2026.

Yuhang Zhou, Lizhu Zhang, Yifan Wu, Mingyi Wang, Bo Peng, Jiayi Liu, Xiangjun Fan, and Zhuokai Zhao. Sage-opd: Selective agent-guided intervention for multi-turn on-policy distillation. *arXiv preprint arXiv:2606.19659*, 2026a.

Yuhang Zhou, Kai Zheng, Haoling Li, Dengyun Peng, Can Xu, and Jingjing Chen. Turnopd: Making on-policy distillation turn-aware for efficient long-horizon agent training. *arXiv preprint arXiv:2607.05804*, 2026b.

APPENDIX CONTENTS

A Method Details and Derivations. 14
A.1 Turn-wise Decomposition of OPD 14
A.2 Policy-gradient Surrogate and Gradient Flow 15
A.3 Power-posterior Interpretation 16
A.4 Complete IG-OPD Algorithm 17

B Implementation Details 18
B.1 Agent Environment and Teacher Training 18
B.2 OPD Data Construction 19
B.3 Training Configurations 20

C Prompts. 22
C.1 System Prompt 22
C.2 Judge Prompt 23
C.3 Tool Scheme 23

D Case Study 26

E Ethics Statement 28

A Method Details and Derivations

A.1 Turn-wise Decomposition of OPD

Interaction and token notation. For an input x, the student produces a trajectory $\tau = (a_1, o_1, \ldots, a_{T-1}, o_{T-1}, a_T)$, where a_t is a policy-generated search action for $t < T$, o_t is the corresponding tool observation, and a_T is the final response. We write H_{t-1} for the history available immediately before a_t and $H_t = (H_{t-1}, a_t, o_t)$ for the history after the observation is returned. Each action is represented autoregressively by generated tokens. Let $\mathcal{I}_t$ denote their positions, $n_t = |\mathcal{I}_t|$, and $N = \sum_{t=1}^{T} n_t$. For $i \in \mathcal{I}_t$, $u_{t,i}$ is the sampled token and $z_{t,i}$ is its complete prefix. Tool observations enter later prefixes as context, but are not generated by the policy and therefore do not receive token loss.

From prefix matching to sampled-token OPD. OPD matches the student to a frozen teacher on prefixes visited by the student itself. At a fixed prefix z, the reverse-KL objective is

$$\mathcal{K}(z) = D_{\mathrm{KL}}(\pi_S(\cdot \mid z) \,\|\, \pi_T(\cdot \mid z)) = \sum_{u \in \mathcal{V}} \pi_S(u \mid z) \log \frac{\pi_S(u \mid z)}{\pi_T(u \mid z)}, \tag{11}$$

where $\mathcal{V}$ is the vocabulary. Define the sampled teacher–student log-probability gap

$$D^{\mathrm{OPD}}(u, z) = \log \pi_T(u \mid z) - \log \pi_S(u \mid z). \tag{12}$$

For $u \sim \pi_S(\cdot \mid z)$, this is an unbiased one-sample estimator of the negative prefix divergence:

$$\mathbb{E}_{u \sim \pi_S(\cdot|z)}\left[D^{\mathrm{OPD}}(u, z)\right] = -\mathcal{K}(z). \tag{13}$$

Sampling therefore replaces the vocabulary-wide sum with the token already generated during the on-policy rollout. Averaging these samples over all generated positions gives

$$\widehat{\mathcal{L}}_{\mathrm{OPD}} = -\mathbb{E}_{x,\, \tau \sim \pi_S(\cdot|x)} \left[\frac{1}{N} \sum_{t=1}^{T} \sum_{i \in \mathcal{I}_t} \underbrace{\log \frac{\pi_T(u_{t,i} \mid z_{t,i})}{\pi_S(u_{t,i} \mid z_{t,i})}}_{D^{\mathrm{OPD}}_{t,i}} \right]. \tag{14}$$

At every position, $D^{\mathrm{OPD}}_{t,i} > 0$ means that the teacher assigns the sampled token greater probability than the student, whereas $D^{\mathrm{OPD}}_{t,i} < 0$ expresses the opposite preference. The signal is dense because every generated position supplies one such comparison.

Regrouping by interaction turns. Equation 14 uses a single mean over the concatenated generated sequence. Partitioning its indices by turn and multiplying each inner sum by n_t/n_t gives the exact identity

$$\begin{aligned} \frac{1}{N} \sum_{t} \sum_{i \in \mathcal{I}_t} D^{\mathrm{OPD}}_{t,i} &= \sum_{t} \frac{n_t}{N} \left(\frac{1}{n_t} \sum_{i \in \mathcal{I}_t} D^{\mathrm{OPD}}_{t,i} \right) \\ &= \sum_{t} \underbrace{\frac{n_t}{N}}_{\text{budget prior}} \underbrace{\left[\frac{1}{n_t} \sum_{i \in \mathcal{I}_t} \log \frac{\pi_T(u_{t,i} \mid z_{t,i})}{\pi_S(u_{t,i} \mid z_{t,i})} \right]}_{\text{turn-mean teacher signal (direction prior)}}. \end{aligned} \tag{15}$$

Let

$$b^{\mathrm{OPD}}_t = \frac{n_t}{N}, \qquad \bar{D}^{\mathrm{OPD}}_t = \frac{1}{n_t} \sum_{i \in \mathcal{I}_t} D^{\mathrm{OPD}}_{t,i}. \tag{16}$$

Then the trajectory score is simply $\sum_t b^{\mathrm{OPD}}_t \bar{D}^{\mathrm{OPD}}_t$. The coefficient b^{OPD}_t is the *budget prior*: equal token weighting allocates supervision to a turn in direct proportion to its length. The turn-mean signal $\bar{D}^{\mathrm{OPD}}_t$ is the *direction prior*: it determines whether the sampled action is moved toward or away from the teacher's preference.

