IGSD: ENVIRONMENT-VERIFIED HINDSIGHT SELF-DISTILLATION FOR SEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation densifies agent training without external teachers: a policy conditioned on privileged hindsight provides step-level guidance for its own unprivileged rollouts. For search agents, however, hindsight can make the teacher prefer a query that does not improve retrieval from the student's state. Existing methods either distill this preference directly or filter it with model-internal scores, but neither strategy verifies the query's executed retrieval consequence. We propose **Information-Gain-Gated Self-Distillation (IGSD)**, which verifies on-policy token proposals with environment feedback before distilling them. Treating each query token as a micro-action, IGSD completes the teacher's token proposal and the student's sampled token into matched queries and executes both from the same failed state with the same retriever. Shared counterfactual controls account for query-conditioned shifts in answer likelihood, so their difference, the *executed paired information gain*, provides a relative utility contrast for the retrieved documents. IGSD uses this contrast as a positive-only soft weight for candidate-pair distillation, while leaving the GRPO objective unchanged and confining verification to training. Across seven single-hop and multi-hop QA benchmarks, IGSD reaches macro-average exact-match accuracies of 42.8% and 47.0% with 3B and 7B policies, respectively, without inference-time verification. These results support environment-verified hindsight as an effective approach to reliable action-level supervision for search agents.

1 INTRODUCTION

Search-augmented reasoning agents solve information-seeking tasks by interleaving internal reasoning with external retrieval operations (Yao et al., 2023; Jin et al., 2025; Chen et al., 2025). Training frameworks such as Search-R1 (Jin et al., 2025) optimize these trajectories with reinforcement learning under trajectory-level outcome rewards (Shao et al., 2024; Guo et al., 2025). Because the reward depends only on the final answer, an informative query and a redundant or misleading query can receive the same trajectory-level credit, leaving intermediate search decisions without fine-grained supervision.

To densify intermediate supervision, recent work introduces *hindsight self-distillation* (HSD) (Ma et al., 2026; Lu et al., 2026). The policy first rolls out without privileged context, while a self-teacher conditioned on hindsight information evaluates or revises intermediate steps and distills supervision back into the unprivileged policy. For search agents, however, a query that proved useful in a successful sibling may be redundant or misleading at the student's failed state: the trajectory may already contain the relevant evidence, or may lack the bridge entity required by the query. Hindsight preference is therefore not sufficient evidence that the proposed query will improve retrieval from the student's current state.

Existing methods can filter hindsight supervision with model-internal signals, such as the detached teacher–student likelihood gap used by SDAR (Lu et al., 2026), yet any such score reflects the model's preference, not whether the proposal actually brings the new evidence the current student needs. This creates a *teacher-supervision reliability gap*: how to verify, before distillation, whether teacher guidance genuinely helps. In search agents, however, every search step naturally yields environment feedback, which makes this question answerable by execution rather than by introspection.

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.

We propose **Information-Gain-Gated Self-Distillation (IGSD)**, which grounds hindsight guidance in environment-derived utility rather than model preference. Working on-policy at the states the student actually visited, the hindsight teacher proposes token-level alternatives; whenever a proposal diverges from the student's sampled action, IGSD lets the student policy complete both candidates into matched queries and executes them with the same retriever. Each execution is scored by how much its retrieved documents raise the policy's answer likelihood, net of a shared counterfactual control, and their difference, the *executed paired information gain*, is an environment-derived utility contrast at the granularity of a single action. IGSD uses it as a positive-only soft weight on candidate-pair distillation, so helpful proposals receive stronger supervision while non-improving ones are omitted. The GRPO objective remains unchanged, and verification is confined to training.

Experiments across seven single-hop and multi-hop QA benchmarks show that IGSD achieves the highest macro-average exact-match accuracy among the methods, reaching 42.8% with the 3B policy and 47.0% with the 7B policy, with no inference-time verification. Our code is available at `https://anonymous.4open.science/r/igsd_code-520A`.

Our main contributions are as follows:

- **Environment-verified hindsight distillation.** We introduce IGSD, which converts hindsight teacher suggestions into local candidate actions and verifies their retrieval consequences from the student's state before distillation.
- **Paired utility gating.** We define executed paired information gain from matched teacher and student query executions with shared counterfactual controls, and use it as a positive-only soft weight on the candidate-pair distillation loss while leaving the GRPO objective unchanged.
- **Empirical and mechanistic analysis.** We validate the effectiveness of IGSD across seven single-hop and multi-hop QA benchmarks at both 3B and 7B scales, and an extensive mechanistic analysis shows stronger observed alignment with terminal outcomes than the tested model-internal signals.

2 Related Work

Search-agent RL. Search-R1 (Jin et al., 2025) trains language models to interleave reasoning and retrieval with a trajectory-level outcome reward. ReSearch (Chen et al., 2025) extends this setting to multi-turn reasoning and search control. Other work introduces denser step- or action-level signals, including IGPO (Wang et al., 2025) and GiGPO (Feng et al., 2025). IGSD differs by adding a paired, environment-executed comparison of candidate search actions from the same local state, while preserving the trajectory-level objective and introducing the action-level signal only through a soft distillation weight.

Hindsight self-distillation. SD-Search (Ma et al., 2026) applies on-policy hindsight self-distillation to search agents, but does not perform per-candidate retrieval verification. Search-E1 (Liang et al., 2026b) alternates GRPO updates with privileged-context self-distillation. SDAR (Lu et al., 2026) gates per-token distillation with a detached teacher–student likelihood gap, which measures an internal preference rather than the utility of retrieved evidence. Skill-SD (Wang et al., 2026) conditions privileged self-distillation on skills for multi-turn agents. IGSD complements these approaches by executing teacher-induced and student-induced queries from the same state and using their paired environment-derived utility to decide whether distillation is applied.

Information gain as supervision. IG-Search (Liang et al., 2026a) and IGPO (Wang et al., 2025) introduce information gain as an RL signal. SIGHT (Zhong et al., 2026) uses it for observation-level de-duplication, reflection, and branching, while SmartSearch (Wen et al., 2026) applies it to query verification and refinement. CRAFT (Meng & Chen, 2026) and LOTAPO (Zhu et al., 2026) estimate token- or process-level credit from sibling rollouts. IGSD instead uses paired executed gain only to weight auxiliary distillation, without changing the policy-gradient objective.

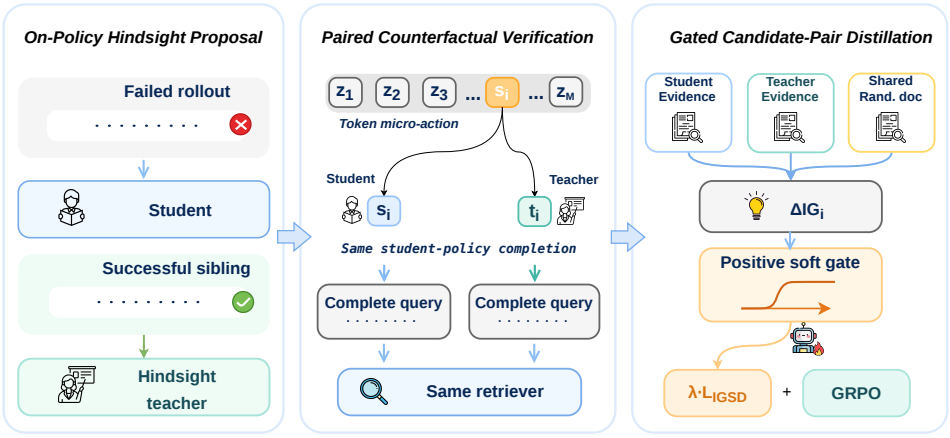

Figure 1: Overview of IGSD. Left: privileged hindsight produces a teacher proposal for a failed student rollout. Middle: the teacher and sampled student tokens are independently completed by the same frozen student policy into matched queries and evaluated with the same retriever. Right: the resulting environment-derived utility contrast $\Delta\mathrm{IG}_i$ gates candidate-pair distillation together with the unchanged GRPO objective.

