SP-GRPO: SHARED-PREFIX GROUP RELATIVE POLICY OPTIMIZATION FOR MULTI-TURN AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning post-training, notably the critic-free Group Relative Policy Optimization (GRPO), has become a cornerstone for enhancing the reasoning capabilities of large language models, yet extending its group-comparison paradigm to long-horizon, multi-turn agent tasks remains challenging, as sparse terminal rewards fail to provide fine-grained, turn-level credit assignment. Existing methods typically attempt isolated fixes by either modifying credit assignment through turn-level advantages computed across independently rolled-out trajectories, or altering rollout structures via tree expansions. By treating rollout dynamics and credit assignment as decoupled components, these approaches suffer from a structural misalignment that introduces context-confounding bias. To address this, we propose Shared-Prefix Group Relative Policy Optimization (SP-GRPO), which aligns single-path environment continuation with turn-level counterfactual evaluation. SP-GRPO concentrates the sampling budget on branching multiple candidate actions from a strictly identical historical state, scores each candidate by information gain, and advances the environment by uniformly selecting a single continuation path. We prove that uniform continuation aligns with the original rollout, and that shared-prefix comparison removes the state-mismatch residual of cross-trajectory baselines while retaining the within-group contrast. Empirically, extensive experiments across seven single-hop and multi-hop QA benchmarks show that SP-GRPO achieves the best average performance at both model scales, outperforming trajectory-level GRPO, tree-based approaches, and existing step-level baselines.

1 INTRODUCTION

Reinforcement learning post-training has emerged as a cornerstone for enhancing reasoning capabilities and decision-making in large language models and autonomous agents (Ouyang et al., 2022; Bai et al., 2022). Among these methods, the critic-free Group Relative Policy Optimization (GRPO) (Shao et al., 2024; DeepSeek-AI, 2025) learns by comparing the outcome-level rewards of a group of complete trajectories sampled under the same prompt, constructing relative advantages purely at the trajectory level. As LLMs develop into multi-turn, long-horizon agents, the quality of intermediate steps becomes increasingly important, driving RL supervision to shift from the outcome level to the turn level.

Existing efforts toward turn-level supervision fall into three paradigms. The first introduces fine-grained feedback for intermediate steps via value critics or process reward models (PRMs) (Lightman et al., 2024; Wang et al., 2024), which policy-gradient RL (Schulman et al., 2016; 2017) converts into advantages by accumulating discounted future returns along the rolled-out trajectory. The second extends group comparison to the step level in a critic-free manner, retroactively grouping actions across trajectories that reach identical or partially shared states (Feng et al., 2025; He et al., 2026b). The third restructures the rollout itself, branching multiple actions at intermediate nodes so that candidate actions share a common prefix (Ji et al., 2025).

Although prior methods introduce turn-level supervision on either the advantage side or the rollout side, a granularity mismatch persists, as the advantage is estimated per turn while the evidence for it is still collected at the trajectory level. Chain-based methods sample only one action per state and match actions across diverged contexts. Tree-based methods spread the budget across branches

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.

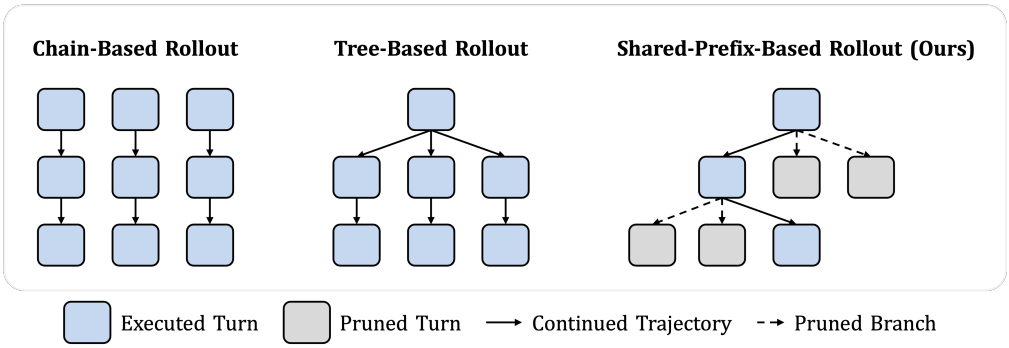

Figure 1: Comparison of rollout patterns. (left) Chain-based rollout with independent trajectories, (middle) tree-based rollout with full branching, and (right) our shared-prefix-based rollout. Blue boxes represent executed turns, while gray boxes denote pruned ones. Solid and dashed arrows indicate continued and pruned trajectories, respectively.

and states, leaving few candidates per state at rapidly growing cost and no immediate per-state evidence, so per-turn advantages can only be derived by accumulating discounted future returns whose variance compounds with the horizon. Faithful turn-level supervision therefore calls for a rollout that is itself organized around turns. This leaves open the central question of *how to obtain turn-level advantages that compare candidate actions under strictly identical contexts through dense per-state branching during rollout, while keeping the sampling budget practical*.

To address this, we propose Shared-Prefix Group Relative Policy Optimization (SP-GRPO), a turn-level rollout framework with same-context turn-level credit assignment in multi-turn agent RL. As illustrated in Figure 1, SP-GRPO maintains a single active trajectory, branches G candidate actions from the strictly identical historical prefix at each turn, evaluates them with an instantaneous process reward for a localized, same-context counterfactual comparison, and continues one candidate uniformly at random. Our main contributions are threefold.

- We propose a controlled turn-level rollout framework that enables fine-grained counterfactual credit assignment for multi-turn agents without requiring an explicit value critic.
- We estimate the turn-level advantage among the candidates via instantaneous process rewards under a shared prefix, eliminating the context confounding inherent in cross-trajectory comparisons and the compounding variance incurred by downstream errors.
- We select the continuation uniformly at random among the search-query candidates, and prove that this rule introduces no selection pressure on candidate actions while keeping rollout complexity linear in the interaction horizon.

2 Related Works

Step-Level Credit Assignment in Agentic RL. To densify supervision beyond terminal outcomes, recent work either trains auxiliary critics or process reward models for turn-level feedback (Zhou et al., 2025; Chae et al., 2025; Xi et al., 2025), or extracts step-level signals from trajectory structure, where GiGPO (Feng et al., 2025) groups actions at exactly matched states, HGPO (He et al., 2026b) re-weights groups by historical-context consistency, MT-GRPO (Zeng et al., 2025) derives turn-level advantages for multi-turn tool use, and SALT (Li et al., 2025) merges parallel trajectories into a graph. All of these methods estimate turn-level credit post hoc from evidence collected under diverged histories, so context confounding is attenuated but not removed, and the compounding variance of accumulated returns remains.

Tree-Structured Rollouts in Agentic RL. Tree-structured exploration, popularized by inference-time search over reasoning paths (Yao et al., 2023; Hao et al., 2023; Zhou et al., 2024), has been increasingly adopted to organize agentic rollouts, from MCTS-guided exploration (Putta et al., 2024) to search trees with group advantages (Ji et al., 2025), information-gain-guided expansion (Zhang et al., 2026), adaptive branching (Dong et al., 2025; He et al., 2026a), or process scorers that locate

worthwhile branches (Li et al., 2026). Branching in these methods is sparse or recursively expensive, and each branch is scored by the back-propagated return of its subtree or steered by node-level process scores, so few candidate actions are ever compared under one shared state. SP-GRPO, inspired by counterfactual multi-agent credit assignment (Foerster et al., 2018; Greensmith et al., 2004), instead performs single-action branching at every turn under a strictly shared prefix, scores candidates by instantaneous feedback, and continues a single query path with linear rollout complexity.

3 METHOD

3.1 OVERVIEW AND PROBLEM FORMULATION

We train a search agent with reinforcement learning, where the agent is an LLM that answers a question by interleaving generated queries with retrieved documents. Given a question $q \in \mathcal{S}$, the policy π_θ at turn $t \in \{1, \ldots, T\}$ produces an action $a_t \in \mathcal{A}$ from the retrieval prefix

$$s_t = (q, a_1, o_1, \ldots, a_{t-1}, o_{t-1}), \tag{1}$$

with $s_1 = q$. Each action is either a search query or a final answer. Issuing a query returns documents $o_t \sim P(\cdot \mid s_t, a_t)$ from the retriever and an instantaneous process reward $r(s_t, a_t)$. The completed trajectory $\tau = (q, a_1, o_1, \ldots, a_{T-1}, o_{T-1}, a_T)$, whose last action a_T is the final answer, receives a scalar outcome reward $R(\tau) \in \{0, 1\}$ given by exact match against the gold answer.

Standard GRPO samples G independent trajectories $\{\tau_i\}_{i=1}^G$ for the same question and uses only trajectory-level advantages $\hat{A}_i = (R(\tau_i) - \mu_R)/(\sigma_R + \epsilon)$. Intermediate actions receive no direct credit. SP-GRPO instead assigns turn-level advantages by branching candidate actions from a shared prefix and continuing the rollout along a single path, comparing candidate queries under the same state with dense turn-level credit, while answer candidates retain the outcome-level comparison.

3.2 SHARED-PREFIX ROLLOUT

Unlike chain-based rollouts that collect independent trajectories, SP-GRPO spends its sampling budget at every turn on a group of candidates sharing the same history rather than on G divergent chains, and continues only a single path, so the compute budget stays fixed and later actions always act on the same state as the compared candidates. Since the candidates of a turn may mix search queries and final answers, we partition them into a process group and an answer group, and compute a group-relative advantage within each. The process reward is an answer-anchored information gain in the same family as the probe of IGPO (Wang et al., 2026a), reflecting whether the action of each turn advances the final answer.

Branching Under an Identical Prefix. Given s_t, we sample G candidates $\{a_{t,j}\}_{j=1}^G \overset{\text{i.i.d.}}{\sim} \pi_{\theta_{\text{old}}}(\cdot \mid s_t)$. Since candidates act on the same prefix, differences in their process rewards are attributable to the candidate action and its immediate consequence, not to mismatched retrieval histories.

Process-Group and Answer-Group Advantages. The G candidates generally mix search queries issued to the search tool and final answers, whose rewards live on incomparable scales. We therefore partition them into a process group $\mathcal{P}_t$, whose candidates are search-tool calls scored by the process reward, and an answer group $\mathcal{Y}_t$, whose candidates are final answers scored by the outcome reward, and compute a group-relative advantage within each group.