The timing of a search interaction makes the second prior explicit. Let $\mathcal{F}_t^-$ be the information available when a_t is sampled and $\mathcal{F}_t = \sigma(\mathcal{F}_t^-, o_t)$ the information available after tool execution. Every prefix $z_{t,i}$ and every $D_{t,i}^{\mathrm{OPD}}$ is $\mathcal{F}_t^-$-measurable. In contrast, the realized utility of the action is determined after o_t arrives and is $\mathcal{F}_t$-measurable. Standard OPD therefore fixes both b_t^{OPD} and $\bar{D}_t^{\mathrm{OPD}}$ before observing the evidence produced by the search action.

IG-OPD replaces these two quantities separately. ITC substitutes a normalized posterior budget q_t for b_t^{OPD}, while IQC adds a signed post-retrieval correction on the tokens that specify a_t. The token multiplier implementing the budget substitution follows directly from Eq. 15:

$$w_t = \frac{q_t}{b_t^{\mathrm{OPD}}} = \frac{Nq_t}{n_t}, \qquad \frac{1}{N}\sum_t \sum_{i\in\mathcal{I}_t} w_t D_{t,i}^{\mathrm{OPD}} = \sum_t q_t \bar{D}_t^{\mathrm{OPD}}. \tag{17}$$

Thus $\lambda = 0$ yields $q_t = 1/T$ and assigns every turn exactly the same aggregate OPD budget, independently of its number of generated tokens.

A.2 Policy-gradient Surrogate and Gradient Flow

Score-function form of sampled OPD. The stop-gradient operation separates score construction from policy optimization. To see the resulting gradient, consider the negative KL at a fixed prefix,

$$J_z(\theta) = \mathbb{E}_{u\sim\pi_\theta(\cdot|z)}\left[\log \pi_T(u \mid z) - \log \pi_\theta(u \mid z)\right]. \tag{18}$$

Differentiating the categorical expectation gives

$$\begin{aligned}\nabla_\theta J_z(\theta) &= \sum_u \nabla_\theta \pi_\theta(u \mid z) D_\theta^{\mathrm{OPD}}(u,z) - \sum_u \pi_\theta(u \mid z)\nabla_\theta \log \pi_\theta(u \mid z) \\ &= \mathbb{E}_{u\sim\pi_\theta}\left[D_\theta^{\mathrm{OPD}}(u,z)\nabla_\theta \log \pi_\theta(u \mid z)\right],\end{aligned} \tag{19}$$

because the second sum equals $\sum_u \nabla_\theta \pi_\theta(u \mid z) = 0$. Let $\bar{\theta}$ denote the fixed behavior parameters used to draw a token, and define the local surrogate

$$J_z^{\mathrm{sur}}(\theta;\bar{\theta}) = \mathbb{E}_{u\sim\pi_{\bar{\theta}}(\cdot|z)}\left[\mathrm{sg}\big(D_{\bar{\theta}}^{\mathrm{OPD}}(u,z)\big)\log \pi_\theta(u \mid z)\right]. \tag{20}$$

Evaluated at the behavior policy, $\nabla_\theta J_z^{\mathrm{sur}}(\theta;\bar{\theta})vert_{heta=\bar{\theta}} = \nabla_\theta J_z(\theta)vert_{heta=\bar{\theta}}$. The detached value is therefore the coefficient of a policy gradient; its effect enters through $\nabla_\theta \log \pi_\theta$.

Calibrated per-token score. At update k, trajectories are sampled from the behavior policy π_{θ_k}. The teacher and behavior-policy log probabilities define

$$D_{t,i}^{\mathrm{OPD}} = \log \pi_T(u_{t,i} \mid z_{t,i}) - \log \pi_{\theta_k}(u_{t,i} \mid z_{t,i}). \tag{21}$$

Let $\mathcal{M}_t \subseteq \mathcal{I}_t$ contain the action-critical tokens of search action a_t, and let $m_t = |\mathcal{M}_t|$. Combining ITC and IQC gives

$$A_{t,i} = \frac{Nq_t}{n_t} D_{t,i}^{\mathrm{OPD}} + \mathbf{1}[i \in \mathcal{M}_t]\frac{\alpha_k g_t}{m_t}. \tag{22}$$

Both q_t and g_t are computed from the frozen teacher and are detached along with $D_{t,i}^{\mathrm{OPD}}$. For one trajectory, the corresponding log-probability surrogate is

$$\mathcal{L}_{\mathrm{sur}}(\theta;\tau) = -\frac{1}{N}\sum_t \sum_{i\in\mathcal{I}_t} \mathrm{sg}(A_{t,i}) \log \pi_\theta(u_{t,i} \mid z_{t,i}), \tag{23}$$

where sg denotes stop-gradient. At $\theta = \theta_k$, substitution of Eq. 22 yields

$$\begin{aligned}\nabla_\theta \mathcal{L}_{\mathrm{sur}} = &-\sum_t \frac{q_t}{n_t}\sum_{i\in\mathcal{I}_t} D_{t,i}^{\mathrm{OPD}} \nabla_\theta \log \pi_\theta(u_{t,i} \mid z_{t,i}) \\ &-\frac{\alpha_k}{N}\sum_{t:m_t>0}\frac{g_t}{m_t}\sum_{i\in\mathcal{M}_t}\nabla_\theta \log \pi_\theta(u_{t,i} \mid z_{t,i}).\end{aligned} \tag{24}$$

The first line assigns total OPD coefficient q_t to turn t. The second line applies the signed evidence correction only to the tokens specifying the search decision. A positive g_t increases their sampled log-probabilities; a negative g_t decreases them. Retrieval, reference answer scoring, and the computation of q_t remain outside the autograd graph, while the student parameters are updated through the two displayed log-probability gradients.