3 METHOD

IGSD augments GRPO with environment-verified action-level supervision. As shown in Figure 1, a hindsight-conditioned teacher proposes local alternatives to a failed student trajectory, IGSD executes the resulting candidate queries under matched conditions, and the utility contrast measured on the retrieved evidence gates candidate-pair distillation while leaving the GRPO objective unchanged. Section 3.1 formalizes the search-agent setting; Section 3.2 constructs hindsight-guided candidates; Section 3.3 defines the paired environment verification that produces the executed information gain; and Section 3.4 presents the gain-gated distillation objective.

3.1 PROBLEM SETUP

Search-agent trajectories. Given a question x, a search policy π_θ generates a multi-turn trajectory τ that interleaves reasoning, search actions, and retrieved observations before producing a terminal answer. After standard answer normalization, the answer evaluator returns a scalar terminal outcome score $r(\tau)$ rather than a strict binary exact-match label; partially matching answer strings can therefore receive nonzero score. We use its answer-correctness signal to label trajectories as successful or failed, while retaining the scalar score for sibling ranking and the hindsight hint. GRPO optimizes this terminal score over rollouts for the same question, but does not identify which intermediate search decisions improve retrieved evidence. IGSD derives such local supervision while leaving the GRPO reward and advantage estimator unchanged.

3.2 HINDSIGHT-GUIDED CANDIDATE CONSTRUCTION

Student trajectory. IGSD constructs auxiliary supervision for a failed trajectory whose rollout group contains a successful sibling and at least one executable search action. It selects the last valid search action for supervision. Let h_S denote the prefix before this action and $z = (z_1, \ldots, z_M)$ its serialized tool call. The tokenizer positions inside the search-query field are $\mathcal{I}_Q = \{k_1, \ldots, k_N\}$, excluding tool-call syntax and other fields. The sampled student token at position k_i is $s_i = z_{k_i}$.

Hindsight teacher. A selected successful sibling provides a search-strategy hint containing its search queries Q^+ and outcome score r^+, denoted by $\mathcal{H}^+ = (Q^+, r^+)$. The hint excludes the sibling's terminal answer, reasoning, retrieved documents, and tool observations. The teacher uses the same frozen policy checkpoint as the rollout policy but is conditioned on the student trajectory prefix and the hindsight hint. When multiple successful siblings are available, we select the one with

the highest terminal reward, breaking ties by rollout order:

$$h_T = \mathcal{P}_T(x, h_S, \mathcal{H}^+), \tag{1}$$

Evaluating teacher and student along the same sampled action prefix gives

$$p_i^T(v) = \pi_{\theta_{\text{old}}}(v \mid h_T, z_{<k_i}), \qquad p_i^S(v) = \pi_{\theta_{\text{old}}}(v \mid h_S, z_{<k_i}). \tag{2}$$

The two distributions share the decision state and differ only in the presence of the hindsight hint.

Candidate micro-actions. We treat each query token as a micro-action. At position k_i, the teacher proposes $t_i = \arg\max_v p_i^T(v)$: where $t_i = s_i$ the teacher endorses the student's action, and each disagreement defines a candidate pair $\mathcal{S}_i = \{t_i, s_i\}$ to be verified on-policy, by executing both actions at the exact state visited by the student's own rollout and comparing the evidence they retrieve (Section 3.3).

3.3 Paired Environment Verification

Matched query completion. IGSD converts each candidate pair into two executable queries by fixing the candidate at position k_i and greedily completing the remainder with the frozen student policy, without the hindsight hint:

$$Q_i^T = \text{Complete}_{\pi_{\theta_{\text{old}}}}(h_S, z_{<k_i} \oplus t_i), \qquad Q_i^S = \text{Complete}_{\pi_{\theta_{\text{old}}}}(h_S, z_{<k_i} \oplus s_i). \tag{3}$$

The two branches share the state, prefix, continuation policy, decoding procedure, and retriever; only the fixed token at k_i differs. The resulting comparison measures the downstream utility of a local token intervention under the shared continuation policy. We call them the teacher-induced and student-induced queries, with retrieved document sets D_i^T and D_i^S.

Document utility and counterfactual control. IGSD scores document set D by the frozen policy's mean reference-answer token log-likelihood. For branch $b \in \{T, S\}$ and answer aliases $\mathcal{A}_i$, define

$$L_i^b(D; \mathcal{A}_i) = \max_{a \in \mathcal{A}_i} \log \pi_{\theta_{\text{old}}}(a \mid h_S, Q_i^b, D). \tag{4}$$

Here, the log-likelihood of an alias denotes the mean over its encoded answer tokens. The maximum lets any valid surface form of the reference answer carry the score. The score L_i^b, however, also responds to the query itself: a query's phrasing can shift the policy's answer likelihood even when its retrieved evidence is unchanged. To reduce this query-dependent component as much as possible while keeping the comparison focused on retrieved evidence, IGSD scores every branch against a shared counterfactual control: for each candidate pair it draws C counterfactual document sets $\widetilde{D}_{ij}$ at random from the corpus, each matched in size and format to the retrieved documents.

Executed paired information gain. The executed paired information gain compares the controlled document utilities of the two branches:

$$\Delta\text{IG}_i = \frac{1}{C} \sum_{j=1}^{C} \Big[\big(L_i^T(D_i^T; \mathcal{A}_i) - L_i^T(\widetilde{D}_{ij}; \mathcal{A}_i)\big) - \big(L_i^S(D_i^S; \mathcal{A}_i) - L_i^S(\widetilde{D}_{ij}; \mathcal{A}_i)\big) \Big]. \tag{5}$$

A positive ΔIG_i indicates that the evidence retrieved by the teacher-induced query contributes more answer support than the evidence retrieved by the student's own query. This environment-derived utility contrast weights the auxiliary distillation loss but does not enter the GRPO reward or advantage estimator.

3.4 Gain-Gated Candidate-Pair Distillation

Positive-only utility gate. The executed gain determines whether a candidate pair contributes to distillation and how strongly it is weighted. IGSD defines the detached soft gate

$$w_i = \text{sg}\left[\max\left(0,\, \tanh\left(\frac{\beta}{2} \Delta\text{IG}_i\right)\right)\right], \tag{6}$$

where sg denotes stop-gradient and β controls the sharpness; candidate pairs with non-positive gain receive zero auxiliary weight.

Candidate-pair target. For each accepted pair, the frozen teacher target and current student distribution are renormalized on $\mathcal{S}_i = \{t_i, s_i\}$:

$$\bar{p}_i^T(v) = \frac{p_i^T(v)}{\sum_{u \in \mathcal{S}_i} p_i^T(u)}, \qquad \bar{q}_i^\theta(v) = \frac{\pi_\theta(v \mid h_S, z_{<k_i})}{\sum_{u \in \mathcal{S}_i} \pi_\theta(u \mid h_S, z_{<k_i})}, \qquad v \in \mathcal{S}_i. \tag{7}$$

The per-position loss is the Jensen-Shannon divergence between these distributions:

$$\mathcal{L}_{\text{pair}}(i) = \frac{1}{2}\,\mathrm{KL}(\bar{p}_i^T \,\|\, \mu_i) + \frac{1}{2}\,\mathrm{KL}(\bar{q}_i^\theta \,\|\, \mu_i), \qquad \mu_i = \frac{\bar{p}_i^T + \bar{q}_i^\theta}{2}. \tag{8}$$

Verification and distillation play complementary roles: executed paired information gain determines whether a candidate pair is retained and how strongly it is weighted, while the Jensen–Shannon divergence transfers the verified teacher preference over the two adjudicated tokens.