Each candidate query is issued to the retriever once, solely for scoring, and its reward is the information gain computed from the retrieved documents, with the full formulation in Eq. 25 (Appendix B). During rollout, the process-group rewards are mapped through a temperature-controlled softmax, $x_{t,j} = \exp(r_{t,j}/\eta)/\sum_{m \in \mathcal{P}_t} \exp(r_{t,m}/\eta)$ for $j \in \mathcal{P}_t$ (η the softmax temperature, set to 1 in our experiments), which converts the gains into a probability distribution over candidates. The process advantages are then z-scored within the turn,

$$\hat{A}_{t,j} = \frac{x_{t,j} - \mu_t}{\sigma_t + \epsilon}, \qquad j \in \mathcal{P}_t, \tag{2}$$

where μ_t and σ_t are the mean and standard deviation computed over the process group, including the pruned query candidates. Each answer candidate, appended to the prefix at its turn, completes

a trajectory and is collected into the answer group along the rollout. After the rollout terminates, the exact-match rewards of the collected trajectories are z-scored within the question-level answer group, $\hat{A}_u^{\text{out}} = (R_u - \mu_{\mathcal{Y}})/(\sigma_{\mathcal{Y}} + \epsilon)$, recovering GRPO's outcome comparison among the completed trajectories of the same question.

Uniform Single-Path Continuation. Although a full group of candidates is sampled at every turn, only a single path is continued, so that later turns remain conditioned on one prefix and the rollout cost stays linear in the number of hops instead of expanding exponentially. Answer candidates never continue the rollout. Each of them completes a trajectory at its own turn and is collected into the answer group, and the continuation index is drawn uniformly from the process group,

$$j^* \sim \text{Uniform}(\mathcal{P}_t), \qquad a_t = a_{t,j^*}. \tag{3}$$

This content-agnostic rule introduces no selection pressure on candidate actions (Proposition 1). The executed query is issued to the search tool and the prefix is updated to $s_{t+1} = (s_t, a_t, o_{t,j^*})$ with the returned documents, and the rollout terminates once the collected answer trajectories meet the outcome quota (G by default) or the tool-call limit is reached. All candidate queries are issued to the retriever, but only the documents of the executed query enter the prefix. The pruned candidates still enter the group-relative baseline and the token-level KL regularization, but their advantages are zeroed and they receive no surrogate gradient. The procedure is summarized in Algorithm 1.

3.3 Training Objective

Executed queries enter the surrogate with their process advantages, collected answer candidates enter with their question-level outcome advantages, and the pruned query candidates contribute solely through the group-relative baseline and the token-level KL term.

Let i index a trained action, namely the executed query a_{t,j^*} of some turn or an answer candidate u collected along the rollout, with advantage $\hat{A}_i$ equal to $\hat{A}_{t,j^*}$ in the former case and $\hat{A}_u^{\text{out}}$ in the latter, and let s_i denote the prefix under which a_i was sampled. Writing $a_{i,k}$ for the k-th token of a_i and $a_{i,<k}$ for its first $k-1$ tokens, the token-level importance ratio is

$$\rho_{i,k}(\theta) = \frac{\pi_\theta(a_{i,k} \mid s_i, a_{i,<k})}{\pi_{\theta_{\text{old}}}(a_{i,k} \mid s_i, a_{i,<k})}. \tag{4}$$

The objective (to be maximized) is

$$\begin{aligned} \mathcal{J}_{\text{SP-GRPO}}(\theta) = \mathbb{E}\Bigg[\sum_i \sum_{k=1}^{|a_i|} \min\Big(\rho_{i,k}(\theta)\, \hat{A}_i,\ \text{clip}\big(\rho_{i,k}(\theta), 1-\varepsilon, 1+\varepsilon\big)\, \hat{A}_i \Big) \Bigg] \\ - \beta\, \mathbb{E}\Bigg[\frac{1}{G} \sum_{t=1}^{T} \sum_{j=1}^{G} \sum_{k=1}^{|a_{t,j}|} D_{\text{KL}}\big(\pi_\theta(\cdot \mid s_t, a_{t,j,<k}) \,\|\, \pi_{\text{ref}}(\cdot \mid s_t, a_{t,j,<k})\big) \Bigg], \end{aligned} \tag{5}$$

where the first sum ranges over the executed query of every turn and over every collected answer candidate, ε is the clipping threshold, and β weights the KL penalty, which averages the token-level divergence to the reference policy over all G candidates.

3.4 Theoretical Analysis

We establish three properties of SP-GRPO, with all proofs deferred to Appendix A.

Selection-Agnostic Continuation. Continuing one query candidate uniformly at random does not favor any candidate action.

Proposition 1 (Selection-agnostic continuation)**.** *Suppose* $\{a_{t,j}\}_{j=1}^{G} \overset{\text{i.i.d.}}{\sim} \pi_{\theta_{\text{old}}}(\cdot \mid s_t)$ *at every turn, every answer candidate is collected as a completed trajectory, and* $j^* \sim \text{Uniform}(\mathcal{P}_t)$ *whenever* $\mathcal{P}_t \neq \emptyset$*. Then every collected answer trajectory is an exact on-policy sample of* $\pi_{\theta_{\text{old}}}$ *given its prefix, and for every search query* a*,*

$$P_{\text{SP-GRPO}}(a_t = a \mid s_t) = \frac{\pi_{\theta_{\text{old}}}(a \mid s_t)}{\pi_{\mathcal{Q}}(s_t)}, \qquad \pi_{\mathcal{Q}}(s_t) := \Pr_{a' \sim \pi_{\theta_{\text{old}}}(\cdot \mid s_t)}\big(a' \textit{ is a search query}\big), \tag{6}$$

so the executed query chain is distributed as an on-policy chain of $\pi_{\theta_{\text{old}}}$ reweighted by state-dependent factors that do not depend on the executed actions.

These reweighting factors depend on the states alone and introduce no preference among actions within a state, so the collected data is trained with the standard clipped importance ratio, with the bounded staleness of minibatch updates controlled by the clipped ratio and the KL penalty as in standard PPO-style training.

Within-State versus Between-State Variance. Comparing queries issued from different retrieval histories confounds query quality with document-context quality. Let $v(s) := \mathbb{E}_{a\sim\pi}[r(s,a)]$ and $A(s,a) := r(s,a) - v(s)$. The law of total variance splits

$$\mathrm{Var}(r) = \underbrace{\mathbb{E}_s[\mathrm{Var}(r \mid s)]}_{\sigma_w^2 \text{ (within-state)}} + \underbrace{\mathrm{Var}_s[v(s)]}_{\sigma_b^2 \text{ (between-state)}} . \tag{7}$$

A cross-trajectory group baseline estimates a pooled mean $\bar{v} = \mathbb{E}[v(S)]$ rather than $v(s)$, leaving a state-mismatch residual $\Delta(s) := v(s) - \bar{v}$.

Theorem 1 (Shared-prefix advantage estimation)**.** *Let $\tilde{A} = r - b_G$ be a mean-centered group advantage with G samples. For a cross-trajectory baseline that averages rewards from independently rolled-out prefixes,*

$$\mathbb{E}[\tilde{A}_{\text{cross}} \mid s, a] = A(s,a) + \Delta(s) + O(G^{-1}), \qquad \mathbb{E}[\Delta(S)^2] = \sigma_b^2, \tag{8}$$

and the baseline inherits both variance terms, $\mathrm{Var}(b_G) = \mathcal{O}((\sigma_w^2 + \sigma_b^2)/G)$. Under a shared prefix s_t, all G queries are i.i.d. given s_t, so $\Delta(s_t) \equiv 0$ and

$$\mathbb{E}[\tilde{A} \mid s_t, a] = \frac{G-1}{G} A(s_t, a), \qquad \mathrm{Var}(b_G \mid s_t) = \mathcal{O}(\sigma_w^2/G). \tag{9}$$

Theorem 1 therefore establishes, for the linear mean-centered estimator $\tilde{A}$, that a shared-prefix baseline removes the state-mismatch residual together with the extra variance σ_b^2, while a cross-trajectory baseline contaminates the turn-level ranking signal. Since $\Delta(s)$ depends on the state alone, its contribution to the policy gradient vanishes by the score-function identity, so the defect of cross-trajectory comparison is not a first-order gradient bias but a contamination of the ranking signal and the extra variance. The estimator actually used in Eq. 2 is not this linear estimator. It applies a within-group softmax to the raw gains and z-scores the resulting probabilities, a nonlinear, group-size-dependent transform whose magnitude depends on the full candidate set and the temperature η, and it should be read as a heuristic within-state ranking signal rather than an unbiased estimate of $A(s_t, a)$. The transform nevertheless operates purely on rewards collected at the same s_t, so it does not reintroduce $\Delta(s)$, and because the softmax is monotone in the raw reward and the z-score is affine, the candidate ordering induced by the information gains is preserved. We thus separate the two claims, namely that shared-prefix grouping provably removes cross-state baseline contamination for a linear estimator, and that Eq. 2 is a useful heuristic within-state ranking signal built on the same grouping (Appendix A.2).

Group-Relative Contrast at a Shared Prefix. Because the normalization in Eq. 2 is mean-centered within each group, $\sum_{j\in\mathcal{P}_t} \hat{A}_{t,j} = 0$ at every turn (and likewise over the answer candidates collected for a question, up to singleton groups). Combined with uniform j^*, the on-path update is a contrast among same-prefix candidates.

Theorem 2 (Group-relative contrast)**.** *Consider a turn t whose G candidates are all search queries, with advantages $\{\hat{A}_{t,j}\}$ mean-centered by Eq. 2. If $j^* \sim \mathrm{Uniform}(\mathcal{P}_t) = \mathrm{Uniform}(\{1,\dots,G\})$, the expected on-path policy gradient at turn t satisfies*

$$\mathbb{E}\big[\nabla_\theta \mathcal{J}_{\text{SP-GRPO}}(\theta) \mid s_t\big] = \frac{1}{G}\sum_{j=1}^{G} \hat{A}_{t,j} \left(\nabla_\theta \log \pi_\theta(a_{t,j} \mid s_t) - \frac{1}{G}\sum_{m=1}^{G} \nabla_\theta \log \pi_\theta(a_{t,m} \mid s_t) \right). \tag{10}$$

Theorem 2 characterizes the update direction and is invariant to any constant shift of the advantages within a group. Both the linear kernel of Theorem 1 and the normalization in Eq. 2 are mean-centered within $\mathcal{P}_t$, so they share the same contrast form among same-prefix candidates, with the

nonlinearity only reweighting candidates within the group. The theorem is stated for turns whose candidates are all queries. At mixed turns the query candidates remain mean-centered within $\mathcal{P}_t$, while an answer candidate enters the surrogate directly with its outcome advantage $\hat{A}_u^{\text{out}}$, contrasted across the completed trajectories of the same question as in GRPO (Appendix A.3). Each query's credit is thus confined to the evidence observed at its turn rather than backed up from future hops, while an answer candidate carries the outcome credit of its completed trajectory.