Clipped on-policy implementation. During optimization, the actor uses the likelihood ratio $\rho_{t,i}(\theta) = \pi_\theta(u_{t,i} \mid z_{t,i})/\pi_{\theta_k}(u_{t,i} \mid z_{t,i})$ and the clipped surrogate

$$\mathcal{L}_{\text{clip}}(\theta) = -\frac{1}{Z} \sum_{\tau \in \mathcal{B}} \sum_{t,i \in \mathcal{I}_t} \min\Big\{ \rho_{t,i} \widehat{A}_{t,i},\ \text{clip}(\rho_{t,i}, 1-\delta, 1+\delta) \widehat{A}_{t,i} \Big\}, \tag{25}$$

where $\widehat{A}_{t,i} = \text{sg}(A_{t,i})$, δ is the clipping threshold, and Z counts valid generated tokens in the reduction batch. At $\theta = \theta_k$, $\rho_{t,i} = 1$ and the gradient reduces to Eq. 24. Away from the behavior policy, clipping bounds the change contributed by each sampled token. Padding positions, tool observations, and prompt tokens are removed by the generation mask before the normalizer Z is computed.

The two calibrations enter this surrogate through distinct algebraic paths: q_t rescales the dense OPD coefficient on all generated tokens in the turn, whereas g_t contributes an additive signed coefficient on $\mathcal{M}_t$. IQC is placed outside the ITC multiplier, so its aggregate score is set by the realized information change rather than by the length-derived or calibrated OPD budget.

A.3 Power-posterior Interpretation

Trajectory factorization. Let P_{env} denote the tool environment. The trajectory distribution induced by a policy π factorizes as

$$p_\pi(\tau \mid x) = \left[\prod_{t=1}^{T-1} \pi(a_t \mid H_{t-1}) P_{\text{env}}(o_t \mid H_{t-1}, a_t) \right] \pi(a_T \mid H_{T-1}). \tag{26}$$

The student and teacher act in the same environment. Hence the environment transition terms cancel in their trajectory likelihood ratio:

$$\log \frac{p_T(\tau \mid x)}{p_S(\tau \mid x)} = \sum_{t=1}^{T} \log \frac{\pi_T(a_t \mid H_{t-1})}{\pi_S(a_t \mid H_{t-1})}. \tag{27}$$

This is the sequence-level counterpart of summing the token gaps in Eq. 12.

Evidence-conditioned target distribution. Write $\Psi_t = \exp(\Phi_t)$ and set $g_T = 0$ for the final response. The post-retrieval gains telescope:

$$G(\tau) = \sum_{t=1}^{T-1} g_t = \sum_{t=1}^{T-1} (\Phi_t - \Phi_{t-1}) = \Phi_{T-1} - \Phi_0. \tag{28}$$

At training step k, augment the sequence-level OPD score with this net information change:

$$R_k(\tau) = \log \frac{p_T(\tau \mid x)}{p_S(\tau \mid x)} + \alpha_k \sum_{t<T} g_t = \log \frac{p_T(\tau \mid x) \Psi_{T-1}(\tau)^{\alpha_k}}{p_S(\tau \mid x)} - \alpha_k \Phi_0. \tag{29}$$

The term $-\alpha_k \Phi_0$ depends on $(x, y^\star)$ but not on the sampled trajectory. Define the normalized evidence-conditioned teacher target

$$\widetilde{p}_{T,k}(\tau \mid x, y^\star) = \frac{p_T(\tau \mid x) \Psi_{T-1}(\tau)^{\alpha_k}}{\mathcal{Z}_k(x, y^\star)}, \qquad \mathcal{Z}_k = \mathbb{E}_{\tau \sim p_T}\big[\Psi_{T-1}(\tau)^{\alpha_k}\big]. \tag{30}$$

For fixed $(x, y^\star)$ and α_k, both $\mathcal{Z}_k$ and Φ_0 are independent of the student parameters. Expanding the reverse KL gives

$$-\mathbb{E}_{\tau \sim p_S}[R_k(\tau)] = D_{\text{KL}}(p_S \,\|\, \widetilde{p}_{T,k}) - \log \mathcal{Z}_k + \alpha_k \Phi_0. \tag{31}$$

Minimizing the gain-augmented sequence objective is therefore equivalent, up to constants independent of the student, to reverse-KL distillation from $\widetilde{p}_{T,k}$. The target retains the teacher trajectory prior and tilts it by the post-retrieval answer potential.

The same result can be written as a product of local evidence updates:

$$\Psi_{T-1}(\tau)^{\alpha_k} = \Psi_0^{\alpha_k} \prod_{t=1}^{T-1} \exp(\alpha_k g_t). \tag{32}$$

Each returned observation contributes the multiplicative factor $\exp(\alpha_k g_t)$: useful evidence raises the trajectory's target mass, while misleading evidence lowers it. For any two trajectories τ and τ', their odds under the calibrated target satisfy

$$\frac{\widetilde{p}_{T,k}(\tau \mid x, y^\star)}{\widetilde{p}_{T,k}(\tau' \mid x, y^\star)} = \frac{p_T(\tau \mid x)}{p_T(\tau' \mid x)} \left(\frac{\Psi_{T-1}(\tau)}{\Psi_{T-1}(\tau')} \right)^{\alpha_k}. \tag{33}$$

At $\alpha_k = 0$, the target is the teacher trajectory distribution. As the cosine schedule increases α_k, post-retrieval answer confidence contributes progressively more strongly to trajectory ranking. The token-level IQC term realizes this evidence tilt on the decision tokens of each search action, while ITC supplies the complementary posterior turn budget.

A.4 Complete IG-OPD Algorithm

Reference-answer probes and masks. The frozen teacher scores only reference-answer tokens under the initial context and each post-observation history, using teacher forcing rather than generating another answer. Their mean log-likelihood yields the potential and adjacent differences g_t defined in the main text. The response has $g_T = 0$ but retains a positive ITC budget. For a search action a_t, IQC masks select the parsed tool-name and supported argument-value spans, mapped back to generated-token positions; reasoning, wrapper tags, and observations are excluded.

Algorithm 1 Information-Gain Calibrated OPD

Input: frozen teacher π_T, initialized student π_θ, training pairs $(x, y^\star)$, ITC strength λ, IQC strength $\alpha_{\max}$, ramp horizon K, and clipping threshold δ.