Joint training objective. Let $\mathcal{V}_Q = \{i \in \mathcal{I}_Q : \Delta\mathrm{IG}_i > 0\}$ denote the accepted query positions. The query-level auxiliary loss and the actor objective are

$$\mathcal{L}_{\text{IGSD}}(Q) = \frac{1}{|\mathcal{I}_Q|} \sum_{i \in \mathcal{V}_Q} w_i \mathcal{L}_{\text{pair}}(i), \qquad \mathcal{L}_{\text{actor}} = \mathcal{L}_{\text{GRPO}} + \lambda\, \mathbb{E}_Q[\mathcal{L}_{\text{IGSD}}(Q)]. \tag{9}$$

We normalize by all query-token positions, so rejected positions contribute zero and the auxiliary term is not inflated solely by the number of accepted disagreements. Verification is training-only; inference uses the standard search agent without a teacher, verification branches, or extra retrieval calls.

4 Experiments

We evaluate IGSD along three axes. **RQ1:** Does verified hindsight distillation improve trained search policies? **RQ2:** Does executed paired information gain align with downstream utility? **RQ3:** How much does each component of the decision chain contribute to the trained policy?

4.1 Experimental Setup

Datasets. Following Search-R1 (Jin et al., 2025), we train on the combined training splits of NQ and HotpotQA, and evaluate on the evaluation sets of seven benchmarks: NQ, TriviaQA, and PopQA (single-hop), and HotpotQA, 2WikiMultiHopQA, MuSiQue, and Bamboogle (multi-hop). We report exact-match (EM) accuracy per dataset and the macro-average across the seven benchmarks.

Baselines. Inference-only references are Direct Generation, Vanilla RAG, and Search-o1 (Li et al., 2025); outcome-reward RL includes Search-R1 (Jin et al., 2025), ReSearch (Chen et al., 2025), and AutoRefine (Shi et al., 2025); process-level supervision includes IGPO (Wang et al., 2025), GiGPO (Feng et al., 2025), CRAFT (Meng & Chen, 2026), SD-Search (Ma et al., 2026), and LOTAPO (Zhu et al., 2026). We also report IGSD w/o Gate (Section 4.4), the same recipe without verification gating.

Implementation details. We use Qwen2.5-3B-Instruct (Qwen Team, 2025) as the main policy and also evaluate IGSD with Qwen2.5-7B-Instruct. All locally trained methods are implemented with SearchAgent-Zero (Li, 2026), which builds on the veRL training framework (Sheng et al., 2025) and adds search-agent-specific components, including the multi-turn search pipeline and tool interface. Training uses batches of 256 prompts and five rollouts per prompt, a learning rate of 10^{-6}, and 400 optimization steps. All locally trained runs use eight NVIDIA A100 80GB GPUs. We use E5-base-v2 (Wang et al., 2022) with the December 2018 Wikipedia corpus (Karpukhin et al., 2020) and retrieve the top three passages per search call, and the policy may take at most six assistant turns per trajectory. GRPO uses the terminal answer evaluator's reward, and evaluation uses greedy decoding with one response per question; all locally trained methods are evaluated at the final checkpoint. Full hyperparameter details are provided in Appendix F.

4.2 Main Results (RQ1)

Table 1: Main results with Qwen2.5-3B-Instruct. Bold and underline indicate the best and second-best values in each column.

	Single-Hop			Multi-Hop				
Method	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle	Avg.
Inference-only								
Direct Generation	0.106	0.288	0.108	0.149	0.244	0.020	0.024	0.134
Vanilla RAG	0.348	0.544	0.387	0.255	0.226	0.047	0.080	0.270
Search-o1	0.238	0.472	0.262	0.221	0.218	0.054	0.320	0.255
Outcome-reward RL								
Search-R1	0.439	0.591	0.423	0.388	0.404	0.185	0.328	0.394
ReSearch	0.387	0.581	0.410	0.302	0.236	0.124	0.266	0.329
AutoRefine	0.436	0.597	0.411	0.397	0.418	**0.187**	0.352	0.400
Process-level supervision								
IGPO	0.445	0.604	0.434	0.389	0.405	0.171	0.312	0.394
GiGPO	0.452	0.605	0.442	0.389	0.387	0.139	0.280	0.385
CRAFT	0.436	0.594	0.414	0.387	0.410	0.176	0.368	0.398
SD-Search	**0.460**	0.612	0.424	0.415	0.428	0.181	0.384	0.415
LOTAPO	0.434	0.585	0.425	0.389	0.395	0.178	0.336	0.392
IGSD and controls								
IGSD w/o Gate	0.455	**0.623**	0.448	0.401	0.407	0.184	0.400	0.417
IGSD	0.458	0.620	**0.461**	**0.420**	**0.443**	0.183	**0.408**	**0.428**

Table 2: Results with Qwen2.5-7B-Instruct. Notation as in Table 1.

	Single-Hop			Multi-Hop				
Method	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle	Avg.
Direct Generation	0.134	0.408	0.140	0.183	0.250	0.031	0.120	0.181
Vanilla RAG	0.349	0.585	0.392	0.299	0.235	0.058	0.208	0.304
Search-o1	0.151	0.443	0.131	0.187	0.176	0.058	0.296	0.206
Search-R1	0.477	0.647	**0.480**	0.452	0.435	0.202	0.424	0.445
SD-Search	0.492	0.652	0.469	0.472	0.438	0.228	0.440	0.456
IGSD w/o Gate	0.489	0.655	0.472	0.465	0.436	0.216	0.440	0.453
IGSD	**0.496**	**0.656**	0.476	**0.478**	**0.489**	**0.232**	**0.464**	**0.470**

At 3B, IGSD reaches a 0.428 macro-average (Table 1). The results support two conclusions. Against outcome-reward RL, IGSD improves over the strongest baseline, AutoRefine (0.400), with gains on three of the four multi-hop benchmarks and a +0.025 gain on 2WikiMultiHopQA, confirming that step-level supervision adds value beyond trajectory-level rewards. Against process-level supervision, IGSD exceeds SD-Search (0.415), with the largest per-dataset separation on PopQA (+0.037), and the ungated control sharpens the attribution: it shares IGSD's teacher context, candidate construction, and auxiliary objective, differing only in the gate, and its 0.417 macro-average therefore assigns the remaining +0.011 to executed-gain weighting. Verified gating is thus a necessary component of the process signal, not a refinement of it. Three-seed averages preserve IGSD as the strongest configuration; the complete results are reported in Appendix C.



Figure 2: Fixed-validation macro-average throughout training for Search-R1, IGSD w/o Gate, and IGSD.

At 7B (Table 2), IGSD remains ahead, reaching a 0.470 macro-average and retaining a clear advantage over its ungated control (0.453). It also exceeds Search-R1 (0.445) and SD-Search (0.456), leading on all four multi-hop benchmarks. The benefit of environment-verified supervision thus transfers to the larger scale.