4 EXPERIMENTS

Section 4.2 reports exact match (EM) against search RL baselines, Section 4.3 analyzes why a shared-prefix sampler and an instantaneous process reward are necessary, and Section 4.4 ablates remaining design choices, with measurement protocols in Appendix C.

4.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate on seven search-augmented QA benchmarks covering single-hop QA (NQ† (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023)) and multi-hop QA (HotpotQA (Yang et al., 2018), 2WikiMultihopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), Bamboogle (Press et al., 2023)).

Baselines. We compare SP-GRPO against representative baselines covering the main lines of progress in search RL, including the R1-Instruct backbone, trajectory-level chain methods (Search-R1 (Jin et al., 2025), ZeroSearch (Sun et al., 2025)), step-level chain methods (StepSearch (Wang et al., 2025), GiGPO (Feng et al., 2025), IGPO (Wang et al., 2026a)), and tree-based methods (Tree-GRPO (Ji et al., 2025), IGRPO (Zhang et al., 2026)).

Training Details. All methods use Qwen2.5-3B/7B-Instruct (Yang et al., 2024) as base models, the E5 retriever (Wang et al., 2022) following Search-R1 (Jin et al., 2025), at most 4 retrieval hops, and 200 training iterations. Chain-based baselines roll out $N = 5$ trajectories per question, while SP-GRPO branches $G = 5$ candidates per turn, matching the per-turn sampling budget. Intermediate search queries are scored by an answer-anchored information-gain probe of the same family as IGPO (Wang et al., 2026a), instantiated as the process reward $r(s_t, a)$ in Eq. 25. Further training details, including the differences from IGPO's formulation, are provided in Appendix B.

4.2 OVERALL PERFORMANCE

Table 1 presents the overall evaluation results, where SP-GRPO establishes the best average performance across both model scales. On Qwen2.5-3B-Instruct, SP-GRPO achieves an average accuracy of 47.4%, outperforming the strongest tree-based baseline IGRPO (45.4%) and chain-based baseline GiGPO (42.3%). On Qwen2.5-7B-Instruct, SP-GRPO reaches an average score of 50.0%, exceeding IGRPO (48.2%) and GiGPO (47.3%), though it trails IGRPO on Bamboogle at both scales. This margin is obtained under the same per-question sampling budget as the baselines ($G{=}N{=}5$), indicating that the gain stems from the quality of same-context turn-level credit rather than from additional compute, and it persists across model scales and out-of-distribution test sets.

4.3 IN-DEPTH ANALYSIS

4.3.1 RQ1: WHY MUST TURN-LEVEL CANDIDATES SHARE A PREFIX?

Prefix Collapse in Chain-Based Rollouts. Although existing research suggests that independently sampled trajectories often exhibit high overlap, we find a severe scarcity of shared prefixes in multi-turn search. As shown in Figure 2, the share of turns in groups of size ≥ 2 peaks near 30% in the first search turn and drops below 13% from the second turn onward, vanishing at the final turn, with the mean group size converging to 1. Post-hoc matching for shared prefixes is thus virtually infeasible after the initial retrieval turn, which prevents chain-based methods from obtaining turn-level relative advantages and directly motivates the shared-prefix rollout of SP-GRPO.

State Heterogeneity Biases Turn-Level Advantages. A tempting alternative is to ignore state heterogeneity altogether and align independent rollouts at the same turn index. We term this variant

Table 1: Main performance comparison across single-hop and multi-hop QA benchmarks. Best results are in **bold**, and second-best results are underlined. † indicates in-domain test sets, and * indicates out-of-distribution evaluation benchmarks. Avg. denotes the test-size-weighted average over the evaluation datasets.

Type	Method	Single-Hop QA			Multi-Hop QA				Avg.
		NQ†	TriviaQA*	PopQA*	HotpotQA†	2Wiki*	MuSiQue*	Bamb.*	
Qwen2.5-3B-Instruct									
RL Training	R1-Instruct	27.0	53.7	19.9	23.7	29.2	7.2	29.3	29.8
Chain-based	Search-R1	34.1	54.5	37.8	32.4	31.9	10.3	26.4	37.4
Chain-based	ZeroSearch	41.4	57.4	44.8	27.4	30.0	9.8	11.1	39.1
Chain-based	StepSearch	—	—	—	34.5	32.0	**17.4**	34.4	—
Chain-based	GiGPO	42.0	59.5	42.4	36.9	37.0	12.6	64.1	42.3
Chain-based	IGPO	40.2	58.1	43.2	35.7	34.9	12.3	64.5	41.4
Tree-based	Tree-GRPO	44.5	60.4	43.7	31.1	31.7	11.9	28.0	40.8
Tree-based	IGRPO	45.6	60.9	**47.5**	38.9	40.2	14.1	**65.7**	45.4
Shared-Prefix	**SP-GRPO (Ours)**	**47.8**	**62.9**	47.3	**41.8**	**42.9**	16.0	33.6	**47.4**
Qwen2.5-7B-Instruct									
RL Training	R1-Instruct	21.0	44.9	17.1	20.8	27.5	6.0	19.2	25.8
Chain-based	Search-R1	39.3	61.0	39.7	37.0	40.1	14.6	36.8	42.5
Chain-based	ZeroSearch	43.6	61.8	**51.5**	34.6	35.2	18.4	27.8	44.8
Chain-based	StepSearch	—	—	—	38.6	36.6	**22.6**	40.0	—
Chain-based	GiGPO	46.4	64.7	46.1	41.6	43.6	18.9	68.9	47.3
Chain-based	IGPO	44.0	60.1	45.4	42.9	42.5	16.8	39.9	45.7
Tree-based	Tree-GRPO	45.6	62.5	42.8	39.9	35.6	17.5	41.6	43.6
Tree-based	IGRPO	48.0	64.7	47.0	44.6	43.8	20.2	**71.8**	48.2
Shared-Prefix	**SP-GRPO (Ours)**	**51.0**	**66.2**	49.2	**45.2**	**45.0**	17.9	36.8	**50.0**

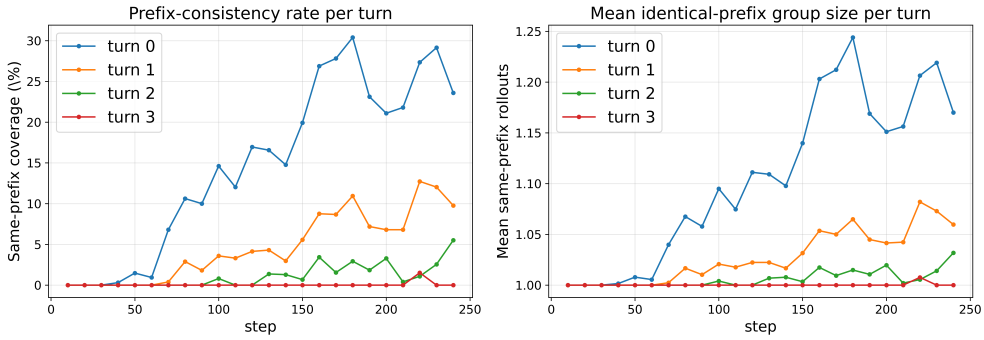

Figure 2: Empirical analysis of exact prefix sharing across $G = 5$ independently sampled chain rollouts. The left panel shows the fraction of turns in groups of at least two turns sharing an identical prefix, and the right panel shows the mean size of such groups.

Step-GRPO, which suffers a dramatic 11.3-point drop from 47.4% to 36.1% (Appendix C) because actions are compared under disparate historical states, conflating environmental favorability with true action quality. Even GiGPO, which groups actions only at exactly matched states, still degrades by 5.1 points, because exact matches are too rare to supply turn-level groups (Figure 2). These results thus separate two failure modes of post-hoc matching, pooling bias under approximate alignment and group starvation under exact matching, and underscore why SP-GRPO forms its groups at sampling time, where a full same-state group is guaranteed at every turn.

Shared-Prefix Rollouts Yield Unconfounded Comparative Signals. SP-GRPO removes this confounding by construction, as a shared prefix guarantees that the candidate queries of a turn are scored under identical states. Every candidate query is issued to the retriever for scoring, but only the continued query a_{t,j^*} enters the clipped surrogate with its process advantage. The pruned queries receive zero advantages and contribute only through the group-relative baseline and the KL term, while answer candidates collected along the rollout enter separately with their outcome advantages. Since the continued query's advantage is determined by the within-group scores in Eq. 2, the $G-1$ pruned candidates still govern the direction and magnitude of its update, yielding an unconfounded

comparative signal that chain-based methods lack. Table 4 (Appendix C) shows that once candidates share a single prefix, SP-GRPO's average improves monotonically with the scale (from 36.8% at $G{=}4$ to 41.0% at $G{=}16$) while the chain-based baselines plateau near 36%, with SP-GRPO overtaking both GiGPO and IGPO at every scale from $G = 4$ onward. At $G = 2$ the single reference candidate makes the within-group comparison too noisy for reliable relative advantages.

4.3.2 RQ2: Why instantaneous rewards over discounted returns?

Process rewards in long-chain reasoning often backpropagate terminal success via GAE. In contrast, our turn-level signal admits a strictly local formulation, since the instantaneous information gain reflects the local progress of the current retrieval turn, leaving no residual benefit for future returns to capture. We validate this choice by ablating advantage formulations that incorporate future returns, all of which trail the instantaneous signal, and by analyzing where such return-based variants fail.

Future Signals Corrupt Local Credit. Formally, future rewards $R_{t'>t}$ depend on the model's subsequent reasoning rather than the isolated quality of a_t, and incorporating them may introduce heavy conditional noise. As shown in Figure 3(a), hybrid variants that inject trajectory-level outcomes into Eq. 2 reduce average EM from 47.4% to 45.8% (mul) and 45.1% (add), where the multiplicative variant scales the instantaneous advantage by the trajectory-level outcome and the additive variant combines the two in a weighted sum (details in Appendix C.3). Since these variants retain the shared-prefix rollout and modify only the advantage, they isolate the effect of incorporating future returns, and a full GAE backup along the continued trajectory likewise yields 46.4% (per-dataset results in Table 5, Appendix C.3).