1. At update k, sample a prompt batch and generate tool-interleaved rollouts with the behavior policy π_{θ_k}. Record generated-token indices $\mathcal{I}_t$, pre-action prefixes, tool observations, and parsed action spans.
2. Score every generated token under π_T and π_{θ_k} on the same prefix, and compute $D_{t,i}^{\mathrm{OPD}}$ using Eq. 21.
3. For each rollout, evaluate Φ_t under $H_0, H_1, \ldots, H_{T-1}$. Set $g_t = \Phi_t - \Phi_{t-1}$ for $t < T$ and $g_T = 0$.
4. Compute the batch gain scale σ_g and the normalized ITC budget q_t with Eq. 4. Set the per-token turn multiplier $w_t = N q_t / n_t$.
5. Parse each search action to construct $\mathcal{M}_t$ and $m_t = |\mathcal{M}_t|$. Set $\alpha_k = \frac{\alpha_{\max}}{2}[1 - \cos(\pi \min\{k/K, 1\})]$.
6. Form the calibrated token score $A_{t,i} = w_t D_{t,i}^{\mathrm{OPD}} + \mathbf{1}[i \in \mathcal{M}_t]\alpha_k g_t / m_t$, using a zero IQC term whenever $m_t = 0$.
7. Detach $A_{t,i}$, construct the token ratios against π_{θ_k}, and update the student by minimizing Eq. 25 over generated-token positions.
8. Replace the behavior policy with the updated student and repeat until the prescribed training horizon is reached.

B Implementation Details

B.1 Agent Environment and Teacher Training

Table 4: Tool environment shared by teacher training, student training, and inference.

Family	Tool	Backend and returned observation
Retrieval	`text_search`	Serper organic search, Jina Reader page extraction, and query-conditioned Qwen3-32B evidence summaries with titles and URLs.
	`image_search`	Serper Lens reverse-image retrieval over the original image or an intermediate image identifier.
Perception	`layout_parsing`	PPStructureV3 with PP-OCRv5 and layout, table, formula, and chart recognition; returns structured blocks, text, and Markdown.
	`crop`	PIL region extraction; returns a new image identifier and region metadata.
Enhancement	`perspective_correct`	OpenCV quadrilateral rectification and homography.
	`sharpen`	Local unsharp masking for small or blurred visual content.
	`super_resolution`	EDSR $4\times$ upsampling for low-resolution regions.

Action–observation environment. We implement multimodal search as a multi-turn action–observation process. At turn t, the policy receives the original question and image together with the complete interaction history. It then emits a reasoning block followed by either one structured tool call or a final response. A tool call is represented as a JSON object enclosed by `<tool_call>` tags, and the corresponding result is appended to the trajectory inside `<observation>` tags. This creates an explicit sequence of policy actions and environment observations that can be processed consistently during data collection, SFT, on-policy distillation, and inference.

Each newly generated image is assigned a stable identifier (`img_1`, `img_2`, ...). Subsequent tools can therefore operate on a crop, rectified view, or enhanced version without serializing the image into the textual context. The policy may execute at most one tool per turn and up to 20 interaction turns per trajectory. Model-generated reasoning, tool-call, and final-response tokens receive optimization masks. System instructions, user content, images, and tool observations remain visible as conditioning context but are excluded from the actor loss.

Search and retrieval stack. The textual retrieval tool first submits the query to the Serper Search API and obtains the top organic results, with five results used by default. For each result, the pipeline retains its title, URL, and search snippet and uses Jina Reader to extract page-level content. The recovered passages are then compressed by a query-conditioned Qwen3-32B summarizer exposed through an OpenAI-compatible endpoint. The resulting observation preserves the source URL and the evidence relevant to the current information need while removing unrelated webpage content.

This retrieval pipeline allows the agent to refine its query over multiple turns without repeatedly inserting entire webpages into the context. The language parameter is passed to Serper, enabling multilingual queries and cross-lingual retrieval through the same interface. Search observations include the result title, source URL, and a focused evidence summary, which allows later turns to verify, refine, or combine information from multiple sources.

The image-retrieval tool uses the Serper Lens endpoint. It accepts either the original image or an intermediate image identifier, performs a reverse visual lookup, and returns candidate entities and visually related pages. This is particularly useful when an object, artwork, landmark, product, or document must first be identified before a factual text query can be formed. The agent prompt instructs the policy to follow visual identification with `text_search` whenever the requested answer depends on an externally verifiable fact.

Visual perception and enhancement stack. Document understanding is served locally through a FastAPI endpoint backed by PaddleOCR PPStructureV3. The service combines PP-OCRv5 text

detection and recognition with dedicated document-layout, table, formula, and chart modules. It returns recognized text, Markdown, reading-order information, and typed layout blocks. Consequently, labels, tables, equations, captions, and dense document regions can be incorporated into the trajectory as structured observations rather than inferred solely from image pixels.

Lightweight visual transformations are executed locally. Cropping uses PIL to extract a specified region. Perspective correction detects a quadrilateral and applies an OpenCV homography. Sharpening uses unsharp masking to improve the legibility of small or blurred text. Super-resolution uses an EDSR $4\times$ model. These operations can be composed into tool chains such as `perspective_correct`→`crop` →`layout_parsing`, while each intermediate image remains addressable through its assigned identifier.

Training the teacher with custom GRPO rewards. Before distillation, we adapt OpenSearch-VL-8B to the tool environment with GRPO. The reward jointly optimizes final-answer correctness, search quality, structured execution, and search efficiency. For every sampled trajectory τ, we define

$$R_T(\tau) = r_{\text{fmt}}\, e(s)\,(0.8 r_{\text{acc}} + 0.2 r_{\text{query}})\,, \qquad e(s) = \begin{cases} 1, & s \le 5, \\ \max\{0.7, 1 - 0.05(s-5)\}, & s > 5, \end{cases} \tag{34}$$

where r_{acc} scores final-answer correctness and r_{query} evaluates evidence relevance, query progression, and the complementarity of visual and textual search. The variable s counts text and image searches.

The format term r_{fmt} is computed from the structural validity of the generated steps. Well-formed reasoning, tool-call, observation, and response blocks receive full credit. A valid direct-response fallback receives partial credit, whereas malformed tool calls and execution-error steps receive zero. The efficiency factor preserves the full task reward for concise trajectories and gradually discounts redundant retrieval after the fifth search without making search count the dominant objective.

