StepOPSD: Step-Aware Online Preference Distillation for Agent Reinforcement Learning

Anonymous ACL submission

Abstract

Reinforcement learning for multi-turn agents suffers from a credit-assignment mismatch: rewards are sparse and trajectory-level, while success often hinges on a few local decisions. Existing online policy distillation (OPD) provides denser token-level supervision, but typically treats heterogeneous agent trajectories as monolithic strings rather than causal interaction units. We present StepOPSD, a post-rollout preference self-distillation framework that takes the agent step as the unit of credit redistribution. StepOPSD decomposes trajectories into action-centered step segments, rescoring them under hindsight-enriched teacher contexts and converting token-level log-probability gaps into sign-preserving advantage shaping with a normalized per-step credit budget before the GRPO update. Across ALFWorld and Search-QA with `Qwen3-1.7B` and `Qwen2.5-3B-Instruct`, StepOPSD attains best or second-best results on subsets most sensitive to local causal errors, including first-place performance on ALFWorld `Heat` (79.1%), `PickTwo` (95.0%), Search-QA `TriviaQA` (61.6%), and tied-best performance on `HotpotQA` (40.4%). The results further reveal a consistent two-knob law: smaller α_{clip} acts as a broadly stabilizing local trust region, whereas the optimal global mixing strength λ_{mix} remains task-dependent. These findings suggest that step-aware distillation is most useful when trajectory-level rewards are weakly aligned with the local action that determines downstream success.

1 Introduction

Agentic post-training for Large Language Models (LLMs) is increasingly shaped by two complementary paradigms (Yao et al., 2023; Schick et al., 2023; Jin et al., 2025b): Reinforcement Learning (RL), which optimizes task outcomes from environment feedback, and teacher-guided distillation, which supplies denser token-level supervision. In long-horizon agents, however, naively combining these signals creates a severe credit-assignment mismatch. A single sparse reward must supervise dozens of decisions, even though success often hinges on only one or two of them.

This mismatch has two structural sources. *First, failure is highly localized.* A malformed query, an invalid action, or a premature answer can ruin a rollout spanning thousands of tokens, yet standard RL broadcasts the same trajectory-level signal across the full response. *Second, agent trajectories are heterogeneous.* Unlike static reasoning traces, they interleave immutable environment observations with model-generated action spans. Treating the full trajectory as a uniform string wastes supervision on non-controllable tokens and blurs the causal boundaries that make hindsight feedback useful.

These observations suggest that the key problem is not merely how to inject a stronger teacher, but *where* credit should be redistributed once hindsight information is available. For multi-turn agents, the natural unit is not the whole response but the action-centered step. A useful distillation signal should therefore respect the step structure of interaction and concentrate optimization mass on the local decisions that actually determine downstream outcomes.

Driven by this view, we present StepOPSD (Step-Aware Online Preference Distillation). The term *preference* is intentional: unlike policy distillation, our method does not ask the student to imitate the teacher distribution. Instead, it uses OPSD-style teacher–student rescoring to induce bounded local preferences over realized step tokens, while the RL advantage still determines the update direction. In this sense, the teacher does not replace the policy objective; it only reshapes where credit is assigned within the sampled trajectory.

Operationally, StepOPSD decomposes com-

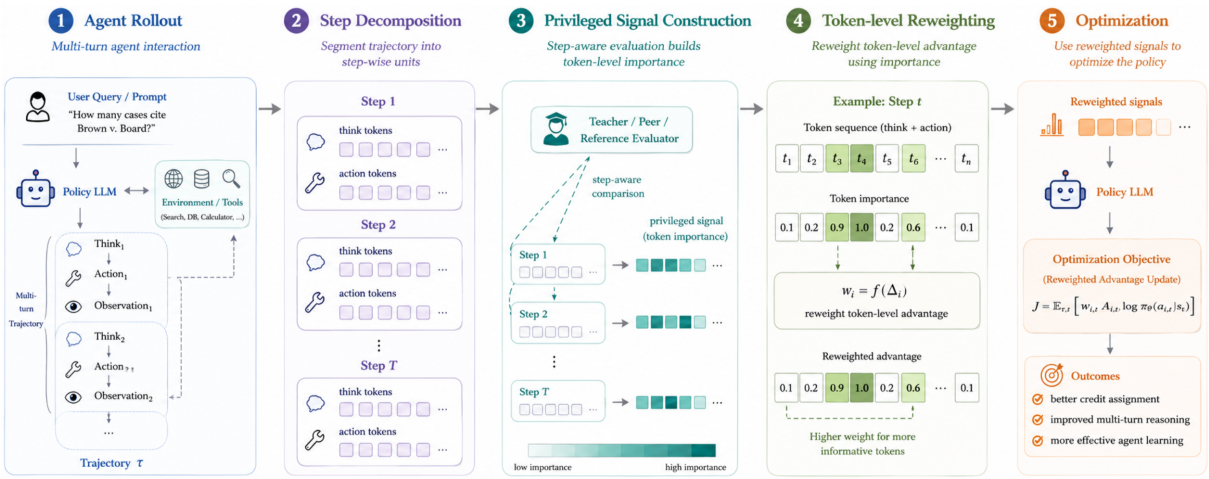

Figure 1: Overview of StepOPSD. StepOPSD decouples online environment interaction from offline credit shaping. While the base GRPO rollout dynamics remain untouched, post-rollout trajectories are structurally parsed into causal action boundaries. By rescoring these isolated spans under hindsight-enriched teacher contexts, StepOPSD translates token-level preference gaps into a direct modulation of the RL advantage—injecting dense, step-aware supervision without distorting the primary trajectory-level reward signal.

pleted trajectories into causal step segments, synthesizes hindsight-enriched teacher contexts for those spans, and converts teacher–student log-probability gaps into step-aware advantage shaping before the GRPO update. Unlike RLSD (Yang et al., 2026), which re-weights all tokens using self-divergence, or SDAR (Lu et al., 2026), which applies soft token-level gating over the full sequence, StepOPSD localizes supervision to the action-centered spans that actually drive agent behavior. The intervention is deliberately non-intrusive: online rollout dynamics remain untouched, and credit assignment is refined only in the post-rollout phase through a specialized reward manager and a lightweight trainer subclass.