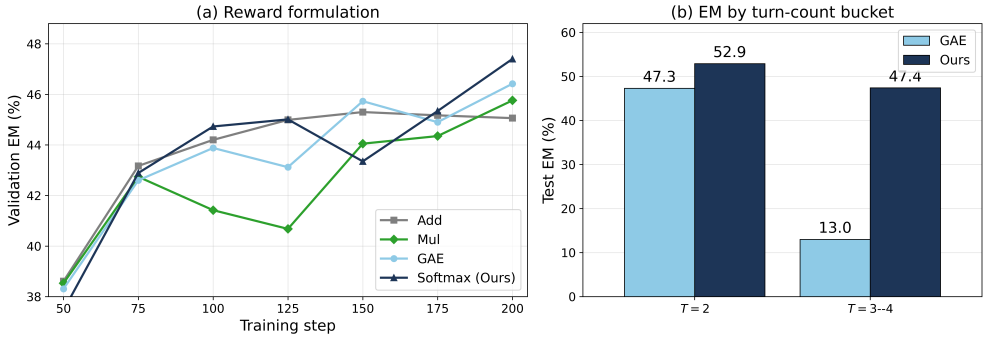

Figure 3: Ablation on advantage formulations and turn-depth analysis. (a) Test EM under different advantage formulations. (b) Test EM grouped by trajectory turn count T.

Discounted Return Degradation. To isolate where return-based advantages fail, Figure 3(b) breaks down performance by turn count T. Within the GAE baseline, EM holds at 47.3% on shallow trajectories ($T = 2$) but collapses to 13.0% once a third retrieval turn is required, whereas the instantaneous formulation exhibits no such depth-induced collapse because it confines credit to the current turn. The probability of reasoning errors or uninformative retrievals compounds with depth, and GAE propagates the final outcome backward through every turn, unfairly penalizing early high-quality actions whenever any subsequent turn fails.

4.3.3 RQ3: Does per-turn scoring inflate rollout overhead?

While scoring candidate queries at each turn requires additional forward passes, SP-GRPO maintains strictly linear time and space complexity in the trajectory depth by enforcing single-path continuation. At the 3B scale it consumes $\sim 1.9 \times 10^3$ tokens per question, the linear spine of a chain rollout plus the G candidate queries at the current tip, versus $\sim 6.0 \times 10^3$ for IGRPO, a $3.2\times$ reduction. Although each training step is more compute-intensive than parallel-chain baselines ($1{,}041$ s versus 673 s per step against GiGPO at the 3B scale), the high-precision advantage signals accelerate policy optimization, and SP-GRPO reaches the peak validation EM of GiGPO in ~ 143 training steps versus ~ 175 (19% fewer) while sustaining a 2.5-point EM margin.

4.3.4 RQ4: DOES SINGLE-PATH CONTINUATION REDUCE EXPLORATION DIVERSITY?

SP-GRPO continues only one candidate per turn, and a natural question is whether keeping multiple active paths improves search diversity and accuracy. Table 2 compares single-path continuation against keeping $K \in \{2, 4\}$ active paths, splitting the same total candidate budget across paths (detailed configurations in Appendix C.5). Average EM drops from 47.4% at $K = 1$ to 41.9% and 39.7% at $K = 2$ and $K = 4$. The candidate budget is held fixed, but each active path continues its own spine, so the executed queries and the retrieval cost grow linearly with K, while the within-group candidate pool shrinks and updates are spent on divergent histories outside the shared prefix, reintroducing the state-mismatch bias analyzed in Section 4.3.1. Appendix C.6 further confirms that single-path continuation preserves exploration, as sibling groups remain as query-diverse as chain-based rollouts.

Continued Paths K	Avg. EM (%)
K=1 (Ours)	**47.4**
K=2	41.9
K=4	39.7

Table 2: Ablation on the number of continued paths K.

4.4 ABLATION STUDY

We ablate key design choices on Qwen2.5-3B-Instruct, reporting average EM. Each variant modifies a single component while retaining the defaults of uniform continuation, information-gain rewards, and surrogate updates on the executed query and the collected answer candidates.

Continuation Action Selection Strategy. We compare uniform selection against Greedy Max-IG, which always advances the highest-IG candidate, and Softmax-over-IG, which samples proportionally to information gain. Softmax-over-IG introduces mild off-policy bias with 1.4% degradation, and Greedy Max-IG collapses with 30.2% degradation (Table 3). The training dynamics in Appendix C.4 attribute these gaps to exploration, as random selection preserves the diversity that sustains stable multi-turn optimization.

Selection	Avg. EM (%)
Random (Ours)	**47.4**
Softmax-over-IG	46.0
Greedy Max-IG	17.2

Table 3: Ablation on continuation selection strategies.

Process Reward Formulation. SP-GRPO employs Information Gain (IG), an answer-aware process reward measuring how effectively the current action alters the target answer distribution. Replacing it with an answer-agnostic token-entropy signal that requires no gold targets reduces performance to 41.0% (Table 7 in Appendix D). The signal reuses the entropy and confidence measurements that T^{2}PO (Wang et al., 2026b) employs to decide when to truncate reasoning and resample a turn, but deployed as a process reward it scores each action by the drop in the policy's own token-level uncertainty, which cannot distinguish a query that resolves the current uncertainty from one that merely expresses it.

Surrogate Updates on Pruned Actions. All G candidates establish the group-relative baseline and the KL penalty, while the clipped surrogate in Eq. 5 covers the executed query a_{t,j^*} and the collected answer candidates. Extending it to the pruned query candidates with their respective advantages reaches 46.2% (Table 7 in Appendix D), slightly below the default, confirming that restricting surrogate updates to the executed query and the answer candidates stabilizes optimization.

5 CONCLUSION

In this paper, we propose Shared-Prefix Group Relative Policy Optimization (SP-GRPO), a reinforcement learning framework that converts turn-level credit assignment in multi-turn search agents into a within-state comparison. At each turn the agent branches G candidate actions under one shared prefix, scores query candidates by information gain and answer candidates by outcomes, and continues a single on-policy path, so supervision is both dense and free of state mismatch while keeps linear complexity. Extensive experiments across seven search-QA benchmarks demonstrate that SP-GRPO consistently outperforms strong chain-based and tree-based baselines. We view the shared-prefix principle of comparing candidate actions under strictly identical contexts as a general building block for credit assignment in broader agentic RL.

AI USE STATEMENT

We disclose that AI tools were used in preparing this paper. Large language models were used to aid and polish the writing of this manuscript, with all AI-assisted text reviewed and revised by the authors. AI-assisted tools were used for retrieval and discovery of related work, with all cited references read and verified by the authors. The authors take full responsibility for the content of this paper.

REFERENCES

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Hyungjoo Chae, Sunghwan Kim, Junhee Cho, Seungone Kim, Seungjun Moon, Gyeom Hwangbo, Dongha Lim, et al. Web-Shepherd: Advancing PRMs for reinforcing web agents. *arXiv preprint arXiv:2505.15277*, 2025.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, et al. ARPO: Agentic reinforced policy optimization. *arXiv preprint arXiv:2507.19849*, 2025.

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

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004.

Shibo Hao, Yi Gu, Haodi Ma, Jiahui Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Bowei He, Yankai Chen, Xiaokun Zhang, and Xue Liu. Branching policy optimization: Sandbox-native language agent reinforcement learning. *arXiv preprint arXiv:2607.14171*, 2026a.

Shuo He, Lang Feng, Qi Wei, Xin Cheng, Lei Feng, and Bo An. Hierarchy-of-groups policy optimization for long-horizon agentic tasks. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Yuxiang Ji, Ziyu Ma, Yong Wang, Guanhua Chen, Xiangxiang Chu, and Liaoni Wu. Tree search for LLM agent reinforcement learning. *arXiv preprint arXiv:2509.21240*, 2025.

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

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Jiazheng Li, Yawei Wang, David Yan, Yijun Tian, Zhichao Xu, et al. SALT: Step-level advantage assignment for long-horizon agents via trajectory graph. *arXiv preprint arXiv:2510.20022*, 2025.

Xintong Li, Sha Li, Yuwei Zhang, Changlong Yu, Rongmei Lin, Hongye Jin, et al. PATR: Process reward informed tree rollout for effective multi-turn RL. *arXiv preprint arXiv:2607.15610*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.

Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent Q: Advanced reasoning and learning for autonomous AI agents. *arXiv preprint arXiv:2408.07199*, 2024.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Yan Zhang, Fei Huang, and Jingren Zhou. Zerosearch: Incentivize the search capability of LLMs without searching. *arXiv preprint arXiv:2505.04588*, 2025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. Information gain-based policy optimization: A simple and effective approach for multi-turn search agents. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Haixin Wang, Hejie Cui, Chenwei Zhang, Xin Liu, Shuowei Jin, Shijie Geng, Xinyang Zhang, Nasser Zalmout, Zhenyu Shi, and Yizhou Sun. T^2PO: Uncertainty-guided exploration control for stable multi-turn agentic reinforcement learning. *arXiv preprint arXiv:2605.02178*, 2026b.

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.

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

Ziliang Wang, Xuhui Zheng, Kang An, Cijun Ouyang, Jialu Cai, Yuhang Wang, and Yichao Wu. Stepsearch: Igniting LLMs search ability via step-wise proximal policy optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Zhiheng Xi, Chenyang Liao, Guanyu Li, Yajie Yang, Wenxiang Chen, Zhihao Zhang, Binghai Wang, et al. AgentPRM: Process reward models for LLM agents via step-wise promise and progress. *arXiv preprint arXiv:2511.08325*, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023.

Siliang Zeng, Quan Wei, William Brown, Oana Frunza, Yuriy Nevmyvaka, and Mingyi Hong. Reinforcing multi-turn reasoning in LLM agents via turn-level credit assignment. *arXiv preprint arXiv:2505.11821*, 2025.

Yijun Zhang, Fan Xu, Jiaxin Ding, Yule Xie, Shiqing Gao, Xin Ding, Haoxiang Zhang, Luoyi Fu, and Xinbing Wang. Information gain-based rollout policy optimization: An adaptive tree-structured rollout approach for multi-turn LLM agents. *arXiv preprint arXiv:2607.06223*, 2026.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. SWEET-RL: Training multi-turn LLM agents on collaborative reasoning tasks. In *International Conference on Machine Learning*, 2025.