Figure 2 traces the effect of verification on training stability. Search-R1 improves early but plateaus after roughly the first 200 steps. IGSD w/o Gate fluctuates throughout training, as unverified hind-

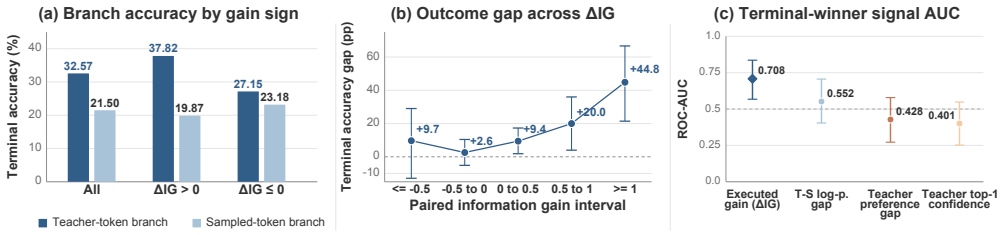

Figure 3: Mechanistic evidence for executed paired information gain. Panels (a)–(c) report branch accuracy by gain sign, teacher–student accuracy gaps across fixed gain intervals, and terminal-winner ROC-AUC for executed and model-internal signals. Error bars in (b) and (c) are 95% question-cluster bootstrap confidence intervals; both branches use independent completions with the same frozen student policy and decoding.

sight supervision can help or hurt depending on the current search state. IGSD continues to improve across the full schedule and shows a smoother validation trajectory: filtering proposals by their executed gain retains useful hindsight supervision while suppressing proposals with unfavorable utility. Complete training-time and search-usage measurements are reported in Appendix D.

4.3 Mechanistic Evidence (RQ2)

We next test whether executed paired information gain aligns with downstream utility. Each candidate pair is evaluated by fixing the candidate tokens in turn, completing both branches independently with the same frozen student policy and decoding procedure, and measuring the exact-match (EM) accuracy of the final answer. Across the 307 evaluated candidate pairs, the teacher-token branch reaches 0.326 EM on average versus 0.215 for the sampled-token branch. Splitting the pairs by gain sign (Figure 3a) reveals a less uniform pattern: for $\Delta\mathrm{IG} \leq 0$, individual pairs exhibit both positive and negative teacher–student utility gaps, whereas for $\Delta\mathrm{IG} > 0$, the teacher-token branch shows a larger aggregate advantage (0.378 vs. 0.199 EM).

Figure 3b examines the same pairs across fixed $\Delta\mathrm{IG}$ intervals. Its vertical axis reports the paired terminal-accuracy gap, defined as teacher-token accuracy minus sampled-token accuracy in percentage points. The gap is small and variable in the low-gain region, including $\Delta\mathrm{IG} \leq 0$, then increases markedly for positive gains, reaching +20.0 points for the 0.5–1 interval and +44.8 points for $\Delta\mathrm{IG} \geq 1$. The pattern supports using executed gain as a graded utility contrast. Both branches share the same student state, frozen completion policy, decoding procedure, retriever, and counterfactual control; only the candidate token fixed at the inspected position differs before the branches are independently completed and executed. The resulting $\Delta\mathrm{IG}$ is therefore used as an environment-derived weight for the auxiliary distillation loss.

Finally, Figure 3c compares executed paired information gain with three model-internal token-probability signals for ranking the terminal-winning branch among the 72 terminal-discordant pairs. The y-axis reports ROC-AUC; the four x-axis entries are executed paired information gain, the teacher–student log-probability gap on the sampled token, the teacher's log-probability difference between its top-1 and the sampled token, and the teacher's log-probability for its top-1 token. Executed gain achieves the highest point estimate (AUC 0.708; 95% question-cluster bootstrap CI [0.568, 0.836]), compared with 0.552 for the strongest internal signal, the sampled-token log-probability gap. This result shows that the executed signal provides the strongest tested alignment with the terminal-winning branch.

4.4 Ablation Studies (RQ3)

We next isolate the contributions of IGSD's decision-chain components. Keeping the default training recipe fixed, we vary supervision scope, paired verification, gain weighting, and distillation support one at a time. Table 3 reports the core comparisons, while Figure 4 and Appendices E–F provide the summary and full per-dataset results. Unless explicitly stated, all variants use the default configuration described in Sections 3 and 4.1.

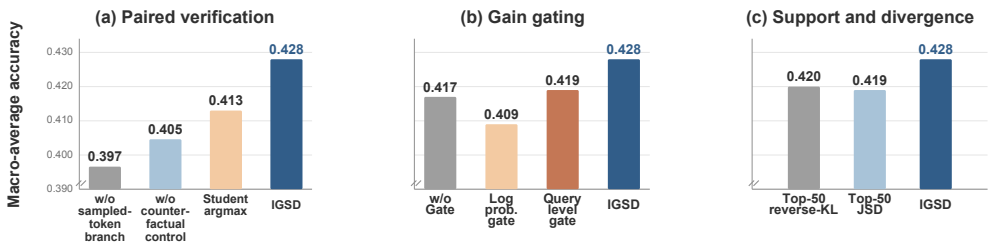

Figure 4: Ablation of IGSD's decision chain with the 3B policy. Panels group variants by paired verification, gain gating, and distillation support and divergence.

Table 3: Component ablations with the Qwen2.5-3B-Instruct policy under the recipe of Section 4.1. The best and second-best values per column are in **bold** and underlined.

Variant	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle	Avg.
Search-R1	0.439	0.591	0.423	0.388	0.404	0.185	0.328	0.394
All-turn verification	0.455	0.615	0.451	0.417	**0.444**	**0.201**	0.368	0.421
w/o sampled-token branch	0.437	0.599	0.412	0.397	0.407	0.156	0.368	0.397
w/o counterfactual control	0.438	0.597	0.408	0.406	0.420	0.180	0.384	0.405
w/o Gate	0.455	**0.623**	0.448	0.401	0.407	0.184	0.400	0.417
Log-probability gate	0.450	0.609	0.438	0.412	0.421	0.186	0.344	0.409
IGSD	**0.458**	0.620	**0.461**	**0.420**	0.443	0.183	**0.408**	**0.428**

Supervision scope. Verification supervises the last valid search turn of a failed trajectory. Verifying every valid turn instead (*all-turn verification*) reaches a macro-average of 0.421, with no improvement over the 0.428 reference. We therefore focus on the last valid turn. By then, the accumulated context makes teacher–student disagreement more likely to expose a concrete query-level alternative, yielding more targeted candidate pairs for distillation.

Paired verification. The paired-verification ablation isolates the two components of the executed comparison. Removing the sampled-token branch while retaining the other verification machinery (*w/o sampled-token branch*) gives a macro-average of 0.397, while removing the counterfactual control (*w/o counterfactual control*) gives 0.405. Replacing the sampled student token with the student policy's arg-max token at the same decision position (*student argmax*) gives 0.413. The full paired construction reaches 0.428 (Figure 4a). Each component removal lowers the macro-average, supporting the matched sampled-token reference and the shared counterfactual control in the executed comparison.

Gain gating. The gate ablation keeps the candidate pairs, teacher target, and loss fixed while changing only the weight signal. Unit weights without verification (*w/o Gate*) reach 0.417. Replacing executed gain with the sampled- token log-probability gap $\log p_T(s_i) - \log p_S(s_i)$ (*log-probability gate*) lowers the result to 0.409. In the query-level variant, the executed gains of all verified token positions in one query are averaged and the same weight is applied to that query's pairs; this reaches 0.419. The per-pair executed-gain gate reaches 0.428. The gap between the per-pair and query-level variants indicates that averaging can wash out local differences between candidate actions, while the lower log-probability-gate result shows that a model-internal preference is not an adequate substitute for executed utility (Figure 4b).

Support and divergence. The final ablation varies the support and divergence of the auxiliary loss. Widening the two-token target to the de-duplicated teacher–student top-50 union (*top-50 JSD*) gives 0.419, while *top-50 reverse-KL* gives 0.420, versus 0.428 for candidate-pair JSD. In these variants, broader supports lower the macro-average, consistent with assigning distillation mass to tokens that were not directly adjudicated by environment execution (Figure 4c; Table 12).

