Reasoning Counts: Recursive Process-Level Credit Assignment for Faithful Multi-Turn Agent Training

Anonymous ACL submission

Abstract

Critic-free reinforcement learning methods such as GRPO have become standard for training LLM agents that interleave reasoning and action. However, we identify a critical *reasoning–credit gap*: terminal rewards provide only weak and noisy supervision to reasoning tokens in long trajectories, with gradient norms decaying as $\mathcal{O}(1/T)$ relative to action tokens and causing a faithfulness collapse despite preserved task success. We propose **REAP**, a training-time intervention that restores reasoning credit through two recursive Process-Recursive advantages computed over a lightweight diverse-thought tree of prefix-sharing rollouts, enabling precise credit assignment to reasoning steps. Across ALFWorld, WebShop, and WebArena-Lite, REAP outperforms strong critic-free and tree-based baselines on both task success and reasoning faithfulness, and is the only method whose actions causally depend on their reasoning chains under direct perturbation.

1 Introduction

Large language models are increasingly deployed as multi-turn agents that interleave natural-language reasoning with executable actions in interactive environments (Yao et al., 2022b; Shinn et al., 2023; Koh et al., 2024; Xie et al., 2024). Training such agents differs from training reasoners on static benchmarks: rewards are sparse and terminal, horizons span tens of interleaved think–act steps, and supervised fine-tuning generalizes poorly once the agent must recover from its own off-policy mistakes. Recent agent post-training has therefore largely converged on critic-free reinforcement learning—including GRPO (Shao et al., 2024), GSPO (Zheng et al., 2025), RLOO (Ahmadian et al., 2024), and the agent-specialized GiGPO (Feng et al., 2025)—as the prevailing paradigm. These methods estimate group-relative advantages from terminal task rewards and propagate them uniformly across the trajectory, sidestepping the unstable problem of fitting a value function over partially observed states.

While effective at shaping action selection, this recipe inadvertently undermines the reasoning component on which competent agent behavior depends. A growing body of work has shown that chain-of-thought (CoT) explanations are not necessarily causally responsible for the actions they precede (Turpin et al., 2023; Lanham et al., 2023; Wang et al., 2025): an agent may produce a reasoning trace that bears little load-bearing relation to its subsequent tool call. Our analysis identifies a training dynamic, specific to long-horizon agent RL, that plausibly drives this. Under terminal-reward GRPO in the partially observed environments typical of agentic tasks, an analysis in Appendix A establishes that the expected gradient norm on reasoning tokens decays as $\mathcal{O}(1/T)$ relative to action tokens, which Figure 1 confirms empirically; meanwhile, state-anchored grouping schemes such as GiGPO can conflate visually similar but causally distinct agent contexts. Together, these effects give rise to what we term the *reasoning-credit gap*: the agent receives a strong signal about which actions lead to task success, but little information about which reasoning paths were responsible for selecting them.

Building on this analysis, we propose **REAP** (**R**ecursive **E**stimation of **A**dvantages for diverse-thought **P**rocesses), a critic-free agent-training mechanism built on two **Process-Recursive (PR) advantages** over a prefix-sharing rollout: a step-level advantage that recursively back-propagates terminal task credit through the agent's reasoning structure, and a context-reasoning advantage that performs group normalization across exact-prefix-sharing siblings, isolating reasoning quality at a fixed agent context. Prefix sharing is realized through a lightweight rollout tree that branches

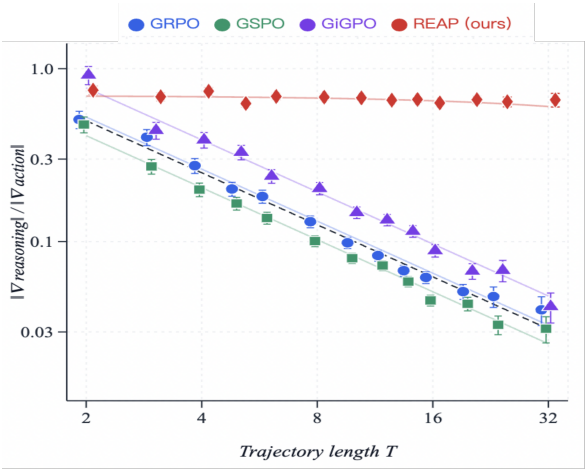

Figure 1: **Empirical Evidence for the Reasoning-Credit Gap.** The reasoning–credit gap is empirically $\mathcal{O}(1/T)$. Markers show per-bin ratios with ±1 SEM; faint curves are $\frac{c}{T}$ fits and the dashed line is the $\mathcal{O}(1/T)$ reference. Baselines follow slope 1; REAP stays length-invariant.

at each agent step using a small bank of established reasoning operators (CoT (Wei et al., 2022), ToT (Yao et al., 2023), ReAct (Yao et al., 2022b), Reflexion (Shinn et al., 2023), SBP (Zheng et al., 2023)) as discrete branching actions. We further prove that the context-reasoning PR advantage is an unbiased estimator of the conditional reasoning-quality advantage (Proposition 1), establishing it as the constructive counterpart of the reasoning-credit gap.