A Theoretical Guarantees and Proofs

This appendix records the statements and proofs of the guarantees summarized in the Method section. Throughout, s_t denotes the retrieval prefix at turn t, $\tau = (q, a_1, o_1, \ldots, a_{T-1}, o_{T-1}, a_T)$ a query–document trajectory initiated at question q, $r(s, a)$ the instantaneous process reward (information gain on search queries), and $R(\tau)$ the terminal exact-match reward. For a fixed turn t, d_t denotes the occupancy distribution over prefixes, i.e., $d_t(s) = \Pr(s_t = s)$, and $v(s) = \mathbb{E}_{a\sim\pi}[\, r(s, a)\,]$ is the state value under the behavior policy, so that $A(s, a) = r(s, a) - v(s)$ is the state-conditioned advantage.

A.1 SELECTION-AGNOSTIC CONTINUATION

Proposition 1. *Suppose* $\{a_{t,j}\}_{j=1}^{G} \overset{\text{i.i.d.}}{\sim} \pi_{\theta_{\text{old}}}(\cdot \mid s_t)$ *at every turn, every answer candidate is collected as a completed trajectory, and* $j^* \sim \text{Uniform}(\mathcal{P}_t)$ *whenever* $\mathcal{P}_t \neq \emptyset$*. Then (i) every collected answer trajectory* $(q, a_1, o_1, \ldots, a_{t-1}, o_{t-1}, a_{t,j})$ *is an exact on-policy sample of* $\pi_{\theta_{\text{old}}}$ *given its prefix, and (ii) for every search query* a*,*

$$P_{\text{SP-GRPO}}(a_t = a \mid s_t) = \frac{\pi_{\theta_{\text{old}}}(a \mid s_t)}{\pi_{\mathcal{Q}}(s_t)}, \qquad \pi_{\mathcal{Q}}(s_t) := \Pr_{a' \sim \pi_{\theta_{\text{old}}}(\cdot \mid s_t)}\big(a' \text{ is a search query}\big), \quad (11)$$

so the executed query chain is distributed as an on-policy chain of $\pi_{\theta_{\text{old}}}$ *reweighted by state-dependent factors that do not depend on the executed actions.*

Proof. *Part (i).* Conditioned on the prefix s_t, candidate $a_{t,j}$ is an i.i.d. draw from $\pi_{\theta_{\text{old}}}(\cdot \mid s_t)$. Collecting it exactly when it is a final answer therefore yields a trajectory whose conditional density given s_t is $\pi_{\theta_{\text{old}}}(a_{t,j} \mid s_t)$, which is precisely the on-policy law of a trajectory ending with answer $a_{t,j}$ at s_t.

Part (ii). Fix s_t and let $m = |\mathcal{P}_t| \geq 1$. Conditioned on m, the m query candidates are i.i.d. draws from $\pi_{\theta_{\text{old}}}(\cdot \mid s_t)$ restricted to search queries. For a query a, let n_a denote the number of candidates equal to a. Since j^* is uniform over $\mathcal{P}_t$,

$$P_{\text{SP-GRPO}}(a_t = a \mid s_t, m) = \frac{1}{m}\,\mathbb{E}[n_a \mid s_t, m] = \frac{\pi_{\theta_{\text{old}}}(a \mid s_t)}{\pi_{\mathcal{Q}}(s_t)}. \quad (12)$$

The right-hand side does not depend on m, so averaging over m yields the claimed marginal. Every realized query candidate receives the same mass $1/m$ regardless of its content, so the rule introduces no preference among candidate actions.

Part (iii). By the chain rule and part (ii), the executed query chain $\tau = (a_1, o_1, \ldots, a_T, o_T)$ satisfies

$$P_{\text{SP-GRPO}}(\tau \mid q) = \prod_{t=1}^{T} \frac{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}{\pi_{\mathcal{Q}}(s_t)}\, P(o_t \mid s_t, a_t) = P_{\pi_{\theta_{\text{old}}}}(\tau \mid q) \prod_{t=1}^{T} \pi_{\mathcal{Q}}(s_t)^{-1}, \quad (13)$$

and each factor $\pi_{\mathcal{Q}}(s_t)^{-1}$ depends on the state s_t alone. ■

Remark (staleness within an iteration). Proposition 1 concerns the behavior policy $\pi_{\theta_{\text{old}}}$ that generated the rollouts. The state-dependent reweighting factors bias the state visitation but not the actions within a state. Within an iteration, the parameters are updated on the same batch, which introduces the usual bounded off-policy staleness of PPO-style methods; this staleness is controlled by the clipped importance ratio and the KL penalty in Eq. 5, and fresh rollouts are collected from the updated policy at the start of every iteration.

A.2 SHARED-PREFIX ADVANTAGE ESTIMATION

The law of total variance decomposes the reward variance into a within-state and a between-state component,

$$\text{Var}(r) = \underbrace{\mathbb{E}_s[\text{Var}(r \mid s)]}_{\sigma_w^2 \text{ (within-state)}} + \underbrace{\text{Var}_s[v(s)]}_{\sigma_b^2 \text{ (between-state)}}. \quad (14)$$

Theorem 1. *Let* $r = r(S, A)$*,* $v(s) = \mathbb{E}_{a \sim \pi}[r(s, a)]$*,* $A(s, a) = r(s, a) - v(s)$*,* $\sigma_w^2 = \mathbb{E}[\text{Var}(r \mid S)]$*,* $\sigma_b^2 = \text{Var}(v(S))$*, and* $\Delta(s) = v(s) - \bar{v}$ *with* $\bar{v} = \mathbb{E}[v(S)]$*. For the mean-centered group advantage* $\tilde{A} = r - b_G$*: (i) a cross-trajectory baseline whose samples are drawn from the occupancy* d_t *satisfies* $\mathbb{E}[\tilde{A}_{\text{cross}} \mid s, a] = A(s, a) + \Delta(s) + O(G^{-1})$ *and* $\mathbb{E}[\Delta(S)^2] = \sigma_b^2$*, with* $\text{Var}(b_G) = \mathcal{O}((\sigma_w^2 + \sigma_b^2)/G)$*; (ii) a shared-prefix baseline satisfies* $\mathbb{E}[\tilde{A} \mid s_t, a] = \frac{G-1}{G} A(s_t, a)$ *and* $\text{Var}(b_G \mid s_t) = \mathcal{O}(\sigma_w^2/G)$*.*

Proof. The law of total variance gives equation 14,

$$\mathrm{Var}(r) = \mathbb{E}_S\big[\mathrm{Var}(r \mid S)\big] + \mathrm{Var}_S\big(v(S)\big) = \sigma_w^2 + \sigma_b^2. \tag{15}$$

In particular $\mathbb{E}[\Delta(S)^2] = \mathrm{Var}(v(S)) = \sigma_b^2$.

Cross-Trajectory Baseline. Let $\{(s_m, a_m)\}_{m=1}^G$ be i.i.d. draws of a prefix from the occupancy d_t and an action $a_m \sim \pi(\cdot \mid s_m)$. The pooled baseline is $b_G = \frac{1}{G}\sum_{m=1}^G r(s_m, a_m)$. For a focal pair $(s, a) = (s_1, a_1)$,

$$\mathbb{E}[b_G \mid s, a] = \frac{1}{G}\, r(s, a) + \frac{G-1}{G}\, \bar{v}, \tag{16}$$

hence

$$\mathbb{E}[\tilde{A}_{\mathrm{cross}} \mid s, a] = \mathbb{E}[r(s, a) - b_G \mid s, a] = \frac{G-1}{G}\big(A(s, a) + \Delta(s)\big) = A(s, a) + \Delta(s) + O(G^{-1}). \tag{17}$$

The term $\Delta(s)$ is a function of the prefix alone, and it credits the average information gain of a retrieval history, not the candidate query issued from that history. Since b_G averages i.i.d. rewards with variance $\sigma_w^2 + \sigma_b^2$,

$$\mathrm{Var}(b_G) = \frac{1}{G}(\sigma_w^2 + \sigma_b^2). \tag{18}$$

Shared-prefix baseline. Fix s_t and sample $a_{t,1}, \ldots, a_{t,G} \overset{\text{i.i.d.}}{\sim} \pi(\cdot \mid s_t)$. Then $b_G = \frac{1}{G}\sum_{j=1}^G r(s_t, a_{t,j})$ and $\mathbb{E}[b_G \mid s_t] = v(s_t)$, so $\Delta(s_t) = 0$ does not appear. For a focal action a,

$$\mathbb{E}[\tilde{A} \mid s_t, a] = \mathbb{E}\Big[r(s_t, a) - \frac{1}{G} r(s_t, a) - \frac{1}{G}\sum_{j=2}^{G} r(s_t, a_{t,j}) \;\Big|\; s_t, a\Big] = \frac{G-1}{G} A(s_t, a). \tag{19}$$

Conditionally on s_t, $\mathrm{Var}(r(s_t, A) \mid s_t) \le \sup_s \mathrm{Var}(r \mid s)$ and

$$\mathrm{Var}(b_G \mid s_t) = \frac{1}{G}\,\mathrm{Var}(r \mid s_t), \tag{20}$$

which in expectation over $s_t \sim d_t$ is $\mathcal{O}(\sigma_w^2/G)$. The softmax–z-score estimator of Eq. 2 is computed from the same-prefix group $\{r(s_t, a_{t,j})\}$ and therefore cannot reintroduce $\Delta(s_t)$. ■

Remark (policy-gradient bias). The residual $\Delta(s)$ is a function of the state, not of the action. For the score-function identity $\mathbb{E}_{a\sim\pi}[\nabla_\theta \log \pi_\theta(a \mid s)] = 0$, the contribution $\mathbb{E}[\nabla_\theta \log \pi_\theta(a \mid s)\, \Delta(s)]$ vanishes. The defect of cross-trajectory comparison is therefore not a first-order bias of the policy gradient, but a contamination of the *turn-level* advantage used as a ranking signal among search queries, together with the extra between-state variance σ_b^2 in the baseline. Shared-prefix branching removes both.