For each prompt x, GRPO samples a group of $G = 8$ trajectories and forms the group-relative advantage

$$\widehat{A}_j = \frac{R_T(\tau_j) - \mu_x}{\sigma_x + \varepsilon}, \qquad \mu_x = \frac{1}{G}\sum_{j=1}^{G} R_T(\tau_j), \tag{35}$$

where μ_x and σ_x are the mean and standard deviation within the rollout group. The teacher actor is optimized using

$$\mathcal{J}_T(\theta) = \mathbb{E}_x \left[\frac{1}{G}\sum_{j=1}^{G} \sum_{i \in \mathcal{I}(\tau_j)} \min\Big(\rho_{j,i}(\theta)\widehat{A}_j,\ \text{clip}\big(\rho_{j,i}(\theta), 0.8, 1.28\big)\widehat{A}_j \Big) \right], \tag{36}$$

where

$$\rho_{j,i}(\theta) = \frac{\pi_\theta(u_{j,i} \mid z_{j,i})}{\pi_{\text{old}}(u_{j,i} \mid z_{j,i})}, \tag{37}$$

and $\mathcal{I}(\tau_j)$ contains only model-generated tokens. Masked token scores are summed within each sequence and averaged across sequences.

Teacher training uses 64 prompts per update and eight responses per prompt, corresponding to 512 sampled trajectories per update. We use a PPO minibatch size of 16, an actor learning rate of 10^{-6}, BF16 precision, and one epoch on eight GPUs. Teacher rollouts use temperature 0.7 and top-$p = 1$. We use the resulting update-60 checkpoint as the distillation teacher and freeze all teacher parameters throughout subsequent OPD and IG-OPD training.

B.2 OPD Data Construction

Source mixture. The OPD corpus contains search-intensive multimodal questions rather than precomputed teacher trajectories. Each record contains a unique training identifier, a question, one or more image references, a reference answer $y^\star$, and source metadata. The clean parent pool combines Search-VL-RL-8K, VDR-Bench-2K, and the FVQA training split.

We sample 3,840 prompts without replacement using a fixed 70:15:15 source ratio, as summarized in Table 5. Search-VL-RL provides broad open-domain visual-search problems, while VDR-Bench

and FVQA increase the coverage of document-centric perception and fact-oriented retrieval. The FVQA component is drawn exclusively from its training split.

Table 5: OPD prompt mixture shared by OPD, IG-OPD, and the on-policy student baselines.

Source	Parent pool	Selected	Share
Search-VL-RL-8K	5,600	2,688	70%
VDR-Bench-2K	1,200	576	15%
FVQA training split	1,200	576	15%
Total	8,000	3,840	100%

Schema validation and within-pool cleaning. We first convert the three sources into a common schema. Question text is canonicalized by normalizing Unicode, letter case, punctuation, and whitespace. Records with an empty question or answer, invalid image references, duplicated training identifiers, or malformed fields are removed before source-stratified sampling.

Deterministic subset construction. Within each cleaned source partition, examples are sampled without replacement using seed 42. The selected partitions are concatenated and globally shuffled using the same seed. The builder verifies that all 3,840 output identifiers are unique and that every selected item has a matching reference-answer entry. It then writes both JSONL and Parquet versions and records a SHA-256 digest of the ordered identifier list.

Reference answers and reusable tool artifacts. Reference answers are stored in an identifier-keyed cache and serve two roles. First, they provide the answer target used by outcome-reward baselines. Second, IG-OPD inserts the same $y^\star$ into the frozen teacher probes that produce H_t and H_t^+ in Appendix A. Dataset construction validates cache coverage for every selected training identifier, eliminating online answer lookup from the training loop.

Image paths point to canonical source assets rather than duplicated image payloads. We additionally cache the initial-image visual-search response by training identifier and image identity. Paired runs can therefore reuse the same initial visual lookup, stabilizing this external component while leaving query-dependent textual retrieval fully on-policy. All subsequent queries and tool-selection decisions are generated by the current student policy.

B.3 Training Configurations

SFT cold start. Both student scales start from Qwen3-VL-Instruct and undergo full-parameter SFT before on-policy training. The vision tower and multimodal projector remain trainable, allowing the model to learn visual inspection, multi-turn reasoning, and the structured tool protocol jointly with language generation. We use the `qwen3_vl` conversation template, a maximum sequence length of 24,000 tokens, maximum image and video budgets of 262,144 and 16,384 pixels, and BF16 precision. SFT uses a global batch size of 128, a learning rate of 10^{-5}, cosine decay with a 0.1 warmup fraction, two epochs, gradient checkpointing, and ZeRO-3 over eight GPUs. The 2B and 4B distillation runs are initialized from their corresponding update-154 SFT checkpoints.

Shared on-policy training backbone. OPD and IG-OPD share the same student initialization, frozen teacher, prompt ordering, rollout engine, optimizer, and distributed configuration. Each update samples 64 prompts and one student trajectory per prompt. The asynchronous SGLang engine executes up to 64 agent trajectories and 256 tool calls in parallel.

Rollouts use temperature 0.7, top-$p = 1$, and disabled top-k truncation. We use tensor parallelism 2, a maximum prompt length of 20,000 tokens, a maximum generated length of 70,000 tokens, and at most 20 agent turns. The long response budget accommodates interleaved reasoning, tool calls, retrieved observations, and the final response without truncating multi-hop search trajectories.

The student actor is trained in BF16 with FSDP on eight GPUs. We use a learning rate of 5×10^{-6}, cosine decay with a 0.1 warmup fraction, a PPO minibatch size of 64, one actor epoch per rollout batch, clipping radius 0.2, and dynamic token-based microbatching. Gradient checkpointing and

CPU parameter and optimizer offloading are used for long-context updates. The teacher remains frozen and can be offloaded outside its scoring windows.

As derived in Appendix A.2, optimization is applied only to student-generated action tokens. Prompts and tool observations determine the policy context but are not prediction targets. No entropy bonus or additional KL reward is used in the student objective.