Taken together, the ablations favor a localized decision chain: the teacher proposal is compared against the action actually sampled by the student, the comparison is made after query execution, and

the resulting paired utility contrast is applied only to the corresponding candidate pair. Removing one of these localization steps weakens the final result in the tested setting. The complete per-dataset variants and counterfactual-draw analysis are deferred to Appendix E.

4.5 Qualitative Cases

Table 4 reports terminal outcomes from the actual teacher-token and sampled-token branches after they disagree at the inspected decision. Each branch fixes its candidate token, independently completes the query under the same frozen student policy, executes retrieval, and follows the resulting continuation to a terminal answer. In Case A, the student query leaves the already identified entity unresolved by using the vague phrase *the actress*. The hindsight teacher instead recognizes Berlinda Tolbert in the retrieved evidence, produces a query grounded on that entity, and retrieves the correct sitcom, The Jeffersons ($\Delta\mathrm{IG} = +3.58$, retained). In Case B, the current evidence has already identified Sigmund Freud. The sampled-token branch directly queries for Freud's father and retrieves Jacob Freud, whereas the hindsight teacher redundantly re-encodes the already resolved relation as a nested query and returns Sigmund Freud again. The negative $\Delta\mathrm{IG}$ therefore rejects this redundant teacher proposal ($\Delta\mathrm{IG} = -0.198$), rather than distilling it. Further cases are given in Appendix G.

Table 4: Representative paired search decisions, one retained and one rejected by the gate. Blue and orange mark the student and teacher query spans; $\Delta\mathrm{IG}$ is teacher minus student information gain.

A Positive $\Delta\mathrm{IG}$ — hindsight resolves the entity — **RETAIN**

Question: An actress in Strange Fruit was better known for a role in a CBS sitcom. Which sitcom is this actress better known for than her role in Strange Fruit? [Reference: The Jeffersons]
State: Strange Fruit is a 2004 film … starring Kent Faulcon as William Boyals and **Berlinda Tolbert** as Emma Ayers (actress identified; her sitcom unresolved).

Student-sampled action	Hindsight-teacher proposal
Search: which CBS sitcom was **the actress** better known for than her role in Strange Fruit	**Search:** which CBS sitcom was **Berlinda Tolbert** better known for than her role in Strange Fruit
Retrieved: Ricki Lake … had a recurring role on the CBS sitcom **The King of Queens** …	**Retrieved: Berlinda Tolbert** … best known for her role as Jenny Willis Jefferson … on the CBS sitcom **The Jeffersons** …
Answer: The King of Queens — incorrect	**Answer:** The Jeffersons — correct
$\Delta\mathrm{IG} = +3.58 > 0$	

B Negative $\Delta\mathrm{IG}$ — hindsight repeats the unresolved relation — **REJECT**

Question: Who was the father of the father of psychoanalysis? [Reference: Jacob Freud]
State: The discipline was established in the early 1890s by Austrian neurologist **Sigmund Freud** and stemmed partly from the clinical work of Josef Breuer and others …

Student-sampled action	Hindsight-teacher proposal
Search: who was the father of **Sigmund Freud**	**Search:** who was the father of the **father of psychoanalysis**
Retrieved: Jacob Koloman Freud (1815–1896) was the father of **Sigmund Freud**, the founder of psychoanalysis. Born in Tysmenytsia in Austrian Galicia …	**Retrieved:** The discipline was established in the early 1890s by Austrian neurologist **Sigmund Freud** and stemmed partly from the clinical work of Josef Breuer and others …
Answer: Jacob Freud — correct	**Answer:** Sigmund Freud — incorrect
$\Delta\mathrm{IG} = -0.1978 < 0$	

5 Limitations

IGSD relies on reference-answer supervision to estimate retrieved-evidence utility, so it applies directly to tasks with an explicit answer or reliable outcome evaluator, but does not directly extend to open-ended tasks lacking an explicit correctness criterion. Training adds branch completion, execution, and scoring, but localized verification and a bounded counterfactual budget keep this cost manageable; inference is unchanged.

6 Conclusion

IGSD introduces environment-verified hindsight self-distillation for search agents. By executing matched teacher and student query proposals from the same student state and applying shared counterfactual controls, it derives an executed paired information gain that weights candidate-pair distillation according to retrieval utility. This grounds auxiliary supervision in on-policy environment feedback while leaving the GRPO objective and inference procedure unchanged. Across seven QA benchmarks and both 3B and 7B policies, IGSD improves macro-average exact-match accuracy over the evaluated baselines, and executed gain shows stronger alignment with terminal utility than the tested model-internal signals. These results establish environment-verified utility as a practical basis for action-level supervision in search agents.

AI USE STATEMENT

We used generative AI tools to polish the language and presentation of this manuscript and to support an initial survey and organization of related work. All technical claims, citations, experimental results, and final wording were reviewed and verified by the authors, who take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

Our code is available at `https://anonymous.4open.science/r/igsd_code-520A`. The paper and appendix provide the benchmark statistics, data split conventions, retrieval and tool settings, prompts, hyperparameters, evaluation protocol, random-seed results, and ablation configurations used in our study. These details are intended to make the experimental setup and comparisons independently checkable.

REFERENCES

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. Research: Learning to reason with search for llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2503.19470.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *arXiv preprint arXiv:2505.10978*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Ö. Arık, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP*, 2020.

Jiacheng Li. Searchagent-zero: Scalable reinforcement learning for multi-turn search agents. https://github.com/NLPJCL/SearchAgent-Zero, 2026.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*, 2025.

Zihan Liang, Yufei Ma, Ben Chen, Zhipeng Qian, Huangyu Dai, Lingtao Mao, Xuxin Zhang, Chenyi Lei, and Wenwu Ou. Ig-search: Step-level information gain rewards for search-augmented reasoning. *arXiv preprint arXiv:2604.15148*, 2026a.

Zihan Liang, Yufei Ma, Ben Chen, Zhipeng Qian, Xuxin Zhang, Huangyu Dai, and Lingtao Mao. Search-e1: Self-distillation drives self-evolution in search-augmented reasoning. *arXiv preprint arXiv:2605.22511*, 2026b.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2605.15155*, 2026.

Yufei Ma, Zihan Liang, Ben Chen, Zhipeng Qian, Huangyu Dai, Lingtao Mao, Xuxin Zhang, Chenyi Lei, and Wenwu Ou. Sd-search: On-policy hindsight self-distillation for search-augmented reasoning, 2026. URL https://arxiv.org/abs/2605.18299.

Zibin Meng and Kani Chen. Craft: Counterfactual credit assignment from free sibling rollouts for self-distilled agentic reinforcement learning, 2026. URL https://arxiv.org/abs/2606.29476.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *EuroSys*, 2025.

Yaorui Shi, Sihang Li, Chang Wu, Zhiyuan Liu, Junfeng Fang, Hengxing Cai, An Zhang, and Xiang Wang. Search and refine during think: Facilitating knowledge refinement for improved retrieval-augmented reasoning, 2025. URL https://arxiv.org/abs/2505.11277.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. Information gain-based policy optimization: A simple and effective approach for multi-turn search agents. *arXiv preprint arXiv:2510.14967*, 2025.

Hao Wang, Guozhi Wang, Han Xiao, Yufeng Zhou, Yue Pan, Jichao Wang, Ke Xu, Yafei Wen, Xiaohu Ruan, Xiaoxin Chen, and Honggang Qi. Skill-sd: Skill-conditioned self-distillation for multi-turn llm agents, 2026. URL https://arxiv.org/abs/2604.10674.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Tongyu Wen, Guanting Dong, and Zhicheng Dou. Smartsearch: Process reward-guided query refinement for search agents. *arXiv preprint arXiv:2601.04888*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR*, 2023.