Remark (from linear advantages to the softmax–z-score estimator). Theorem 1 analyzes the linear, mean-centered estimator $\tilde{A} = r - b_G$. The normalization in Eq. 2 replaces this estimator with a group-wise monotone transform of the same within-state rewards, namely a temperature softmax followed by an affine within-group z-score, centered so that $\sum_{j\in\mathcal{P}_t} \hat{A}_{t,j} = 0$. Because the transform is computed exclusively within a single prefix s_t, it inherits the two properties established above, namely $\Delta(s_t) \equiv 0$ and the removal of the between-state variance term. The nonlinearity rescales advantages within the group and introduces bounded higher-order terms, but it never mixes rewards across states, so the within-state contrast interpretation of the linear analysis carries over to the estimator used in practice.

A.3 Group-Relative Contrast

Theorem 2. *Consider a turn t whose G candidates are all search queries ($\mathcal{Y}_t = \emptyset$), with advantages $\{\hat{A}_{t,j}\}$ mean-centered by Eq. 2 so that $\sum_{j=1}^G \hat{A}_{t,j} = 0$. If $j^* \sim \mathrm{Uniform}(\mathcal{P}_t) =$*

$\mathrm{Uniform}(\{1,\dots,G\})$, *the expectation, over* j^* *and the candidate draw, of the gradient of the surrogate component of Eq. 5, evaluated at* $\theta = \theta_{\mathrm{old}}$ *and conditioned on* s_t*, satisfies*

$$\mathbb{E}\big[\nabla_\theta \mathcal{J}_{\text{SP-GRPO}}(\theta) \mid s_t\big] = \frac{1}{G}\sum_{j=1}^{G} \hat{A}_{t,j}\left(\nabla_\theta \log \pi_\theta(a_{t,j} \mid s_t) - \frac{1}{G}\sum_{m=1}^{G} \nabla_\theta \log \pi_\theta(a_{t,m} \mid s_t)\right). \tag{21}$$

The KL regularizer in Eq. 5 does not depend on the advantages and is omitted.

Proof. The clipped surrogate in Eq. 5 is applied only to the executed action $a_t = a_{t,j^*}$. At $\theta = \theta_{\mathrm{old}}$ one has $\rho_{t,k} = 1$, so the clip is inactive and the token-level scores telescope to $\nabla_\theta \log \pi_\theta(a_{t,j^*} \mid s_t)$. Hence the on-path gradient at turn t is $\hat{A}_{t,j^*}\, \nabla_\theta \log \pi_\theta(a_{t,j^*} \mid s_t)$. Conditioning on s_t and the realized candidate set $\{a_{t,j}, \hat{A}_{t,j}\}_{j=1}^{G}$, taking the uniform draw of j^* gives

$$\mathbb{E}_{j^*}\big[\hat{A}_{t,j^*}\, \nabla_\theta \log \pi_\theta(a_{t,j^*} \mid s_t) \mid s_t, \{a_{t,j}, \hat{A}_{t,j}\}\big] = \frac{1}{G}\sum_{j=1}^{G} \hat{A}_{t,j}\, \nabla_\theta \log \pi_\theta(a_{t,j} \mid s_t). \tag{22}$$

Averaging over the candidate draw given s_t preserves this form. Since the turn's candidates are all queries, mean-centering over the G candidates holds by the z-scoring in Eq. 2 applied to the softmax probabilities of the process group. For any vector c that does not depend on the candidate index,

$$\sum_{j=1}^{G} \hat{A}_{t,j}\big(\nabla_\theta \log \pi_\theta(a_{t,j} \mid s_t) - c\big) = \sum_{j=1}^{G} \hat{A}_{t,j}\, \nabla_\theta \log \pi_\theta(a_{t,j} \mid s_t), \tag{23}$$

because $\sum_{j=1}^{G} \hat{A}_{t,j} = 0$. Choosing $c = \frac{1}{G}\sum_{m=1}^{G} \nabla_\theta \log \pi_\theta(a_{t,m} \mid s_t)$ yields the contrastive form. ■

Remark (turns mixing queries and answers). The contrast form is a property of the mean-centered query block and extends to mixed turns. Since $\sum_{j\in\mathcal{P}_t} \hat{A}_{t,j} = 0$ at every turn, the query part of Eq. 22 satisfies

$$\sum_{j\in\mathcal{P}_t} \hat{A}_{t,j}\, \nabla_\theta \log \pi_\theta(a_{t,j} \mid s_t) = \sum_{j\in\mathcal{P}_t} \hat{A}_{t,j}\left(\nabla_\theta \log \pi_\theta(a_{t,j} \mid s_t) - \frac{1}{|\mathcal{P}_t|}\sum_{m\in\mathcal{P}_t} \nabla_\theta \log \pi_\theta(a_{t,m} \mid s_t)\right), \tag{24}$$

so an executed query is contrasted against the mean gradient of the candidates within $\mathcal{P}_t$. An answer candidate instead completes its trajectory at its own turn and enters the surrogate directly with its question-level outcome advantage $\hat{A}_u^{\mathrm{out}}$, which is mean-centered over the trajectories collected for the same question after the rollout terminates; answer candidates are therefore contrasted across completed trajectories of the same question, as in GRPO, rather than within a single turn, and empty or singleton groups are skipped (a singleton group is assigned a zero advantage).

B Experiment Details

Hyperparameters. The maximum prompt length is 4096 tokens, and the maximum response length is 512 tokens. The learning rate is 1×10^{-6} for the actor, with 10% linear warmup and a constant schedule thereafter. We adopt a rule-based outcome reward, assigning $R = 1$ for an exact match and $R = 0$ otherwise; invalid actions are penalized with -0.01. Intermediate search queries are scored by information gain (Wang et al., 2026a), instantiated as Eq. 25. We set the training batch size to 256 questions and use a group size of $G = 5$ candidates per turn. The rollout temperature is 1.0. The mini-batch size is 256, and the KL coefficient is $\beta = 0.001$. Process advantages first pass the candidate gains through the process-only softmax with temperature $\eta = 1$ and then apply the within-turn z-score of Eq. 2; answer advantages are z-scored within the question-level answer group after the rollout terminates and enter the surrogate on the answer tokens, and a group containing a single candidate is assigned a zero advantage. Pruned query candidates participate in the group mean and standard deviation but receive zero advantage in the clipped surrogate. The continuation index j^* is drawn uniformly from the process group $\mathcal{P}_t$.

Information-Gain Probe. The probe $p_{\text{ans}}(c) \in [0, 1]$ scores a context c by the mean token probability of the gold answer appended behind a fixed reasoning template under teacher forcing (Wang et al., 2026a). The reward of candidate j at turn t is its information gain, computed from the documents retrieved for its query,

$$r_{t,j} := r(s_t, a_{t,j}) = p_{\text{ans}}(s_t \oplus a_{t,j} \oplus o_{t,j}) - p_{\text{ans}}(s_t), \tag{25}$$

where $o_{t,j} \sim P(\cdot \mid s_t, a_{t,j})$ and $\oplus$ denotes concatenation; the parent term is omitted at the first turn. Each candidate query is issued to the retriever exactly once, solely for scoring, and the documents retrieved for pruned query candidates are discarded after scoring rather than entering the executed prefix.

Computing Details. For search-augmented QA, both Qwen2.5-3B and Qwen2.5-7B use $4\times$H20 GPUs, each trained for 200 iterations. All methods are implemented with verl and use vLLM as the rollout engine with tensor parallelism of 1 and 60% GPU memory utilization.

C Analysis Protocols and Results

This appendix records the setups and full per-dataset results for the analyses and ablations in the main text. Unless noted, results use Qwen2.5-3B-Instruct trained for 200 iterations with the hyperparameters above, and the evaluation protocol follows Section 4.2.

C.1 State Heterogeneity

The Step-GRPO baseline in Section 4.3.1 ablates the shared-prefix premise while retaining a per-step process signal. It uses a standard multi-turn chain rollout with $G{=}5$ independent chains per prompt, where each turn's step reward is the IGPO-style information gain along its own chain. Advantages apply GRPO's group normalization separately within two groups per prompt, namely terminal steps and non-terminal steps, so the baseline for each query is the average of the queries at the same hop of the other chains, rather than the sibling candidates sharing its exact prefix.

Aligning turns by index alone degrades every benchmark relative to SP-GRPO (Table 1). Step-GRPO reaches 37.0, 52.5, and 36.5 EM on NQ, TriviaQA, and PopQA, and 32.6, 27.3, 12.0, and 27.0 on HotpotQA, 2Wiki, MuSiQue, and Bamboogle, averaging 36.1 overall, which is 6.2 points below even GiGPO (42.3), whose exact-match grouping avoids cross-state comparison at the price of scarce groups. The degradation is most severe on multi-hop benchmarks (HotpotQA $41.8 \rightarrow 32.6$, 2Wiki $42.9 \rightarrow 27.3$, MuSiQue $16.0 \rightarrow 12.0$), where independently rolled-out chains diverge immediately after the first retrieval turn (Figure 2), so a shared hop index indexes radically different retrieval histories. This confirms that the pooling residual $\Delta(s)$ analyzed in Theorem 1 is not an abstract artifact. Injecting it into the turn-level ranking signal reverses the gains of step-level supervision.

C.2 Scaling Behavior

To examine how the methods perform at different sampling scales, we vary the per-prompt rollout budget for the chain-based baselines GiGPO and IGPO and for SP-GRPO. For the chain-based baselines the scale denotes the number of independently sampled trajectories per prompt, whereas for SP-GRPO it is the number of candidates branched per turn under single-path continuation ($K{=}1$). The budget ranges over $\{2, 4, 5, 8, 16\}$, where 5 is the default configuration of the main results (Table 1); all results are evaluated at training step 50.

According to the results in Table 4, at $G{=}2$ SP-GRPO (25.7) trails both chain-based baselines, since with a single competing candidate the within-group baseline is too noisy to yield a reliable relative advantage. From $G{=}4$ onward, SP-GRPO's average improves monotonically with the scale (36.8 at $G{=}4$, 38.8 at $G{=}5$, 40.3 at $G{=}8$, 41.0 at $G{=}16$), while GiGPO plateaus around 36 and IGPO, despite improving with the scale, stays at least 4 points below SP-GRPO at every setting. Larger candidate groups sharpen the same-state contrast that supplies the turn-level advantage, whereas additional independent chains enlarge the trajectory group without enlarging any per-state group, so the chain-based baselines cannot convert a larger budget into finer credit assignment.

Table 4: Per-dataset EM (%) at training step 50 for GiGPO, IGPO, and SP-GRPO at different sampling scales G. Within each scale, best results are in **bold**, and second-best results are underlined.