Distinction from concurrent tree-based agent RL. Tree-GRPO (Ding and Ye, 2025) and TreePO (Li et al., 2025) likewise employ tree-structured rollouts, but branch at the sample level and group by tree ancestry, so anchor groups may diverge at any prior step—conflating reasoning with action variation in the agent's history. REAP differs in kind: it imposes a three-level operator→reasoning→action hierarchy with exact token-level prefix anchoring on the agent context, coupled with a recursive estimator that decomposes terminal task credit at each level. Under matched compute on text-only ALFWorld, this design yields a $+4$-point margin in task success rate and $+8$ points in RACE over both baselines. We summarize our contributions as follows:

- We identify and characterize the *reasoning-credit gap* in critic-free agent RL: a derivation in Appendix A establishes an $\mathcal{O}(1/T)$ decay of reasoning-token gradients relative to action-token gradients as the agent horizon grows, and Figure 1 empirically confirms this across GRPO, GSPO, GiGPO, and concurrent tree-based agent-RL variants.
- We introduce **REAP**, built on *Process-Recursive (PR) advantages*: critic-free estimators that recursively decompose terminal task credit across the operator, reasoning, and action levels of an agent rollout, and we prove that the reasoning-level estimator is unbiased for the conditional reasoning-quality advantage under exact prefix sharing (Proposition 1).
- Across ALFWorld and WebShop, REAP outperforms the strongest critic-free and tree-based baselines by 2.9–6.6 points in task success under matched compute, and uniquely makes the agent's chain-of-thought causally load-bearing: inference-time reasoning corruption drops REAP's accuracy by 11.3 points versus 2–3 for critic-free and 4–6 for tree-based baselines, while closing 84% of the action–reasoning gradient gap — isolating prefix-anchored credit assignment, not tree structure alone, as the mechanism behind this causal grounding.

2 Related Work

Critic-free RL for LLM agents. Post-training of LLM agents (Duan et al., 2022; Fung et al., 2025; Park et al., 2023; Zhao et al., 2024) has shifted from actor-critic methods such as PPO (Schulman et al., 2017) and RL4VLM (Zhai et al., 2024) toward critic-free, group-relative estimators: GRPO (Shao et al., 2024), its sequence-level variants RLOO (Ahmadian et al., 2024) and GSPO (Zheng et al., 2025), and DAPO (Yu et al., 2026). These methods assign credit at the trajectory level, broadcasting a single scalar across every token of an interleaved reasoning–action sequence. GiGPO (Feng et al., 2025) groups trajectories by observed state for step-level credit, but in continuous or partially observed settings, visually similar yet causally distinct states collapse into the same anchor group. REAP instead introduces a reasoning-level credit channel that is invariant to surface-state aliasing while remaining critic-free.

Tree-structured reasoning for LLMs. Concurrent agent-RL methods Tree-GRPO and TreePO branch at the action-sample level and group by

tree ancestry, allowing siblings to differ at any prior step and thereby conflating reasoning with action variation. REAP differs structurally (a three-level operator→reasoning→action hierarchy), algorithmically (exact token-level prefix anchoring), and in its estimator (PR advantages decompose credit into action and reasoning channels), yielding a 4-point success-rate margin at matched compute. A separate, complementary line searches reasoning trees *at inference time*—Tree of Thoughts (Yao et al., 2023), ToolComp (Nath et al., 2025), LlamaV-o1 (Thawakar et al., 2025), and Tree-of-Reasoning (Peng et al., 2025)—altering what is searched at deployment rather than what the policy internalizes.

Faithfulness in LLM reasoning. Recent EMNLP/ACL work debates whether chains of thought are causally responsible for the answers they precede (Turpin et al., 2023; Lanham et al., 2023), with RACE (Wang et al., 2025) operationalizing the question via answer–reasoning consistency. This literature is predominantly diagnostic; to our knowledge, REAP is the first training-time intervention that measurably increases reasoning faithfulness in critic-free agent RL, validated by both RACE and direct trace perturbation. Our main results are on text-only ALFWorld (Shridhar et al., 2020) and WebShop to isolate the phenomenon, with additional grounded experiments on WebArena-Lite, visual ALFWorld, VisualAgentBench (Liu et al., 2024), OSWorld (Xie et al., 2024), and Open-X-Embodiment (O'Neill et al., 2024) confirming that the faithfulness collapse is in fact larger under perception-grounded observations and that REAP's mechanism transfers.