Table 6: Training configurations for the 2B and 4B students. OPD and IG-OPD share the on-policy backbone; the final rows list the calibration terms enabled by IG-OPD.

Category	**Parameter**	**SFT**	**OPD / IG-OPD**
Model	Student scale	2B / 4B	2B / 4B
	Initialization	Qwen3-VL-Instruct	SFT update 154
	Teacher	—	OpenSearch-VL-8B, frozen
	Trainable parameters	Full model	Full student
	Precision	BF16	BF16
Data	Training instances	9,843 trajectories	3,840 prompts
	Global batch size	128 trajectories	64 prompts
	Image / video pixel budget	262,144 / 16,384	Input dependent
Sequence	Maximum sequence length	24,000	—
	Maximum prompt / response length	—	20,000 / 70,000
	Maximum agent turns	—	20
Optimization	Learning rate	10^{-5}	5×10^{-6}
	LR schedule / warmup fraction	Cosine / 0.1	Cosine / 0.1
	Training horizon	2 epochs	1 epoch
	PPO minibatch / actor epochs	—	$64/1$
	PPO clipping	—	0.2
	Entropy / extra KL coefficient	—	$0/0$
	Gradient checkpointing	Yes	Yes
Rollout	Engine	—	SGLang, asynchronous
	Responses per prompt	—	1
	Temperature / top-p	—	$0.7/1.0$
	Tensor parallelism	—	2
	Parallel trajectories / tool calls	—	$64/256$
IG-OPD	ITC strength λ	—	0.1
	IQC strength / ramp horizon	—	$10/60$
	Gain / potential stabilizer	—	$10^{-6}/10^{-3}$
	IQC placement	—	Outside the ITC weight
System	GPUs	8	8
	Parallel strategy	ZeRO-3	FSDP
	Checkpoint interval	End of epoch	10 updates

C PROMPTS

C.1 SYSTEM PROMPT

The same system instruction is used for teacher data collection, on-policy student rollouts, and inference. The prompt defines the action–observation protocol, visual-inspection strategy, retrieval policy, and output format. At runtime, the user message contains the tool schemas inside `<tools>` tags, followed by the question and any attached image. The following box preserves the operational instructions while retaining one representative tool-use example.

Agent System Prompt

```
You are an advanced Visual Investigation Agent. Answer user
questions with maximum precision by proactively using image-
processing and retrieval tools.

The attached image in this user message is named img_1. Tool
outputs that create new images are named img_2, img_3, and so on.
Always operate on the latest and clearest image version.

Core Philosophy: "Verify, Don't Guess"

1. Tool-First Mindset
If text is small, crop it. If text is blurry, sharpen it.
If the image is tilted, use perspective_correct. Do not rely
only on the internal visual encoder when a tool can provide a
clearer observation.

2. Chain Your Tools
A non-trivial visual query may require several operations. For
example:

perspective_correct -> crop -> layout_parsing

Continue until the relevant visual information is sufficiently
clear.

3. External Validation
Whenever the answer depends on facts that are not purely visible
in the image, call text_search. After image_search identifies an
entity, use text_search to retrieve and verify the factual detail.

4. Thinking Protocol
Before each action, emit a <think> block that:
- analyzes the request;
- assesses image quality, legibility, geometry, and target size;
- identifies the remaining information gap;
- selects one next action.

5. Tool Calling Format
Tool signatures are supplied inside <tools>...</tools>. Emit
exactly one JSON tool call inside <tool_call>...</tool_call>:

<tool_call>
{"name": "<function>", "arguments": {<arguments>}}
</tool_call>

Wait for the corresponding <observation> before selecting the
next action.

6. Result Interpretation
- Treat successful layout_parsing output as extracted document
  evidence.
- Treat image_search output as an initial entity-identification
  result.
- Treat text_search output as query-focused external evidence.
- Use titles and URLs in the retrieved observations when
  synthesizing the answer.

7. Output Rules
- Emit at most one tool call per turn.
- Never emit a tool call without a preceding <think> block.
- When sufficient evidence has been collected, place the final
  answer inside <response>...</response>.
```

```
Example:

<think>
The document in img_1 is skewed, so I will correct its
perspective before extracting the text.
</think>
<tool_call>
{"name":"perspective_correct",
 "arguments":{"image":"img_1"}}
</tool_call>
```

For examples without an attached image, the image-reference paragraph is replaced by an instruction stating that `img_1` is unavailable. The remaining reasoning, retrieval, and output rules are unchanged.

C.2 JUDGE PROMPT

We use `gpt-4o-2024-08-06` as the answer judge. The judge receives the original question, the reference answer, and the complete final report. It determines whether the report contains an answer consistent with the reference and returns a binary decision together with a short explanation.

Answer Judge Prompt

```
System:
You are a precise but high-recall answer judge.

User:
You are evaluating whether a deep-research report contains the
correct answer to the given question.

[Question]
{question}

[Correct Answer]
{correct_answer}

[Deep Research Report]
{response}

Determine whether the deep-research report contains the correct
answer anywhere in its content.

1. Read the complete report.
2. Locate the answer expressed by the report.
3. Compare it with the provided correct answer.
4. Accept equivalent entities, aliases, units, numerical forms,
   dates, and semantically equivalent wording.
5. The answer does not need to appear in a dedicated final-answer
   sentence.
6. Return "yes" when the report contains an answer consistent with
   the correct answer.
7. Return "no" when the answer is absent or contradicted by the
   report.

Output exactly:
correct: [yes/no]
reasoning: [brief explanation]
```

The returned field `correct` is parsed as the binary correctness label. Outputs that do not follow the required form are resubmitted with an explicit formatting reminder.

C.3 TOOL SCHEME

The tool registry is serialized into an OpenAI-style JSON array and inserted into the user message inside `<tools>` tags. Each entry specifies the function name, its role, the accepted argument schema, required fields, and default values. The scheme below follows the tool registry used by the on-policy rollout environment. The `web_search` entry is a compatibility alias of `text_search`; both execute the same Serper–Jina–Qwen retrieval pipeline. Here, we show part of our tool scheme.