G	Method	Single-Hop QA			Multi-Hop QA				Avg.
		NQ†	TriviaQA*	PopQA*	HotpotQA†	2Wiki*	MuSiQue*	Bamb.*	
2	GiGPO	**31.5**	**50.6**	**32.3**	**27.1**	**25.7**	**7.6**	**22.9**	**32.7**
	IGPO	27.7	45.3	28.5	23.2	21.3	6.5	19.3	28.6
	SP-GRPO (Ours)	24.3	43.7	27.1	21.1	15.0	5.6	15.9	25.7
4	GiGPO	**34.5**	**53.9**	**36.3**	29.3	29.3	**9.9**	21.0	36.1
	IGPO	31.1	50.1	32.9	28.0	23.8	8.4	20.9	32.4
	SP-GRPO (Ours)	**34.5**	53.8	35.4	**32.0**	**31.8**	9.6	**21.6**	**36.8**
5	GiGPO	34.0	54.4	35.7	29.8	28.2	8.4	20.8	35.7
	IGPO	32.7	52.5	31.6	28.3	24.7	9.0	19.6	33.0
	SP-GRPO (Ours)	**37.3**	**56.9**	**38.8**	**33.8**	**31.3**	**10.6**	**25.7**	**38.8**
8	GiGPO	35.3	53.7	35.4	30.9	30.8	8.6	14.5	36.3
	IGPO	35.2	53.3	35.5	30.1	28.2	8.5	22.5	35.5
	SP-GRPO (Ours)	**39.1**	**58.7**	**40.5**	**34.7**	**33.1**	**11.1**	**24.6**	**40.3**
16	GiGPO	36.1	55.2	37.7	33.9	34.5	10.9	23.4	38.8
	IGPO	36.3	54.8	35.3	31.7	30.6	10.5	22.2	36.8
	SP-GRPO (Ours)	**41.3**	**57.6**	**41.0**	**35.4**	**34.7**	**12.9**	**24.2**	**41.0**

C.3 Future Returns

We report three ways of introducing future returns into the turn-level advantage. All variants keep the shared-prefix rollout and the process-group normalization unchanged and modify only how the per-turn advantage couples to the trajectory-level outcome signal. Let

$$s_i \;=\; \frac{\sigma_i - \mu_t}{\sigma(\mathcal{P}_t) + \epsilon}, \qquad o_u \;=\; \frac{R_u^{\max} - \bar{R}_u}{\sigma(\mathcal{R}_u) + \epsilon}, \tag{26}$$

where σ_i is candidate i's softmax probability within the round-t process group $\mathcal{P}_t$ (its group mean and std denoted $\mu_t, \sigma(\mathcal{P}_t)$), and o_u is the outcome advantage of trajectory u built from the best reward $R_u^{\max}$ and the mean/std of the outcome group $\mathcal{R}_u$ of the same prompt. The default (ours) uses the pure turn-level signal $\hat{A}_i = s_i$. *Additive Hybrid* mixes in the outcome signal with equal weight,

$$\hat{A}_i \;=\; s_i + w\, o_u, \qquad w = 1.0; \tag{27}$$

Multiplicative Gated uses it as a gate,

$$\hat{A}_i \;=\; s_i \cdot o_u, \tag{28}$$

so that the largest positive credit is reserved for turns that are both process-favored ($s_i > 0$) and belong to a winning trajectory ($o_u > 0$).

The GAE variant leaves the sampler, winner selection, and the group normalization identical, and only replaces the *score* fed into the process group, so that each candidate receives a GAE-style discounted accumulation of future gain residuals in place of the instantaneous process score. With no critic, the TD residual uses the empirical state value $V_t = \frac{1}{G}\sum_j \Phi_j$, the mean instantaneous gain of the G i.i.d. candidates forked from the same state:

$$\delta_i = \Phi_i - V_t, \qquad A_i = \delta_i + \gamma\, A_{i^+}, \tag{29}$$

where i^+ is the next spine node (the non-pruned, non-EOS child of i), accumulated bottom-up along the executed chain with discount factor $\gamma = 1$, i.e. the full Monte-Carlo return of gain residuals. Pruned and last-round nodes keep $A_i = \delta_i$; outcome rows and the surrogate scope are untouched. This isolates the *credit-assignment horizon* as the primary changed factor.

As shown in Table 5, all three variants that incorporate future returns trail the process-only default on average (46.4 GAE, 45.8 multiplicative, 45.3 additive, versus 47.4), and the losses concentrate on multi-hop benchmarks (MuSiQue $16.0 \to 14.8/14.3/13.9$, HotpotQA $41.8 \to 40.6/39.8/38.3$), precisely where the downstream randomness coupled in by the future signal compounds over depth. The per-dataset exceptions are TriviaQA under the multiplicative gate (63.0) and PopQA under GAE

Table 5: Hybrid query–outcome advantages on Qwen2.5-3B-Instruct (not used in the main method). Best results are in **bold**, and second-best results are underlined.

Method	Single-Hop QA			Multi-Hop QA				Avg.
	NQ†	TriviaQA*	PopQA*	HotpotQA†	2Wiki*	MuSiQue*	Bamb.*	
Process-only softmax (ours)	**47.8**	62.9	47.3	**41.8**	**42.9**	**16.0**	**33.6**	**47.4**
Additive Hybrid	45.7	61.5	45.5	38.3	39.7	13.9	29.6	45.1
Multiplicative Gated	46.3	**63.0**	46.0	39.8	39.6	14.3	31.9	45.8
GAE	47.1	61.9	**47.5**	40.6	40.8	14.8	31.8	46.4

Table 6: Winner-selection ablation, per-dataset EM (%) (K=1, maximum response length 512, G=5); the selection variants are evaluated at step 200, while Random (Ours) reports the default run of Table 1. Best results are in **bold**, and second-best results are underlined.

Selection	Single-Hop QA			Multi-Hop QA				Avg.
	NQ†	TriviaQA*	PopQA*	HotpotQA†	2Wiki*	MuSiQue*	Bamb.*	
Softmax	47.5	62.4	46.6	40.2	39.5	15.6	**34.4**	46.0
Greedy	22.0	32.5	11.0	13.9	14.0	2.5	12.8	17.2
Random (Ours)	**47.8**	**62.9**	**47.3**	**41.8**	**42.9**	**16.0**	33.6	**47.4**

(47.5), suggesting that on single-hop tasks, where one retrieval usually suffices, weak coupling to future success acts as a mild variance filter rather than a source of confounding. These results corroborate the analysis in Section 4.3.2. The instantaneous information gain already captures the local progress of each retrieval turn, and mixing in future returns contaminates it rather than complementing it.

C.4 Continuation Selection

All three variants share the sampler, rewards, storage, and the subsequent advantage computation of SP-GRPO; only the winner-sampling rule differs.

All three rules select the spine continuation among the query candidates forked at the current tip, using the round's expand probability

$$\pi_j \;=\; \frac{\exp(g_j/\eta)}{\sum_{k\in\mathcal{P}_t}\exp(g_k/\eta)}, \tag{30}$$

where g_j is candidate j's instantaneous information gain, $\mathcal{P}_t$ is the process group of non-answer candidates at turn t (so the distribution is renormalized over the non-answer subset), and η is the softmax temperature; we use η=1. The rules differ only in how π is used:

- **Softmax** samples $j^* \sim \pi$, so the winner is drawn with probability proportional to its gain, favoring informative queries while retaining exploration;
- **Random** samples $j^* \sim \mathrm{Uniform}(\mathcal{P}_t)$ and ignores π entirely, so the continuation is content-agnostic, and no selection pressure is fed back into the trajectory distribution;
- **Greedy** sets $j^* = \arg\max_j \pi_j$, so the candidate with the highest gain is always continued (at η=1, $\arg\max_j \pi_j = \arg\max_j g_j$).

Everything else (group baseline, KL penalty, surrogate scope) is identical across the three variants.

Table 6 reports the full per-dataset comparison. Uniform selection attains the best average (47.4), Softmax-over-IG introduces mild off-policy bias and drops to 46.0, whereas Greedy Max-IG collapses catastrophically to 17.2; the per-dataset gaps between uniform and softmax selection remain small, while greedy selection fails on every benchmark, consistent in ranking with the summary in Table 3. Figure 4 further reveals that these gaps stem from exploration dynamics. Greedy Max-IG collapses to an entropy of $\sim$0.06 with the gradient norm exploding to $\sim$4.6, where the subsequent entropy rebound reflects training instability rather than recovered exploration. Softmax-over-IG decays stably to $\sim$0.08, and uniform selection retains the highest entropy ($\sim$0.13), demonstrating

that random selection preserves the on-policy exploration diversity that sustains stable multi-turn optimization.

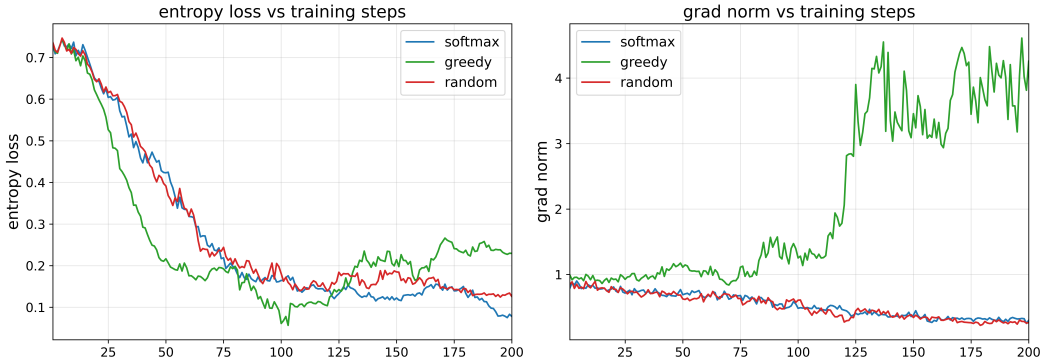

Figure 4: Training dynamics of Greedy Max-IG, Softmax-over-IG, and Random selection strategies. The left panel shows entropy loss, and the right panel shows the gradient norm.

C.5 Single-Path versus Multiple Continuation Paths

Table 2 compares single-path continuation (K=1) against keeping $K \in \{2, 4\}$ active paths. The total per-turn candidate budget is fixed and split evenly across the K continued paths, so additional paths come at the cost of a smaller candidate group per shared prefix; all other hyperparameters follow Appendix B. Under a fixed budget, maintaining extra paths therefore shrinks the within-group candidate pool and spends updates on divergent histories outside the shared prefix.