To summarize, our primary contributions are three-fold:

- **Step-Aware Online Preference Distillation:** We introduce StepOPSD, which shifts distillation from monolithic responses to action-centered step segments and repurposes teacher rescoring as post-rollout credit redistribution rather than policy imitation.
- **Modular, Non-Intrusive Architecture:** A drop-in implementation atop Search-R1/GRPO that integrates diverse hindsight sources and span-selection strategies without changing online rollout dynamics.
- **Empirical Validation:** StepOPSD attains multiple best and second-best results on the subsets most sensitive to local decision errors, including first-place scores on ALFWorld `Heat` and `PickTwo`, and on Search-QA `TriviaQA` and `HotpotQA`. Beyond average gains, the results reveal a two-knob law: tighter local clipping through smaller α_{clip} is broadly stabilizing, whereas the optimal global mixing strength λ_{mix} remains task-dependent, with weaker shaping favoring embodied control and stronger shaping favoring retrieval-centric QA.

2 Method

2.1 The Credit Assignment Bottleneck

The standard paradigm for agentic fine-tuning relies on Group Relative Policy Optimization (GRPO) (Schulman et al., 2017; DeepSeek-AI, 2025), where a policy π_θ generates a multi-step trajectory $\tau = (s_1, a_1, o_1, \ldots, s_T, a_T)$, and a single terminal reward $R(\tau)$ is broadcast across all tokens. This is fundamentally misaligned with the nature of long-horizon tasks. A rollout spanning thousands of tokens often fails due to a single localized error—a malformed query, an invalid action, or a premature conclusion. Standard RL blindly smears the penalty across the entire sequence, forcing the agent to guess which step was fatal. Instead of learning a dense value model (which is notoriously unstable and prone to hallucination in agentic domains), StepOPSD addresses this bottleneck

through post-rollout hindsight distillation. Rather than introducing a standalone objective, StepOPSD acts as a surgical intervention within the existing pipeline:

$$\begin{aligned}&\text{rollout} \rightarrow \text{reward} \rightarrow \text{step extraction} \rightarrow \\ &\text{teacher–student rescoring} \rightarrow \\ &\text{advantage shaping} \rightarrow \text{policy update.}\end{aligned} \tag{1}$$

By decoupling the online interaction from offline credit shaping, we inject step-aware supervision directly into the GRPO advantage, preserving the stability of the base RL algorithm while surgically correcting localized errors.

2.2 Isolating Causal Action Spans

Agent trajectories are highly heterogeneous, interleaving immutable environment observations (o_t) with policy-generated actions (a_t). Applying distillation to the entire trajectory wastes modeling capacity on observation tokens that the policy cannot control. To resolve this, we structurally parse the completed trajectory into atomic step segments aligned with task-specific tags (e.g., `<action>...</action>`). In our implementation, we strictly employ `action_only` extraction for embodied tasks (ALFWorld), where rigid commands are the sole causal drivers. For knowledge-intensive tasks (Search-QA), we utilize `clean_step_no_observation`, which includes the agent's internal reasoning but masks out the retrieved external knowledge.

2.3 Hindsight-Privileged Rescoring

To correct a failed step, the agent must understand what it should have done. We achieve this by contrasting the agent's standard generation probability against a hindsight-privileged teacher. For each extracted step k, we define a student context c_k^S (the causal prefix) and a teacher context $c_k^T = c_k^S \oplus h_k$, where h_k is the injected hindsight information. While hindsight theoretically spans multiple forms—from a weak binary success flag to an expensive oracle demonstration—we exclusively adopt *peer-trajectory hindsight*. When a GRPO group contains both successes and failures, we condition the teacher context of a failed trajectory on the first successful peer from that exact same group. This critical design choice provides a dense, informative cue without incurring the distribution shift or annotation cost of external oracle demonstrations.

2.4 Quantifying Step-Level Credit

We quantify the localized "regret" of an action by measuring how much the hindsight information alters the likelihood of the generated step. For each token $z_{k,j}$ within step k, we compute the log-probability gap between the hindsight-aware teacher π_T and the student π_S:

$$\begin{aligned}\Delta_{k,j} = &\log \pi_T(z_{k,j} \mid c_k^T, z_{k,<j}) \\ &- \log \pi_S(z_{k,j} \mid c_k^S, z_{k,<j}).\end{aligned} \tag{2}$$

This Δ isolates critical decision nodes: tokens where the teacher strongly diverges from the student. To prevent moving-target instability during continuous RL, we instantiate the teacher π_T as a `stale_ref_policy` (the policy parameters from a fixed number of steps ago, e.g., 10 steps), ensuring the log-probability gap reflects a stable reference distribution.