Tool Scheme

```
<tools>
[
  {
    "type": "function",
    "function": {
      "name": "crop",
      "description":
        "Crop a specific region from an image.",
      "parameters": {
        "type": "object",
        "properties": {
          "image": {
            "type": "string",
            "description": "Image reference, e.g. img_1"
          },
          "x": {
            "type": "integer",
            "description": "Starting X coordinate"
          },
          "y": {
            "type": "integer",
            "description": "Starting Y coordinate"
          },
          "width": {
            "type": "integer",
            "description": "Width of the crop region"
          },
          "height": {
            "type": "integer",
            "description": "Height of the crop region"
          }
        },
        "required": ["image", "x", "y", "width", "height"]
      }
    }
  },

  {
    "type": "function",
    "function": {
      "name": "layout_parsing",
      "description":
        "Perform document layout parsing on an image to extract
         structured text.",
      "parameters": {
        "type": "object",
        "properties": {
          "image": {
            "type": "string",
            "description": "Image reference, e.g. img_1"
          },
          "file_path": {
            "type": "string",
            "description": "Absolute image path, when available"
          },
          "use_chart_recognition": {
            "type": "boolean",
            "default": false
          },
          "use_doc_orientation_classify": {
            "type": "boolean",
            "default": false
          }
        },
        "required": []
      }
    }
  },

  {
    "type": "function",
    "function": {
      "name": "image_search",
      "description":
        "Analyze an image using visual search to identify objects.",
      "parameters": {
        "type": "object",
```

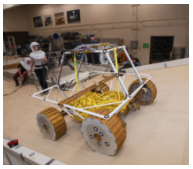
```
        "properties": {
          "url": {
            "type": "string",
            "description":
              "Image reference, e.g. img_1, or a direct URL"
          }
        },
        "required": ["url"]
      }
    }
  },

  {
    "type": "function",
    "function": {
      "name": "text_search",
      "description":
        "Search the web and return query-relevant passages from
         the top-k results.",
      "parameters": {
        "type": "object",
        "properties": {
          "q": {
            "type": "string",
            "description": "Search query"
          },
          "query": {
            "type": "string",
            "description": "Alternative field for q"
          },
          "hl": {
            "type": "string",
            "default": "en"
          },
          "lang": {
            "type": "string",
            "description": "Alias for hl",
            "default": "en"
          },
          "top_k": {
            "type": "integer",
            "default": 5
          }
        },
        "required": []
      }
    }
  }
]
</tools>
```

The rollout parser canonicalizes the aliases `query`→ `q` and `lang`→`hl`. Visual tools return a new image identifier when they produce an image, while retrieval tools return textual observations that remain in the trajectory context for subsequent reasoning.

D CASE STUDY

We present four successful trajectories produced by the 4B IG-OPD student. The cases cover visual identification, document understanding, image enhancement, and cross-entity multi-hop retrieval. Retrieved passages are abridged for readability, while all retained tool calls and their corresponding observations are shown. Figure 6 shows how the student identifies NASA's VIPER rover from visual cues and progressively verifies the exact date on which its cancellation was announced. Figure 7 demonstrates a longer tool chain involving region cropping, layout parsing, visual search, and textual verification. The student combines multilingual signage and retrieved location evidence to identify Myeongdong Night Market.



input image

User Question: In which day of 2024 was the lunar rover project in the picture announced to be canceled?

T1 <think> I will use the visible cues to identify the project. The rover has a white tubular frame, gold thermal blankets, large rugged wheels, and stands in a NASA clean room. </think>
<tool_call> {"name":"text_search","arguments":{"q":"NASA lunar rover white frame gold thermal blanket cancelled in 2024","hl":"en","top_k":5}} </tool_call>
<observation> The results identify NASA's VIPER rover as a cancelled robotic lunar mission. </observation>

T2 <think> The results point to VIPER, but I should verify the exact cancellation date requested by the question. </think>
<tool_call> {"name":"text_search","arguments":{"q":"NASA VIPER lunar rover cancelled date 2024 announcement","hl":"en","top_k":5}} </tool_call>
<observation> The Planetary Society and SpaceNews report that NASA announced the cancellation on July 17, 2024. </observation>

T3 <think> I have the date; I should confirm that the rover pictured is VIPER rather than another NASA rover. </think>
<tool_call> {"name":"text_search","arguments":{"q":"VIPER lunar rover NASA white frame design","hl":"en","top_k":3}} </tool_call>
<observation> NASA describes VIPER as a lunar rover with three instruments and a one-meter drill; related results show the VIPER chassis in NASA's Johnson Space Center clean room. </observation>

T4 <think> The visual identity is consistent with VIPER. I will make one focused search to corroborate the announcement date. </think>
<tool_call> {"name":"text_search","arguments":{"q":"VIPER NASA canceled July 17 2024 announcement date","hl":"en","top_k":3}} </tool_call>
<observation> NASA's announcement and independent mission coverage corroborate July 17, 2024. </observation>

T5 <think> The pictured rover is VIPER, and multiple sources agree on the announcement date. I now have sufficient evidence. </think>
<response> The lunar rover is NASA's VIPER. NASA announced its cancellation on July 17, 2024. </response>

Gold / judged answer: July 17, 2024

Figure 6: A successful MMSearch trajectory. The student identifies the pictured lunar rover as NASA's VIPER and corroborates its cancellation date using multiple retrieval steps.



input image

User Question: What is the name of the market?