Wenlin Zhong, Jinluan Yang, Yiquan Wu, Yi Liu, Jianhang Yao, and Kun Kuang. Sight: Reinforcement learning with self-evidence and information-gain diverse branching for search agent. *arXiv preprint arXiv:2602.11551*, 2026.

Qiang Zhu, Jiajun Wu, and Longyi Wang. Lotapo: Leave-one-turn attribution for self-generated process rewards in multi-turn search reasoning. *arXiv preprint arXiv:2607.13501*, 2026.

A Experimental Details

A.1 Data Statistics

We use the same seven-benchmark collection and split convention as SD-Search (Ma et al., 2026). The datasets are sourced from the FlashRAG collection. Following Search-R1 (Jin et al., 2025), the training set combines the training splits of NQ and HotpotQA. We use the test split when it is available and otherwise use the development split for evaluation.

Table 5: Dataset statistics for the seven question-answering benchmarks used in our experiments. A dash indicates that the corresponding split is not available.

Dataset	Train	Dev	Test	Eval split
NQ	79,168	8,757	3,610	test
TriviaQA	78,785	8,837	11,313	test
PopQA	–	–	14,267	test
HotpotQA	90,447	7,405	–	dev
2WikiMultiHopQA	15,000	12,576	–	dev
MuSiQue	19,938	2,417	–	dev
Bamboogle	–	–	125	test

A.2 Counterfactual Sampling Protocol

During training, counterfactual document sets are sampled from the batch-level pool of passages retrieved in the current training batch. Duplicate passages are removed before sampling. For each candidate pair, a counterfactual set contains the same number of passages as the corresponding executed branch; passages retrieved by that branch are excluded, and the remaining passages are sampled uniformly without replacement. The same sampled set is used for the teacher and student branches of a pair, and the default number of shared draws is $C = 3$. Counterfactual passages are serialized with the same tool-response template, passage ordering, and truncation rule as the corresponding real retrieval response.

A.3 Verification Coverage

Table 6 separates candidate formation from environment verification. A training instance enters the auxiliary pipeline when a failed rollout has a successful sibling and a valid last search turn. Within the supervised query span, only teacher–student disagreements form candidate pairs; positions where the teacher agrees with the sampled token receive no paired verification or auxiliary distillation. Candidates with valid matched executions are then scored by executed paired information gain, and only pairs with positive gain receive a nonzero auxiliary weight. Rates in the first block use token-level totals over the 400-step training schedule. The terminal audit in the second block is a separate diagnostic subset, not an estimate of candidate prevalence; its sizes are given in the definitions.

Table 6: Candidate formation, paired verification, and diagnostic-audit coverage. Training-funnel rates use token-level totals over the 400-step schedule; the terminal audit is a separate source-stratified subset.

Stage	Definition	Rate
Training-time candidate funnel		
Query-token positions	All supervised query-span positions	100%
Teacher–student disagreements	$t_i \neq s_i$	9.22% of query tokens
Submitted to paired verification	Disagreements sent to matched evaluation	99.68% of disagreements
Valid paired evaluations	Submitted pairs passing both execution and scoring checks	99.90% of submitted
Positive-gain pairs	Valid pairs with $\Delta\mathrm{IG} > 0$	53.45% of valid pairs
Positive-gate query tokens	Positions with nonzero auxiliary weight	4.91% of query tokens
Diagnostic terminal audit		
Full diagnostic candidate pool	Candidate pairs before terminal-audit subsampling (4,984 pairs)	100% of pool
Terminal-ready audit pairs	Source-stratified, quota-targeted valid-terminal subset (307 pairs)	6.16% of pool
Question clusters	Questions represented in the audit (70 clusters)	–

A.4 Prompts and Tool Interface

Table 7 summarizes the prompt components used in our experiments. The base question prompt, privileged teacher instruction, and Hermes-style search schema enter the trajectory at different stages. The question itself is supplied by the dataset; it is not rewritten by IGSD.

Table 7: Prompts and search-tool interface used in our experiments.

Component	Prompt or template
Base system prompt	`You are a helpful and harmless assistant. Answer the given question. Reason between <thought> and </thought> after each new observation. If external information is needed, emit one JSON object inside <tool_call> tags. The search tool returns results inside <tool_response> tags. Break complex questions into specific sub-questions, keep queries unique, and finish with <answer>...</answer> when no further search is needed.`
Search tool schema	`Hermes/XML tool call: <tool_call>{''name'': ''search'', ''arguments'': {''query_list'': [''fully-formed semantic query'', ...]}}</tool_call>. The schema exposes one function, search; query_list is a non-empty list of strings.`
Hindsight teacher instruction	`You are the query-repair branch of a search agent. Propose exactly one next search action for the failed trajectory prefix below. A successful sibling's search queries and score are strategy hints only. Do not answer the question, discuss the hint, or emit thought. Output exactly one <tool_call> for the search tool. The prompt then appends: Failed trajectory prefix; Successful sibling score; Successful sibling search queries.`

B Signal Diagnostics

Table 8 evaluates every signal against the same labels: which branch of a terminal-discordant pair ultimately answers correctly. Of the 307 evaluated pairs, 53 are teacher-only correct, 19 student-only correct, 47 both correct, and 188 both wrong; the 72 discordant pairs, drawn from 70 questions, define the winner labels. Confidence intervals are 95% question-cluster percentile bootstrap intervals over 20,000 resamples. The table compares executed gain with model-internal token scores; only executed gain is used by IGSD. Executed gain has the highest AUC among the tested signals and, under its zero-threshold sign, selects the winning branch on 47 of 72 pairs (65.3%, CI $[54.2\%, 76.1\%]$).

Table 8: Predicting the paired terminal winner from model-internal token-probability scores and executed gain.

Signal family	Signal	Terminal-winner AUC (95% CI)
Internal confidence	Teacher top-1 log-probability	0.401 [0.252, 0.548]
Internal likelihood	Teacher–student sampled-token log-probability gap	0.552 [0.404, 0.706]
Internal preference	Teacher preference gap	0.428 [0.272, 0.579]
Environment execution	Executed paired information gain	**0.708 [0.568, 0.836]**

Figure 5 breaks down the terminal-discordant pairs by which branch is correct. Positive executed gain contains 34 teacher-only correct and 6 student-only correct pairs, compared with 19 and 13, respectively, in the non-positive group. The gate therefore retains a substantially higher share of teacher-winning outcomes.

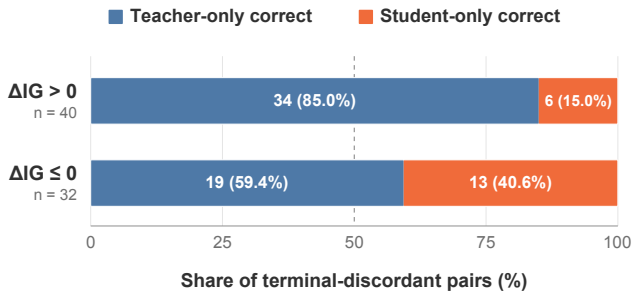

Figure 5: Terminal-outcome composition among terminal-discordant matched pairs. Bars partition pairs by which branch is correct; labels give counts and within-group percentages. Positive executed gain contains a larger share of teacher-winning outcomes.

Figure 6 tracks executed gain and three-way action agreement over training. The executed gain rises early, reaches a mid-training maximum, and then decreases while remaining positive; agreement among the teacher top-1 action, student top-1 action, and sampled student action increases over the same horizon. Together, the curves show that useful teacher guidance is strongest early in training while the three actions increasingly coincide later in training.