2.5 Credit-Aware Advantage Shaping

The log-probability gap Δ must be translated into an optimization signal that modulates the GRPO advantage without destroying its core mathematical properties. Let A_ℓ be the base token-level advantage. We first construct a raw multiplicative weight and then project it onto a symmetric local trust region controlled by α_{clip}:

$$\begin{aligned}w_\ell^{\text{raw}} &= 2 \cdot \sigma(\text{sign}(A_\ell) \cdot \Delta_\ell), \\ w_\ell &= \Pi_{[1-\alpha_{clip},\, 1+\alpha_{clip}]}(w_\ell^{\text{raw}}),\end{aligned} \tag{3}$$

The final shaped advantage mixes the base RL signal with the hindsight weight via a decay parameter λ_{mix}:

$$\tilde{A}_\ell = (1 - \lambda_{mix})A_\ell + \lambda_{mix}(w_\ell A_\ell). \tag{4}$$

This formulation strictly preserves the sign of the original RL advantage. It only redistributes the *magnitude* of the optimization mass—amplifying updates on tokens where the teacher agrees, and dampening updates where the teacher diverges.

2.6 Causal Equivalence via Step Normalization

Finally, because steps naturally vary in token length, a purely token-wise advantage shaping would allow a long, verbose reasoning step to mathematically dominate a short, concise action step, simply by accumulating more Δ mass. We counter this by introducing an `equal_step_mean_abs`

constraint. The shaping weights within each extracted step are rescaled such that the mean absolute modification budget is uniform across all steps in the trajectory. The philosophical rationale is explicit: in the absence of dense sub-step rewards, the most principled prior is that each reasoning step serves as an equally critical causal link to the final outcome. The Δ signal is only used to distribute this fixed budget *internally* among the tokens of a given step, preventing verbosity from hijacking the credit assignment.

3 Experiments

3.1 Benchmarks

To comprehensively evaluate our method, we adopt two diverse agentic environments: ALFWorld (Shridhar et al., 2021) and Search-QA (Jin et al., 2025b). **ALFWorld** is a text-based embodied AI benchmark featuring 3,827 household tasks spanning six categories: Pick and Place (Pick), Look at Obj in Light (Look), Pick Clean then Place in Recep (Clean), Pick Heat then Place in Recep (Heat), Pick Cool then Place in Recep (Cool), and Pick Two Obj and Place (Pick2). **Search-QA** aggregates multiple search-augmented question-answering datasets. It includes single-hop datasets—NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2023)—as well as multi-hop reasoning tasks—HotpotQA (Yang et al., 2018), 2Wiki (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023).

3.2 Implementation Details

We optimize `Qwen3-1.7B` and `Qwen2.5-3B-Instruct` with StepOPSD for up to 150 steps on 4 A800 GPUs. For ALFWorld, we use standard splits, sampling 16 tasks per batch with 8 rollouts per prompt, a maximum prompt length of 2,048 tokens, and a response cap of 512 tokens. For Search-QA, following Search-R1 (Jin et al., 2025b), we use E5 as the dense retriever. The training split includes NQ and HotpotQA, while the remaining datasets are reserved for out-of-domain evaluation. Each batch contains 128 tasks with a maximum prompt length of 4,096 tokens.

Our implementation uses several configurations to stabilize credit shaping. We use a stale reference policy (`stale_ref_policy`, refreshed every 10 steps) as the teacher to avoid moving-target instability. Distillation targets isolate causal boundaries (`action_only` for ALFWorld and `clean_step_no_observation` for Seacrh-QA) rather than full trajectories, and hindsight contexts are enriched with the first correct peer trajectory from the same rollout group. We modulate the RL advantage with a sigmoid weight function, using a default instantiation with $\alpha_{clip} = 0.2$ and an initial $\lambda_{mix} = 0.2$ that linearly decays over 50 steps; later ablations show that tighter α_{clip} is often more stabilizing, whereas the best λ_{mix} remains environment-dependent. After the decay reaches zero, the stale teacher is still refreshed and StepOPSD statistics are still logged, but they no longer affect the actor update. We therefore read post-step-50 StepOPSD curves as diagnostics of teacher-student drift under a continuing stale teacher, rather than evidence that shaping remains equally active late in training. We also apply step-wise advantage normalization (`equal_step_mean_abs`) to prevent long steps from dominating the optimization landscape. We set the GRPO KL penalty to 0.01 and the invalid action penalty coefficient to 0.1.

3.3 Baselines

To ensure a rigorously fair comparison under identical infrastructure, we benchmark StepOPSD against three groups of approaches across our base models: **(1) Training-free methods:** *Skill-Prompt* retrieves task-relevant skills via keyword matching (KM) and prepends them to the input prompt at inference time. **(2) Post-training methods:** This includes standard GRPO (Schulman et al., 2017), vanilla OPSD, and *Skill-GRPO*. *Skill-GRPO* augments GRPO by retrieving skills and injecting them into the training prompt; at test time it can run with (*Skill-GRPO**) or without retrieved skills. **(3) Hybrid methods:** These fuse RL with privileged knowledge distillation. We compare against naive *GRPO+OPSD*, *Skill-SD* (Wang et al., 2026), *RLSD* (Yang et al., 2026), and the current state-of-the-art *SDAR* (Lu et al., 2026). *GRPO+OPSD* simply adds the OPSD distillation loss as an auxiliary objective on top of GRPO training.