C.6 Query Diversity under Single-Path Continuation

We also investigate whether single-path continuation sacrifices search diversity. Figure 5 confirms it does not at the query level, as the distinct-query fraction within SP-GRPO sibling groups closely tracks that of chain-based rollouts throughout training (81.6% versus 88.3% at step 10, declining to 38.1% versus 46.4% at step 200 as both policies sharpen), a persistently narrow gap of 8–10 points despite all G candidates sharing an identical retrieval history. Moreover, measured with the same similarity criterion that GiGPO uses to identify identical states, the average pairwise similarity among distinct queries is only 0.27, indicating that they differ beyond mere paraphrases and constitute genuinely distinct search intents. Each forked candidate therefore probes a different facet of the remaining uncertainty, and the single continuation naturally alters the downstream context, providing sufficient exploration without multi-branch overhead.

D Ablation Protocols and Results

Token Entropy process reward. This variant replaces the ground-truth-anchored information gain with an answer-agnostic, self-calibrated uncertainty signal in the style of T^{2}PO. Each candidate's own response is forwarded once through the policy; per response token t, we compute the full-vocabulary Shannon entropy $H_t = -\sum_v p_t(v) \log p_t(v)$ and the top-j confidence $C_t = \frac{1}{j} \sum_{i \le j} \log p_t(v_i)$ (j=8). Both are min–max normalized within the response and fused as

$$M_t = \alpha\, H_t^n + (1-\alpha)(1 - C_t^n)\,, \qquad \alpha = 0.4, \tag{31}$$

with Φ_i the geometric mean of M_t over candidate i's response tokens, computed in log space as $\exp\big(\frac{1}{|a|} \sum_t \log M_t\big)$ with M_t clamped away from zero (high Φ = the policy was uncertain while generating it). In short, this variant replaces the ground-truth-anchored gain of Eq. 25 with an answer-agnostic Φ-drop. For the candidate continued at a turn, the score is

$$\Delta = \Phi_{\text{node}} - \frac{1}{|\text{ch}|} \sum_{k \in \text{ch}} \Phi_k, \tag{32}$$

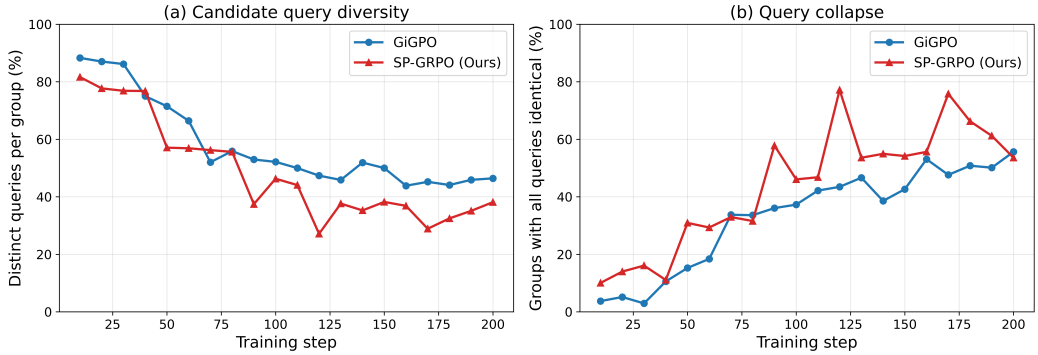

Figure 5: Candidate query diversity of SP-GRPO versus chain-based rollouts over training. (a) Fraction of distinct queries per candidate group. (b) Fraction of groups with all-identical queries.

Table 7: Process-reward and surrogate-scope ablations, full per-dataset EM (%). The ablation variants are evaluated at step 200, while SP-GRPO (Ours) reports the default run of Table 1. Best results are in **bold**, and second-best results are underlined.

Variant	Single-Hop QA			Multi-Hop QA				Avg.
	NQ†	TriviaQA*	PopQA*	HotpotQA†	2Wiki*	MuSiQue*	Bamb.*	
SP-GRPO (Ours)	**47.8**	**62.9**	**47.3**	41.8	**42.9**	16.0	**33.6**	**47.4**
w/ Token Entropy	43.4	57.3	44.5	32.0	32.9	11.7	26.4	41.0
w/ pruned-action surrogate	45.0	61.7	46.0	**41.9**	41.1	**17.6**	32.8	46.2

where ch are the G candidates forked from the continued node at the next turn, and the pruned query candidates of the group store a zero score. The signal is deferred by one round because the children are generated only after the continuation; it measures how much the executed step reduces the policy's uncertainty at the next fork, mirroring the parent–child answer-probability difference of the gold-target information gain in Eq. 25 with the teacher-forced answer probability replaced by the policy's own uncertainty. No ground truth is involved anywhere. The sampler and winner rule are unchanged; the deferred, path-based computation and the zero scores are simply how the Φ-drop is instantiated, not protocol changes beyond the reward replacement itself.

Surrogate on pruned actions. The default zeroes the advantage of the pruned query candidates (`zero_pruned_adv`), so the clipped surrogate in Eq. 5 updates only the executed query a_{t,j^*} and the collected answer candidates; the pruned query candidates still participate in the group mean and the KL term. The ablation (`keep_pruned_adv`) additionally keeps each pruned query candidate's computed group-relative advantage $\hat{A}_i$, so the surrogate backpropagates through all query candidates at s_t using their respective $\hat{A}_i$. This tests whether gradient signal on unexecuted branches helps or injects noisy credit.

As shown in Table 7, replacing the ground-truth-anchored gain with the answer-agnostic Φ-drop costs 6.4 points on average (47.4 → 41.0), with the largest drops on multi-hop benchmarks (HotpotQA 41.8 → 32.0, 2Wiki 42.9 → 32.9). An uncertainty signal computed from the policy's own responses cannot distinguish a query that resolves the current uncertainty from one that merely expresses it, and the distinction matters most when several hops must be chained. Keeping the surrogate on pruned actions costs 1.2 points (47.4 → 46.2). The pruned query candidates' advantages are computed under the shared prefix, but their tokens were never followed by environmental consequences, so backpropagating through them injects credit for counterfactual continuations the environment never took.

E Complete Algorithm

Algorithm 1 makes two rollout details explicit. First, the process/answer partition is decided only after all G candidates are sampled and stepped through the environment: answer candidates are identified by the EOS flag returned by the environment step, and they pass through the environment without issuing a retrieval. Second, the rollout terminates early once the number of collected answer trajectories reaches the outcome quota (G by default), even before the horizon T is exhausted; the remaining query candidates are then pruned without continuation, which shortens the average rollout and underlies the token efficiency reported in Section 4.3.3.

Algorithm 1 Shared-Prefix Group Relative Policy Optimization (SP-GRPO) for search agents

Require: Initial policy π_θ, reference model π_{ref}, gold-answer probe p_{ans} (Eq. 25), outcome reward R

Require: Group size G, horizon T, clip ε, KL coefficient β, softmax temperature η, learning rate ℓ

for iteration $= 1, 2, \ldots$ **do**
  Buffer $\mathcal{D} \leftarrow \emptyset$
  for each question q in batch **do**
    prefix $s \leftarrow (q)$; trajectory set $\mathcal{A}_q \leftarrow \emptyset$
    for $t = 1$ **to** T **do**
      Sample $\{a_{t,j}\}_{j=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid s)$ ▷ identical retrieval prefix
      $o_{t,j} \sim P(\cdot \mid s, a_{t,j})$ for all G candidates ▷ environment step; answer candidates emit EOS and retrieve nothing
      Partition into process group $\mathcal{P}_t$ and answer group $\mathcal{Y}_t$ ▷ after sampling, by the EOS flag
      for $j \in \mathcal{P}_t$ **do**
        $r_{t,j} \leftarrow p_{\text{ans}}(s \oplus a_{t,j} \oplus o_{t,j}) - p_{\text{ans}}(s)$ ▷ score with the retrieved documents
      end for
      Append $\{(s, a_{t,j})\}_{j \in \mathcal{Y}_t}$ to the trajectory set $\mathcal{A}_q$ ▷ each answer candidate completes a trajectory here
      if $|\mathcal{A}_q| \geq G$ **then**
        break ▷ outcome quota met; prune the remaining query candidates
      end if
      $x_{t,j} \leftarrow \exp(r_{t,j}/\eta) / \sum_{m \in \mathcal{P}_t} \exp(r_{t,m}/\eta)$ for $j \in \mathcal{P}_t$
      $\hat{A}_{t,j} \leftarrow (x_{t,j} - \mu_t)/(\sigma_t + \epsilon)$ for $j \in \mathcal{P}_t$ ▷ Eq. 2, within-turn statistics
      $j^* \sim \text{Uniform}(\mathcal{P}_t)$ ▷ content-agnostic continuation
      $a_t \leftarrow a_{t,j^*}$
      Store $(s, \{a_{t,j}\}_{j=1}^{G}, \{\hat{A}_{t,j}\}_{j \in \mathcal{P}_t}, j^*, \mathcal{Y}_t)$ into $\mathcal{D}$
      $s \leftarrow \text{concat}(s, a_t, o_{t,j^*})$ ▷ keep the executed query's documents
    end for
    $R_u \leftarrow R(u)$ for each collected trajectory $u \in \mathcal{A}_q$ ▷ exact-match outcome reward
    $\hat{A}_u^{\text{out}} \leftarrow (R_u - \mu_{\mathcal{Y}})/(\sigma_{\mathcal{Y}} + \epsilon)$ for $u \in \mathcal{A}_q$, attached to the answer candidates in $\mathcal{D}$ ▷ question-level answer group
  end for
  for minibatch in $\mathcal{D}$ **do**
    Set $\hat{A}_{t,j} \leftarrow 0$ for pruned query candidates $j \in \mathcal{P}_t,\ j \neq j^*$ ▷ excluded from the surrogate
    Compute token-level ratios $\rho_{i,k}(\theta)$ on executed queries and answer candidates
    Compute clipped surrogate of Eq. 5 using $\hat{A}_{t,j^*}$ on executed queries and $\hat{A}_u^{\text{out}}$ on answer candidates
    Compute token-level $\text{KL}(\pi_\theta \,\|\, \pi_{\text{ref}})$ on all G candidates
    $\theta \leftarrow \theta + \ell \nabla_\theta \mathcal{J}_{\text{SP-GRPO}}$
  end for
  $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
end for