Figure 6: Mean signed executed gain and three-way action agreement over training.

C Repeated-Run Results

Table 9 reports three-seed means ± standard deviations for the four core configurations.

Table 9: Repeated-run results with three random seeds (Qwen2.5-3B-Instruct, mean ± standard deviation). The best mean in each column is in bold.

Method	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle	Avg.
Search-R1	0.438 ± 0.009	0.587 ± 0.006	0.429 ± 0.010	0.382 ± 0.010	0.420 ± 0.012	0.190 ± 0.011	0.340 ± 0.027	0.398 ± 0.006
SD-Search	0.455 ± 0.010	0.615 ± 0.006	0.427 ± 0.008	0.418 ± 0.015	0.419 ± 0.010	0.178 ± 0.008	0.400 ± 0.022	0.416 ± 0.008
IGSD w/o Gate	0.450 ± 0.008	**0.621** ± 0.009	0.450 ± 0.012	0.402 ± 0.015	0.400 ± 0.009	**0.191** ± 0.013	0.384 ± 0.026	0.414 ± 0.008
IGSD	**0.460** ± 0.008	0.618 ± 0.010	**0.459** ± 0.010	**0.421** ± 0.012	**0.439** ± 0.013	0.190 ± 0.009	**0.404** ± 0.022	**0.427** ± 0.006

D Efficiency and Search Usage

Table 10 separates the measured 400-step training cost from inference-time tool use; both are wall-clock measurements under the corresponding experimental settings. Table 11 gives the dataset-level breakdown of evaluation-time search calls for the same three methods (Search-R1, SD-Search, and IGSD).

Table 10: Training and inference efficiency over 400 optimization steps. Search-call columns report inference-time tool use on the same 51,713-example greedy evaluation set; wall-clock values are hours:minutes:seconds and calls/example is reported to three decimals.

Method	Wall-clock time (400 steps)	Relative time	Search calls	Calls/example
Search-R1	23:41:12	1.00×	84,393	1.632
SD-Search	51:35:04	2.18×	105,629	2.043
IGSD	29:42:53	1.25×	85,366	1.651

Table 11: Average evaluation-time search calls per example by dataset.

Dataset	Search-R1	IGSD	SD-Search
NQ	1.1626	1.1573	2.0003
TriviaQA	1.2766	1.2547	2.0069
PopQA	1.2267	1.2931	1.9990
HotpotQA	1.8913	1.8967	2.0330
2WikiMultiHopQA	2.2491	2.2609	2.1036
MuSiQue	2.3645	2.4075	2.2420
Bamboogle	1.9840	1.9920	2.0480
Overall	1.6319	1.6508	2.0426

E Extended Ablations

E.1 Hindsight Context Ablation

Figure 7 gives the complete per-dataset comparison of hindsight contexts. The variants change only the privileged information supplied to the self-teacher. The default "Queries + score" context is the strongest setting (macro-average 0.428). Removing the sibling outcome score ("Queries only") lowers the average to 0.417, adding retrieved documents ("Queries + score + docs") reaches 0.413, and removing hindsight entirely ("No hindsight") reaches 0.400. Supplying the ground-truth answer ("Queries + score + answer") gives the lowest average, 0.385. This is an intentionally leaky context: exposing the answer can encourage answer-conditioned behavior and reward hacking rather than useful retrieval, so we include it as a stress test of privileged information.

E.2 Complete Variant Comparisons

Table 12 collects the complete per-dataset results for the variants discussed in Section 4.4: its reference and ablation rows are those of Table 3, and the remaining blocks add the support and draw-count variants. The first block is the IGSD reference; each subsequent block varies one stage of the method. All rows use the 3B policy and the training recipe of Section 4.1.

The supervision-scope block verifies every valid search turn instead of only the last one, with the teacher hint, the gate, and the loss unchanged. The next block varies how a teacher proposal is verified. IGSD compares the teacher's top-1 token with the student's sampled token under a per-candidate matched completion; the w/o sampled-token-branch row removes the matched sampled-token reference, the w/o counterfactual-control row removes the shared counterfactual control, and the student-argmax row replaces the sampled student token with the student's arg-max token at the same position. The gate block varies how the executed gain weights supervision: *log-probability gate* substitutes the sampled-token log-probability gap $\log p_T(s_i) - \log p_S(s_i)$ for $\Delta\mathrm{IG}_i$, the binary-gate row replaces the rectified-tanh gate with a sign gate, *query-level gate* averages the gain over a query's verified pairs, and *w/o Gate* issues no verification branches. The support-and-divergence block widens the two-token target to the de-duplicated union of the teacher and student top-50 tokens and to a top-50 reverse-KL objective. The final block varies the number of counterfactual draws.

Table 12: Complete per-dataset comparisons of the supervision-scope, verification, gating, support, and counterfactual-draw variants with the 3B policy. Best value in each column is in bold.

Variant	NQ	Triv.	Pop.	Hot.	2Wiki	Mus.	Bam.	Avg.
Reference								
IGSD (default)	0.458	0.620	**0.461**	0.420	0.443	0.183	**0.408**	**0.428**
Supervision scope								
All-turn verification	0.455	0.615	0.451	0.417	0.444	0.201	0.368	0.421
Paired verification								
Student argmax	0.456	0.618	0.443	0.409	0.425	0.179	0.360	0.413
w/o sampled-token branch	0.437	0.599	0.412	0.397	0.407	0.156	0.368	0.397
w/o counterfactual control	0.438	0.597	0.408	0.406	0.420	0.180	0.384	0.405
Gate								
Log-probability gate	0.450	0.609	0.438	0.412	0.421	0.186	0.344	0.409
Binary gate	0.456	0.617	0.444	0.419	0.439	0.197	0.376	0.421
Query-level gate	**0.469**	0.617	0.430	0.421	0.434	0.175	0.384	0.419
w/o Gate	0.455	**0.623**	0.448	0.401	0.407	0.184	0.400	0.417
Support and divergence								
Top-50 JSD	0.455	0.611	0.441	**0.424**	**0.445**	0.188	0.368	0.419
Top-50 reverse-KL	0.457	0.613	0.433	0.414	0.428	**0.204**	0.392	0.420
Counterfactual draws								
$C = 1$ counterfactual draw	0.460	0.614	0.439	0.419	0.432	0.187	0.400	0.422
$C = 5$ counterfactual draws	0.449	0.609	0.422	0.416	0.416	0.183	0.384	0.411

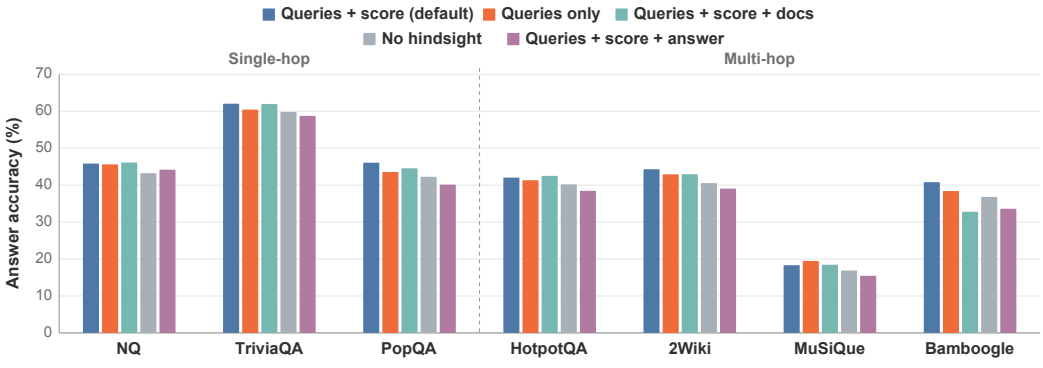

Figure 7: Complete hindsight-context comparison. The default teacher context is successful sibling queries plus outcome score; other bars add, remove, or replace this privileged information as described in the text.