3.4 Main Results

Table 1 reports the main comparison on ALFWorld and Search-QA across `Qwen3-1.7B` and `Qwen2.5-3B-Instruct`. The central pattern is not merely that embodied tasks prefer one setting while retrieval tasks prefer another. Rather, the

Method	Pick	Look	Clean	Heat	Cool	Pick2	Avg	NQ	Triv	Pop	Hotp	2Wk	MuS	Bam	Avg
Qwen3-1.7B															
Vanilla	25.0	22.2	3.1	0.0	21.4	4.2	12.5	29.4	46.9	37.0	23.5	19.6	6.4	10.5	24.8
Skill-Prompt*	10.3	**50.0**	16.1	0.0	5.0	9.4	14.1	29.4	46.5	36.2	22.9	20.8	4.3	10.1	24.3
OPSD	26.3	33.3	9.1	0.0	4.5	5.3	14.1	4.2	8.3	4.6	6.6	15.3	0.7	1.2	5.8
GRPO	**71.1**	41.7	36.4	40.0	31.8	31.6	42.1	40.0	**58.9**	43.5	35.4	30.3	12.0	65.7	40.8
Skill-GRPO	27.6	**54.5**	22.7	27.3	0.0	19.2	21.1	39.2	58.6	43.9	35.2	28.2	11.5	**66.1**	40.4
Skill-GRPO*	31.4	42.9	51.9	8.3	11.5	7.1	28.1	38.0	58.4	43.9	36.3	29.0	12.5	**66.9**	40.7
GRPO+OPSD	38.2	**50.0**	30.8	28.6	30.0	21.1	32.0	**40.7**	**58.9**	45.0	**37.0**	**34.6**	**13.3**	65.7	**42.2**
Skill-SD	52.9	37.5	**69.2**	**42.9**	**60.0**	**36.8**	52.3	39.1	57.5	**45.4**	34.8	34.1	10.7	64.1	40.8
RLSD	50.0	37.5	61.5	19.0	50.0	21.1	42.2	38.6	57.3	43.0	34.5	34.1	11.5	65.3	40.6
SDAR	**73.5**	25.0	**76.9**	33.3	40.0	**36.8**	**53.9**	39.7	**58.9**	45.3	35.9	**35.5**	**12.6**	65.3	41.9
StepOPSD ($\lambda_{mix} = 0.05$)	64.7	44.4	56.5	**60.9**	42.1	**55.0**	**56.3**	40.5	58.6	**45.6**	34.9	29.8	10.8	64.5	41.4
StepOPSD ($\lambda_{mix} = 0.2$)	58.8	44.4	52.2	26.1	**52.6**	20.0	43.8	**40.6**	**59.4**	44.4	**37.1**	32.0	11.6	64.1	**42.0**
Qwen2.5-3B-Instruct															
Vanilla	44.4	11.1	6.2	15.4	28.6	12.5	21.9	24.6	48.1	31.0	26.3	25.3	7.2	59.7	31.7
Skill-Prompt*	51.7	66.7	48.4	0.0	4.3	10.0	28.9	23.7	46.2	30.6	24.4	22.1	7.5	12.5	23.9
OPSD	48.8	41.7	16.7	0.0	15.8	16.7	23.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0
GRPO	91.2	62.5	**96.2**	61.9	65.0	47.4	70.0	43.5	58.8	43.0	36.8	32.2	11.7	**72.5**	42.6
Skill-GRPO	**98.9**	71.4	58.8	**70.6**	40.7	29.2	60.2	44.3	59.6	44.3	39.0	36.1	14.5	66.4	43.5
Skill-GRPO*	94.3	57.1	**100.0**	66.7	73.1	57.1	80.5	44.3	59.6	44.3	39.0	36.1	14.5	14.9	36.1
GRPO+OPSD	**100.0**	**82.4**	85.7	**75.0**	70.0	60.0	81.2	**44.9**	61.2	45.2	**40.4**	38.5	**16.0**	66.1	**44.6**
Skill-SD	88.2	50.0	**96.2**	52.4	65.0	57.9	73.4	44.4	60.4	44.0	**39.5**	**40.4**	15.4	64.9	44.1
RLSD	87.9	**75.0**	90.9	**75.0**	73.1	68.4	79.7	41.5	58.6	42.3	**40.4**	**40.2**	**16.8**	**66.9**	43.8
SDAR	97.1	62.5	**100.0**	61.9	**75.0**	**84.2**	**84.4**	44.8	58.1	44.3	38.6	36.2	15.7	66.1	43.4
StepOPSD ($\lambda_{mix} = 0.05$)	82.4	66.7	82.6	52.2	73.7	75.0	73.4	43.6	61.2	43.8	39.2	38.1	15.8	64.5	43.7
StepOPSD ($\lambda_{mix} = 0.05$, $\alpha_{clip} = 0.05$)	–	–	–	–	–	–	–	**45.0**	**61.6**	**46.2**	**39.5**	39.5	14.4	65.3	**45.7**
StepOPSD ($\lambda_{mix} = 0.2$)	45.0	55.6	87.0	65.2	57.9	**85.3**	69.5	44.4	**61.4**	**46.8**	**40.4**	38.1	14.1	65.0	44.3
StepOPSD ($\lambda_{mix} = 0.2$, $\alpha_{clip} = 0.05$)	97.1	66.7	87.0	**79.1**	**78.9**	**95.0**	**83.6**	–	–	–	–	–	–	–	–

Table 1: Performance on ALFWorld and Search-QA across the 1.7B and 3B model scales. We report success rate (%) on ALFWorld and accuracy (%) on Search-QA. We additionally include two informative 3B StepOPSD variants with altered α_{clip}; – indicates the corresponding task was not run for that follow-up.

benefit of StepOPSD varies systematically with the *specific failure mode* of each subset: the larger the mismatch between a sparse trajectory-level reward and the true local decision that determines success, the larger the value of step-aware credit shaping.