T1 <think> The image shows a Korean street-food stall. I will inspect the visible signs and pricing area for a location cue. </think>
<tool_call> {"name":"crop","arguments":{"image":"img_1","x":650,"y":150,"width":1900,"height":500}} </tool_call>
<observation> The upper stall region is returned as img_2. </observation>

T2 <think> The identifying sign is lower in the frame, so I will crop that region more tightly. </think>
<tool_call> {"name":"crop","arguments":{"image":"img_1","x":400,"y":500,"width":2500,"height":250}} </tool_call>
<observation> The lower sign region is returned as img_3. </observation>

T3 <think> I will inspect the remaining lower strip containing the multilingual steak advertisement. </think>
<tool_call> {"name":"crop","arguments":{"image":"img_1","x":350,"y":600,"width":2000,"height":200}} </tool_call>
<observation> The final sign crop is returned as img_4. </observation>

T4 <think> The lettering is still small; layout parsing can recover the multilingual text from the original image. </think>
<tool_call> {"name":"layout_parsing","arguments":{"image":"img_1"}} </tool_call>
<observation> The parser recovers the phrase "If the steak is not good, I will not take any money" together with Korean, Chinese, and Japanese text. </observation>

T5 <think> The stall style and multilingual signage suggest a major Seoul night market. I will use visual search for the location. </think>
<tool_call> {"name":"image_search","arguments":{"url":"img_1"}} </tool_call>
<observation> The returned visual matches explicitly mention Myeongdong, one of Seoul's best-known shopping and street-food districts. </observation>

T6 <think> I will verify that this street-food setting corresponds to the named market. </think>
<tool_call> {"name":"text_search","arguments":{"q":"Myeongdong Seoul street food steak market","hl":"en","top_k":5}} </tool_call>
<observation> Travel and local sources describe Myeongdong Night Market as a dense street-food market in central Seoul. </observation>

T7 <think> The visual match, signage, and external sources identify the location consistently. </think>
<response> The market is Myeongdong Night Market in Seoul, South Korea. </response>

Gold / judged answer: Myeongdong Night Market

Figure 7: A successful MMSearch+ trajectory. The student progressively localizes the street-food scene through visual inspection, OCR evidence, reverse-image retrieval, and textual verification.

Moreover, Figure 8 illustrates how the student uses a small textual cue on an athlete's bib to formulate a targeted search. It then verifies both the athlete's identity and the corresponding prize amount using independent sources. Figure 9 presents a cross-entity reasoning chain. The student enhances the portrait, identifies Jonathan Nolan, retrieves his relationship to Christopher Nolan, and then determines the highest-ranked film directed by the latter.



input image

User Question: How much in USD did the 200m champion earn at last Saturday's elite track event?

T1
<think> The runner's bib contains the most useful identity cue. I will crop it before forming a search query. </think>
<tool_call> {"name":"crop","arguments":{"image":"img_1","x":350,"y":200,"width":2500,"height":200}} </tool_call>
<observation> The bib region is returned as img_2 and appears to contain a form of the surname Hudson-Smith. </observation>

T2
<think> I will apply layout parsing to the enlarged bib region to recover its text. </think>
<tool_call> {"name":"layout_parsing","arguments":{"image":"img_2"}} </tool_call>
<observation> The parser isolates the runner and bib region but does not return a reliable transcription. </observation>

T3
<think> The visible surname pattern, Puma kit, and event context provide enough cues for a targeted retrieval query. </think>
<tool_call> {"name":"text_search","arguments":{"q":"Hunsworth Noson 200m champion track event prize money","hl":"en","top_k":5}} </tool_call>
<observation> The results identify Matthew Hudson-Smith as the Grand Slam Track 200m winner and report a first-place award of $100,000. </observation>

T4
<think> I will verify the athlete, event, and prize using a query containing the recovered identity. </think>
<tool_call> {"name":"text_search","arguments":{"q":"Matthew Hudson-Smith Grand Slam Track Kingston 200m $100000 prize money","hl":"en","top_k":5}} </tool_call>
<observation> BBC and athletics coverage confirm that Hudson-Smith won the Kingston long-sprints competition and received $100,000. </observation>

T5
<think> The athlete and prize are corroborated by multiple sources. </think>
<response> The 200m champion, Matthew Hudson-Smith, earned $100,000. </response>

Gold / judged answer: $100,000

Figure 8: A successful FVQA trajectory. The student localizes the athlete's bib, identifies Matthew Hudson-Smith, and verifies the $100,000 prize through focused retrieval.



input image

User Question: Which movie directed by the brother of this man has the highest ranking?

T1
<think> The face is the key clue, but the photograph is soft. I will sharpen it before attempting identification. </think>
<tool_call> {"name":"sharpen","arguments":{"image":"img_1","amount":1.5}} </tool_call>
<observation> The enhanced portrait is returned as img_2. </observation>

T2
<think> I will use the sharpened portrait for visual identification. </think>
<tool_call> {"name":"image_search","arguments":{"url":"img_2"}} </tool_call>
<observation> Visual matches identify the speaker as Jonathan Nolan, including a Wikipedia portrait and references to The Dark Knight. </observation>

T3
<think> The question asks about his brother, so I will verify the family relationship. </think>
<tool_call> {"name":"text_search","arguments":{"q":"Jonathan Nolan brother","hl":"en","top_k":5}} </tool_call>
<observation> The results identify film director Christopher Nolan as Jonathan Nolan's brother. </observation>

T4
<think> I now need the highest-ranked film directed by Christopher Nolan. </think>
<tool_call> {"name":"text_search","arguments":{"q":"Christopher Nolan movies ranking highest IMDb","hl":"en","top_k":5}} </tool_call>
<observation> IMDb ranking results place The Dark Knight first, ahead of Inception, Interstellar, and The Prestige. </observation>

T5
<think> Jonathan Nolan is identified in the image, Christopher Nolan is his brother, and the ranking is explicit. </think>
<response> The highest-ranked movie directed by his brother is The Dark Knight. </response>

Gold / judged answer: The Dark Knight

Figure 9: A successful BrowseComp-VL trajectory involving image enhancement, visual identification, relation retrieval, and ranking-based factual verification.

E ETHICS STATEMENT

This work advances the study of multimodal agents that acquire external evidence through search and image-processing tools. Our experiments use established research benchmarks and involve no human-subject studies or collection of private personal data. By design, the experimental agent operates in a read-only environment: it retrieves publicly accessible information and processes benchmark images, while remaining strictly confined to information retrieval and analysis without side effects on external systems. We preserve source URLs for traceability and use datasets and external services in accordance with their applicable licenses and terms. These design choices enable transparent, reproducible, and responsible evaluation of tool-augmented agents. We conduct and report this work in accordance with the ICLR Code of Ethics.