F Hyperparameter Sensitivity

Figure 8 varies one hyperparameter at a time around the default configuration. The default reaches the highest macro-average in each sweep.

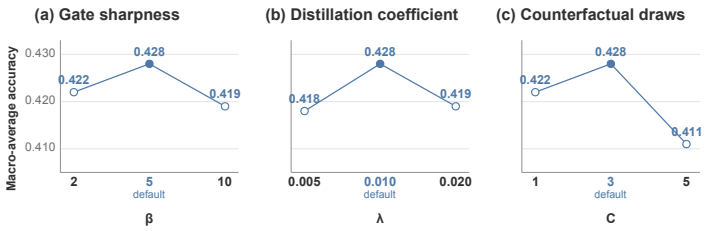

Figure 8: Hyperparameter sensitivity of IGSD with the 3B policy. Gate sharpness, distillation coefficient, and counterfactual draws are varied one at a time.

Table 13 summarizes the reproducibility-critical settings for the main seven-benchmark experiments. Following common reporting practice, we include the optimization, retrieval, and method-specific parameters that affect the training objective or agent behavior.

Table 13: Training, retrieval, and IGSD hyperparameters.

Setting	Main experiments
Training configuration	
Policy	Qwen2.5-3B-Instruct; Qwen2.5-7B-Instruct
Hardware	8× NVIDIA A100 80GB
Training data	NQ + HotpotQA
Training batch size	256
Rollouts per prompt	5
Actor learning rate	10^{-6}
Optimization steps	400
Maximum prompt / response length	4096 / 3000
Maximum assistant turns	6
Rollout temperature / top-p	1.0 / 1.0
Retrieval configuration	
Retriever	E5-base-v2
Corpus	Wikipedia (December 2018)
Passages per query	3
IGSD-specific configuration	
Teacher context	Successful sibling queries + outcome score
Supervised turn	Last valid search turn
Distillation objective	Candidate-pair JSD
Distillation span	Search-query tokens
Counterfactual draws C	3
Gate sharpness β	5
Distillation coefficient λ	0.01

G Additional Case Summaries

Table 14 gives three additional positive-gain paired trajectories and one negative-gain example with independently completed terminal answers. Each case retains the question and state, the sampled-token and teacher-token search branches, retrieved snippets, and the resulting ΔIG. Blue and orange mark the local student and teacher query choices, while gray text explains their retrieval consequences.

Table 14: Additional paired trajectories. The table includes three positive-gain cases and one negative-gain case; both branches are completed independently with the same frozen student policy and decoding procedure.

Positive ΔIG — advances to the missing organizational hop **RETAIN**

Question: What company is Frederic Ewen's employer part of? [Reference: City University of New York]
State: A previous search asks who Frederic Ewen and his employer are; the retrieved evidence identifies him as an English professor at **Brooklyn College** (employer resolved; parent organization unresolved).

Student-sampled action	**Hindsight-teacher proposal**
Search: what is the `employer of Frederic Ewen`	**Search:** what is `Brooklyn College`
Retrieved: Frederic Ewen (1899–1988) was an English professor at Brooklyn College from 1930 to 1952. During the McCarthy period Ewen was forced to resign his teaching position . . .	**Retrieved:** Brooklyn College is a university of the **City University of New York**, located in the Flatbush neighborhood of Brooklyn, New York City . . .
Continuation: ". . . his employer was Brooklyn College."	**Continuation:** "Brooklyn College is a part of the City University of New York. Therefore Frederic Ewen's employer is part of it."
Repeats the solved employer lookup and returns the intermediate entity.	*Queries the resolved entity and advances to the missing hop.*
Answer: Brooklyn College — incorrect	**Answer:** City University of New York — correct
$\Delta\text{IG} = +3.7049 > 0$	

Positive ΔIG — grounds the target country **RETAIN**

Question: Who is the president of the newly declared independent country in the Commission of Truth and Friendship with the country where Gado-gado is found? [Reference: Francisco Guterres]
State: Retrieved evidence identifies Gado-gado as Indonesian and the commission as an Indonesia–East Timor body (target country resolved; president unresolved).

Student-sampled action	**Hindsight-teacher proposal**
Search: who is the president of the newly declared independent country in the commission with `Indonesia`	**Search:** who is the president of `East Timor`
Retrieved: The Indonesia–Timor Leste Commission on Truth and Friendship was established jointly by the governments of Indonesia and East Timor . . .	**Retrieved: Francisco Guterres** is an East Timorese politician who has been President of East Timor since 2017 . . .
Continuation: ". . . the president of Indonesia is Susilo Bambang Yudhoyono."	**Continuation:** ". . . the president of East Timor is Francisco Guterres, and East Timor is the country in the commission with Indonesia."
The long relation leaves Indonesia as the dominant but incorrect anchor.	*Names the resolved country and removes the competing Indonesia anchor.*
Answer: Susilo Bambang Yudhoyono — incorrect	**Answer:** Francisco Guterres — correct
$\Delta\text{IG} = +3.3852 > 0$	

Positive ΔIG — carries forward the resolved director **RETAIN**

Question: Who is the spouse of the director of California (1947)? [Reference: Maureen O'Sullivan]
State: Retrieved evidence identifies **John Farrow** as the director (director resolved; spouse unresolved).

Student-sampled action	**Hindsight-teacher proposal**
Search: who was the spouse of the director of `California (1947)`	**Search:** who was the spouse of `John Farrow`
Retrieved: California is a 1947 American western film directed by John Farrow and featuring Ray Milland, Barbara Stanwyck, and . . .	**Retrieved:** Prudence Farrow married teacher Albert Morrill Bruns in December 1969. She is the daughter of film director John Farrow and actress Maureen O'Sullivan . . .
Continuation: ". . . the director was John Farrow. John Farrow was married to Marsha Hunt."	**Continuation:** ". . . the spouse of John Farrow was Maureen O'Sullivan."
Keeps the film as the anchor and mistakes an associated actress for the spouse.	*Carries the resolved director into the query and retrieves the spouse relation directly.*
Answer: Marsha Hunt — incorrect	**Answer:** Maureen O'Sullivan — correct
$\Delta\text{IG} = +2.6150 > 0$	

Negative ΔIG — anchors on the wrong newspaper format **REJECT**

Question: Newspapers come in various sizes, broadsheet being the largest, and tabloid being smaller, but what name is given to the smallest size? [Reference: compact]
State: The question distinguishes broadsheet from tabloid and asks for the smaller format (format relation unresolved).

Student-sampled action	**Hindsight-teacher proposal**
Search: **Newspapers come in various sizes, broadsheet being the largest, and tabloid being smaller, but what name is given to the smallest size**	**Search:** **what is the name of the smallest size of newspapers, where broadsheet is the largest and tabloid is smaller**
Retrieved: A tabloid is a newspaper with a compact page size smaller than broadsheet. Some small-format papers refer to themselves as **compact** newspapers . . .	**Retrieved:** A tabloid is a newspaper with a compact page size smaller than broadsheet . . . The query remains anchored on the already-mentioned **tabloid** format.
Preserves the contrast among newspaper formats and identifies the requested smaller format.	*Retrieves the tabloid page but treats the question's tabloid reference as the answer.*
Answer: compact — correct	**Answer:** tabloid — incorrect
$\Delta\text{IG} = -1.0606 < 0$	