Fine-grained patterns on ALFWorld. Among the six embodied subsets, `Pick` and `Look` are relatively short-horizon and less credit-ambiguous: once the agent correctly grounds the target object or light source, the remaining path is direct. Standard RL or skill-augmented baselines therefore remain competitive there, e.g., SDAR reaches 73.5% on `Pick` at 1.7B and GRPO+OPSD reaches 82.4% on `Look` at 3B. By contrast, `Heat`, `Cool`, and `Clean` involve latent state transitions that are invisible in the final reward decomposition: an agent may correctly find, carry, and place the object, yet still fail because one intermediate action did not actually change its state. Here, trajectory-level RL is especially wasteful because it punishes the whole rollout for what is often a single local mistake. Step-aware shaping is most effective in exactly this regime. The clearest case is `Heat` at 1.7B, where StepOPSD with $\lambda_{mix} = 0.05$ reaches 60.9%, well above SDAR (33.3%) and GRPO (40.0%), indicating that the post-hoc teacher signal corrects the decisive appliance-interaction step rather than globally suppressing exploration. At 3B, the reduced-α_{clip} run ($\lambda_{mix} = 0.2$, $\alpha_{clip} = 0.05$) strengthens this picture, reaching the best `Heat` score (79.1%) and the best `Cool` score (78.9%). `Clean` shows a milder version of the same pattern, suggesting that its key subgoal is easier for competitors to verbalize once discovered. `Pick2` exposes a different bottleneck: not hidden state transition, but cross-subgoal bookkeeping. After partially solving object 1, the agent must preserve progress while initiating and completing object 2. On 1.7B, the gentler $\lambda_{mix} = 0.05$ achieves the best score (55.0%), showing that weaker shaping better protects fragile exploration. On 3B, the reduced-α_{clip} run reaches the best `Pick2` score (95.0%), suggesting that once the base policy is more competent, strong global guidance remains effective if local over-correction is tightly bounded. Together, these patterns explain why ALFWorld benefits strongly from tighter local clipping at 3B, even though the low-λ_{mix} rows still reveal what happens when that control is absent.

Fine-grained patterns on Search-QA. The Search-QA results show an equally sharp but different decomposition. `NQ` is relatively short-chain and entity-centric, so the main bottleneck is often whether the agent retrieves the answer once, not whether it sustains a long tool-use chain. StepOPSD is therefore competitive but not always dominant there under its default setting. The reduced-α_{clip} 3B follow-up sharpens this picture: with $\lambda_{mix} = 0.05$ and $\alpha_{clip} = 0.05$, StepOPSD reaches the best `NQ` score (45.0%), suggesting

that this subset benefits from preserving search freedom once local over-correction is suppressed. TriviaQA, by contrast, is highly sensitive to lexical formulation, aliases, and evidence phrasing; a slightly better search query often flips the entire example. This is exactly where step-level shaping should matter most, and StepOPSD indeed reaches the best TriviaQA scores at both scales, including 61.6% for the reduced-α_{clip} 3B run. PopQA follows a similar logic: many questions are factual and seemingly shallow, but they are long-tail and brittle to query wording, so better credit assignment at the precise retrieval step again helps StepOPSD reach the best 1.7B score (45.6% with $\lambda_{mix} = 0.05$) and remain near-best at 3B, with 46.8% for $\lambda_{mix} = 0.2$ and 46.2% for the reduced-α_{clip} $\lambda_{mix} = 0.05$ follow-up. HotpotQA makes the mechanism even clearer. Because the second hop depends on an intermediate bridge entity, one malformed first query can poison the rest of the rollout. A denser teacher signal is therefore disproportionately valuable, enabling StepOPSD with $\lambda_{mix} = 0.2$ to achieve the best 1.7B score (37.1%) and tie the best 3B score (40.4%). By contrast, 2Wiki and especially MuSiQue remain harder for our current variant. Their failures are not always caused by one bad search action; they often arise from multi-hop evidence composition under stronger distractors and deeper dependency chains. In those cases, local step correction still helps, but its advantage is diluted when the dominant error lies in global evidence aggregation rather than in a single retrieval decision. Bamboogle is also revealing: once the model issues one sufficiently targeted search, strong RL baselines already perform well, so the room for improvement shrinks, and even our reduced-α_{clip} 3B follow-up remains below GRPO's 72.5%.

A consistent task-dependent law across scales. Taken together, Table 1 reveals a sharper empirical law than a simple weak-vs.-strong shaping dichotomy. Embodied subsets with rigid physical dynamics consistently benefit from tighter local control: at 3B, keeping $\lambda_{mix} = 0.2$ but tightening α_{clip} to 0.05 raises the finalized ALFWorld average to 83.6% and produces the best Heat, Cool, and Pick2 scores. Retrieval subsets remain more sensitive to the interaction between global mixing and local clipping. Under the default α_{clip}, $\lambda_{mix} = 0.2$ yields the stronger 3B Search-QA average (44.3% vs. 43.7%), but once α_{clip} is tightened to 0.05, the lighter-λ_{mix} variant rises to the best 3B Search average (45.7%). The transferable pattern is therefore not that one global λ_{mix} universally wins, but that tighter local clipping through smaller α_{clip} is broadly beneficial, while the optimal global λ_{mix} remains task-dependent. This is the central empirical message of StepOPSD: step-aware distillation is most effective when trajectory-level rewards are maximally misaligned with the local causal action that determines downstream success.

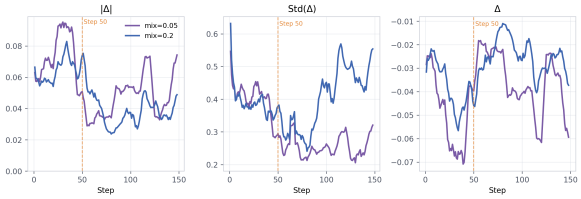

Figure 2: StepOPSD dynamics on ALFWorld for Qwen3-1.7B. Around step 50, heavy shaping ($\lambda_{mix} = 0.2$) induces a variance explosion in the teacher-student gap.

3.5 Training Diagnostics: Pre- and Post-Step 50 Phase Transition

End-task accuracy alone does not reveal how StepOPSD works. We therefore track the standard deviation of the teacher-student gap, Std(Δ), over training. The key pattern is a phase transition around step 50, when the base RL policy begins to mature. This turning point must be read with the implementation detail above: by step 50, the explicit StepOPSD mixing coefficient has decayed to zero, so later curves mainly reflect student drift relative to a stale teacher still refreshed every 10 steps, rather than a still-active shaping term.

In Search-QA, sparse rewards leave the student more dependent on the teacher. Before step 50, all models show similar variance (Std(Δ) $\approx$ 0.40–0.49). After step 50, the weakly shaped student ($\lambda_{mix} = 0.05$) jumps to 0.61, while stronger shaping ($\lambda_{mix} = 0.2$) remains stable at about 0.44. Without enough guidance, exploration drifts.

In ALFWorld, the pattern reverses. Once the student learns basic controls, heavy shaping ($\lambda_{mix} = 0.2$) creates a "tug-of-war" between hindsight paths and RL exploration, pushing Std(Δ) from 0.49 to 0.61 after step 50. By contrast, milder shaping, via lower λ_{mix} or tighter clipping, lets the policy optimize reward without over-constraint and drives Std(Δ) down (e.g., from 0.46 to 0.26).

This mirrored phase transition across environments supports our theoretical variance bound (The-

Figure 3: Three-way StepOPSD ablation on ALFWorld with `Qwen2.5-3B-Instruct`. We compare the heavy baseline ($\lambda_{mix} = 0.2$, $\alpha_{clip} = 0.2$), the lower-λ_{mix} variant ($\lambda_{mix} = 0.05$, $\alpha_{clip} = 0.2$), and the reduced-α_{clip} variant ($\lambda_{mix} = 0.2$, $\alpha_{clip} = 0.05$). The validation panel marks peak and final checkpoints; the others show training dynamics and stage-wise summaries over steps 0–50, 50–100, and 100+.

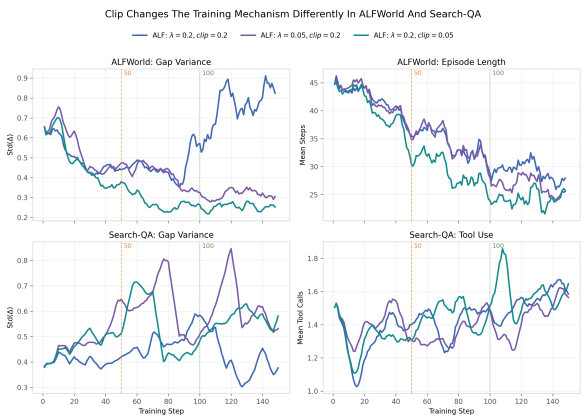

Figure 4: Mechanistic view of weight clipping at 3B. In ALFWorld, tightening α_{clip} under fixed $\lambda_{mix} = 0.2$ prevents the late-stage explosion of $\mathrm{Std}(\Delta)$ and shortens embodied trajectories. In Search-QA, tightening α_{clip} under fixed $\lambda_{mix} = 0.05$ lowers gap variance relative to the loose-clip counterpart and increases tool use. The phase markers at steps 50 and 100 show that clipping matters most after the policy enters the mature regime.

orem 2): StepOPSD reduces gradient variance only when distillation strength is calibrated to the density of the environment reward.

3.6 Ablation Studies: Weight Clipping as a Cross-Task Stabilizer

We use ALFWorld at 3B because it gives the cleanest three-way comparison among StepOPSD control knobs: a heavy baseline ($\lambda_{mix} = 0.2, \alpha_{clip} = 0.2$), a lower-$\lambda_{mix}$ variant ($\lambda_{mix} = 0.05, \alpha_{clip} = 0.2$), and a reduced-$\alpha_{clip}$ variant ($\lambda_{mix} = 0.2, \alpha_{clip} = 0.05$). Figure 3 shows that they stay close early but separate clearly after step 50. The reduced-α_{clip} run yields the best mature-phase validation and the highest logged peak.

The dynamics show that the main issue is not shaping itself, but unbounded local correction during the active phase. After step 50, the heavy baseline reaches a late-stage gap variance of $\mathrm{Std}(\Delta) = 0.770$ over steps 101–150, whereas the reduced-α_{clip} run stays at 0.252. Because λ_{mix} has already decayed to zero, we read these later curves as drift against stale-teacher snapshots rather than continued active shaping. The reduced-α_{clip} run matters because it leaves behind a more stable mature policy.

Figure 4 shows the same mechanism across tasks. In ALFWorld, clipping turns strong shaping into bounded strong shaping: the reduced-α_{clip} run keeps the lowest late-stage variance and shortens the mature policy from 29.2 to 24.3 actions, without changing the near-perfect valid-action ratio. In Search-QA, the effect is different but consistent. Under weak global mixing, tightening α_{clip} lowers late-stage gap variance, raises tool use from 1.45 to 1.59, and shortens responses from 98.9 to 86.4 tokens. The gain therefore comes less from formatting correctness than from reallocating behavior toward actual search.

Taken together, the 3B ablations reveal a more specific trade-off than "more vs. less distillation." The global mixing coefficient λ_{mix} controls how strongly the teacher anchors the update, whereas the clipping coefficient α_{clip} controls how far any local token correction can deviate from the RL signal. Tight clipping helps in both environments because it prevents a small subset of tokens from dominating the post-rollout correction; what changes across tasks is whether this bounded correction stabilizes embodied action plans or encourages more tool-centric retrieval behavior.

4 Related Work

4.1 Agentic RL and Credit Assignment

Agentic language models act through temporally extended trajectories rather than single-pass responses. In formulations such as ReAct, these trajectories are naturally structured as interleaved reasoning, actions, and observations (Yao et al., 2023; Schick et al., 2023). Recent RL work has scaled this paradigm to search agents and other long-horizon settings (Jin et al., 2025b,a), but also makes the credit-assignment problem more severe: rewards are sparse and trajectory-level, while success often depends on a few local decisions. Our work targets this mismatch directly.

4.2 On-Policy Distillation and Self-Distillation

Policy distillation transfers behavior through distribution matching (Rusu et al., 2015), while more recent on-policy variants apply supervision on student-generated rollouts. In particular, OPSD-style methods use privileged or hindsight contexts to define a stronger teacher on the student's own trajectories (Zhao et al., 2026; Penaloza et al., 2026). For multi-turn agents, Skill-SD, RLSD, and SDAR all study how dense self-distillation can complement RL (Wang et al., 2026; Yang et al., 2026; Lu et al., 2026). StepOPSD belongs to this line, but localizes supervision to extracted action-centered step spans instead of distilling an entire trajectory.

4.3 Reward Shaping and Advantage Reweighting

Our method is also related to reward shaping and policy-preserving auxiliary guidance. Classical work showed that shaping signals can accelerate learning without changing the original optimum when designed appropriately (Ng et al., 1999; Wiewiora, 2003). Modern policy optimization likewise uses comparative signals to modulate updates without replacing the base objective (Schulman et al., 2017; Rafailov et al., 2024; Azar et al., 2024). StepOPSD follows this principle: RL remains the backbone, while teacher-derived token gaps modulate post-rollout advantages.

4.4 Step-Structured Supervision for Agents

Multi-turn agent trajectories are heterogeneous: observations, tool calls, reasoning, and answers do not play the same causal role. Supervising the full sequence uniformly can therefore blur the boundary between context and action. Recent agent distillation methods introduce richer training-time contexts, such as privileged information, skill summaries, or token gating (Penaloza et al., 2026; Wang et al., 2026; Lu et al., 2026). Our contribution is to make the supervision unit explicit by extracting action-centered step segments and injecting their teacher-student comparison directly into advantage shaping.

5 Conclusion

This paper presents StepOPSD, a step-aware extension of Online Preference Distillation for multi-turn agent reinforcement learning against Online Policy Distilation (OPD). Instead of distilling an entire trajectory uniformly, StepOPSD isolates the action-centered spans that causally determine agent behavior and feeds hindsight teacher signals back into token-level advantage shaping. The method is faithful to the underlying GRPO training stack, modular in implementation, and naturally aligned with the heterogeneous structure of agent trajectories.

Empirically, our results show that the main value of StepOPSD is not a uniform gain on every benchmark average, but a more precise correction of the subsets most sensitive to local decision errors. Across ALFWorld and Search-QA, StepOPSD consistently achieves multiple best and second-best results, while revealing a sharper two-knob law: tighter local clipping through smaller α_{clip} is broadly beneficial, whereas the optimal global λ_{mix} remains task-dependent. This supports our central claim that step-aware distillation is most effective when sparse trajectory-level rewards are weakly aligned with the local causal action that determines downstream success.

Limitations

While StepOPSD demonstrates strong potential for step-aware credit assignment, we acknowledge two main limitations in the current study.

First, our ablation studies indicate that the optimal mixing strength (λ_{mix}) varies across different environment types—requiring gentler shaping for action-centric tasks and stronger shaping for knowledge-intensive retrieval tasks. This introduces an additional hyperparameter tuning step when deploying the method to entirely new domains.

Second, the current framework focuses strictly

on post-rollout advantage shaping. While this ensures stable integration with the underlying RL backbone, it does not dynamically intervene during the generation phase. Exploring how to use the step-level distillation signal to guide real-time decoding remains an important direction for future work.

References

Mohammad Gheshlaghi Azar, Le Zhou, Aviral Kumar Singh, and 1 others. 2024. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Bowen Jin, Jinsung Yoon, Priyanka Kargupta, Sercan O. Arik, and Jiawei Han. 2025a. An empirical study on reinforcement learning for reasoning-search interleaved llm agents. *arXiv preprint arXiv:2505.15117*.

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

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. 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)*, pages 1601–1611.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. 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)*, pages 9802–9822.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 278–287. Morgan Kaufmann.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. 2026. Privileged information distillation for language models. *Preprint*, arXiv:2602.04942.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. 2015. Policy distillation. *arXiv preprint arXiv:1511.06295*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

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

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

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

Eric Wiewiora. 2003. Potential-based shaping and q-value initialization are equivalent. *Journal of Artificial Intelligence Research*, 19.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. 2026. Self-distilled rlvr. *Preprint*, arXiv:2604.03128.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. 2026. Self-distilled reasoner: On-policy self-distillation for large language models. *Preprint*, arXiv:2601.18734.

A Theoretical Analysis

In this appendix, we provide a formal analysis of StepOPSD's properties, specifically focusing on how it modulates the reinforcement learning gradients without shifting the optimal policy, and how it mitigates the credit assignment problem in long-horizon agent trajectories.

Let the standard policy gradient for a token y_t be $g_t = A_t \nabla_\theta \log \pi_\theta(y_t|s_t)$. StepOPSD modifies this advantage to $\tilde{A}_t = \Psi_t A_t$, where $\Psi_t = 1 - \lambda_{mix} + \lambda_{mix} w_t$ is the shaping multiplier, $w_t = 2\sigma(\text{sign}(A_t) \cdot \Delta_t) \in (0, 2)$ is the weight function, and $\Delta_t = \log \pi_{ref}(y_t|s_t^+) - \log \pi_\theta(y_t|s_t)$ is the teacher-student gap.

A.1 Property 1: Sign Preservation and Directional Consistency

A critical requirement for any advantage shaping method is that it must not introduce pathological local optima that deviate from the true environment reward.

Proposition 1 (Sign Preservation)**.** *For any $\lambda_{mix} \in [0, 1)$, the shaped advantage $\tilde{A}_t$ strictly preserves the sign of the original RL advantage A_t, i.e., $sign(\tilde{A}_t) = sign(A_t)$.*

Proof. By definition, the shaping multiplier is $\Psi_t = 1 - \lambda_{mix} + \lambda_{mix} w_t$. Since $w_t > 0$ and $\lambda_{mix} \in [0, 1)$, we are guaranteed that $1 - \lambda_{mix} > 0$, and thus $\Psi_t > 0$. Consequently, multiplying A_t by a strictly positive scalar Ψ_t preserves its sign: $\tilde{A}_t = 0 \iff A_t = 0$, and $\tilde{A}_t > 0 \iff A_t > 0$. □

Theorem 1 (Directional Consistency)**.** *Let $g_{RL} = \mathbb{E}[A_t \nabla_\theta \log \pi_\theta]$ and $g_{StepOPSD} = \mathbb{E}[\tilde{A}_t \nabla_\theta \log \pi_\theta]$. Under the assumption that the shaping multiplier Ψ_t is bounded and positive, the expected StepOPSD gradient maintains a non-negative cosine similarity with the exact RL gradient, acting as a valid descent direction for the original RL objective.*

Proof. Because $\tilde{A}_t = \Psi_t A_t$ and $\Psi_t > 0$ (Proposition 1), the gradient step for every single token is scaled by a strictly positive magnitude. In expectation, reweighting the components of a gradient vector by strictly positive scalars ensures that the modified vector lies in the same half-space as the original vector ($\langle g_{RL}, g_{StepOPSD} \rangle \geq 0$). Thus, StepOPSD optimizes toward the same global reward maximum as standard GRPO, merely altering the traversal path across the loss landscape. □

A.2 Property 2: Variance Reduction in Multi-Turn Credit Assignment

In multi-turn tasks, the standard advantage A_t for a step k is estimated using the trajectory return, introducing high variance due to the compounding uncertainty of future steps $k + 1 \ldots K$.

Theorem 2 (Variance Bound via Hindsight Teacher)**.** *Let $A_t = A_t^* + \epsilon_t$ where A_t^* is the oracle local advantage of the token and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is the noise from future rollout variability. If the teacher's gap Δ_t provides an unbiased signal of A_t^* such that the weight w_t dampens (i.e., $\Psi_t < 1$) when $sign(A_t) \neq sign(A_t^*)$, the variance of the StepOPSD gradient estimator is bounded below the standard estimator: $Var(g_{StepOPSD}) < Var(g_{RL})$.*

Proof. The variance of the standard gradient is proportional to $\mathbb{E}[A_t^2] = (A_t^*)^2 + \sigma^2$. In StepOPSD, the multiplier Ψ_t suppresses the advantage magnitude precisely when the trajectory-level return contradicts the local step-level teacher evaluation ($\text{sign}(A_t) \cdot \Delta_t < 0$, yielding $w_t < 1$ and thus $\Psi_t < 1$). By actively discounting these high-noise, low-confidence token updates, the contribution of the noise term ϵ_t to the gradient variance is scaled down by $\mathbb{E}[\Psi_t^2 | \text{contradiction}] < 1$, yielding a lower overall gradient variance. □

This mathematical property formally explains why the stronger mixing strength ($\lambda_{mix} = 0.2$) achieves superior performance in complex Search-QA tasks: it effectively mitigates the credit-assignment noise σ^2 inherent to long, multi-document reasoning chains.