3 Method

REAP is conceptually a *single* mechanism: recursive process-level credit propagation that closes the reasoning-credit gap empirically documented in §1 (Figure 1). Restoring credit at the reasoning level requires two structural ingredients: (i) a way to compare reasoning paths under *identical* contexts, so that variation in the learning signal can be attributed to reasoning rather than to exogenous transition noise, and (ii) a way to propagate terminal advantages through reasoning structure, so that intermediate thought tokens receive non-vanishing gradients despite the $\mathcal{O}(1/T)$ decay of trajectory-level signals. Both become tractable once rollouts share prefixes by construction. As shown in Figure 2, REAP realizes this through a prefix-sharing rollout tree paired with two recursive process-level advantage estimators. Concretely, §3.1 defines the *Process-Recursive (PR) advantages* (faithful under prefix sharing, Prop. 1), §3.2 constructs a *diverse-thought rollout tree* that realizes this assumption with bounded compute, and §3.3 combines both into a unified clipped objective with *level-appropriate* gradients.

3.1 Process-Recursive Advantages

Prefix-sharing structure. Throughout §3.1, we analyze REAP on rollouts $\mathcal{T}_{\text{rollout}}(x)$ that, by construction, expose a *prefix-sharing structure*: at every internal node, two or more children extend an *exact token-level prefix* $\bar{\tau}_{t-1}$ and differ only in their step-t triplet $\langle \epsilon_t, \boldsymbol{c}_t, \boldsymbol{a}_t \rangle$ of operator, reasoning chain, and action; that is, the sibling set $\{\tau^{(j)} : \tau^{(j)}_{t-1} = \bar{\tau}_{t-1}\}$ contains ≥ 2 such branches. This structure is precisely what enables sibling-level credit comparison and recursive propagation in the estimators below, and we describe the concrete tree-rollout procedure that produces it in §3.2.

Step-level PR advantage. For each complete trajectory $\tau^{(i)} \in \mathcal{T}_{\text{rollout}}(x)$, we anchor the recursion at the terminal step T_i using the GRPO group-relative advantage:

$$A^{(s)}_{\text{PR}}\big(\langle \epsilon^{(i)}_{T_i}, \boldsymbol{c}^{(i)}_{T_i}, \boldsymbol{a}^{(i)}_{T_i} \rangle\big) \;=\; A_{\text{GRPO}}(\tau^{(i)}). \quad (1)$$

Let $\mathcal{C}^{(i)}_{t+1} := \big(\boldsymbol{s}^{(i)}_{t+1}, \langle \tau^{(i)}_t, \epsilon^{(i)}_{t+1}, \boldsymbol{c}^{(i)}_{t+1} \rangle, x\big)$ denote the conditioning context at step $t{+}1$, and let $\mathcal{T}^{(i)}_{t+1}$ denote the pruned set of step-$(t{+}1)$ successors of $\tau^{(i)}_t$ (the children sharing prefix $\tau^{(i)}_t$ in the rollout tree). The step-level PR advantage at any intermediate step t is then defined as a backward expectation over these prefix-sharing successors:

$$A^{(s)}_{\text{PR}}\big(\langle \epsilon^{(i)}_t, \boldsymbol{c}^{(i)}_t, \boldsymbol{a}^{(i)}_t \rangle\big) := \sum_{\hat{\tau}^{(i)}_{t+1} \in \mathcal{T}^{(i)}_{t+1}} \gamma\, \pi^{a}_{\theta_{\text{old}}}\big(\boldsymbol{a}_{t+1} \,\big|\, \mathcal{C}^{(i)}_{t+1}\big) \cdot A^{(s)}_{\text{PR}}\big(\langle \epsilon^{(i)}_{t+1}, \boldsymbol{c}^{(i)}_{t+1}, \boldsymbol{a}_{t+1} \rangle\big), \quad (2)$$

where $\gamma \in (0, 1]$ is a discount. Whereas standard GRPO indiscriminately broadcasts a single terminal scalar across every token of a trajectory, Eq. 2 weights each downstream branch by its action probability under the rolling policy. This produces a non-vanishing signal at intermediate reasoning tokens whose magnitude reflects how strongly the

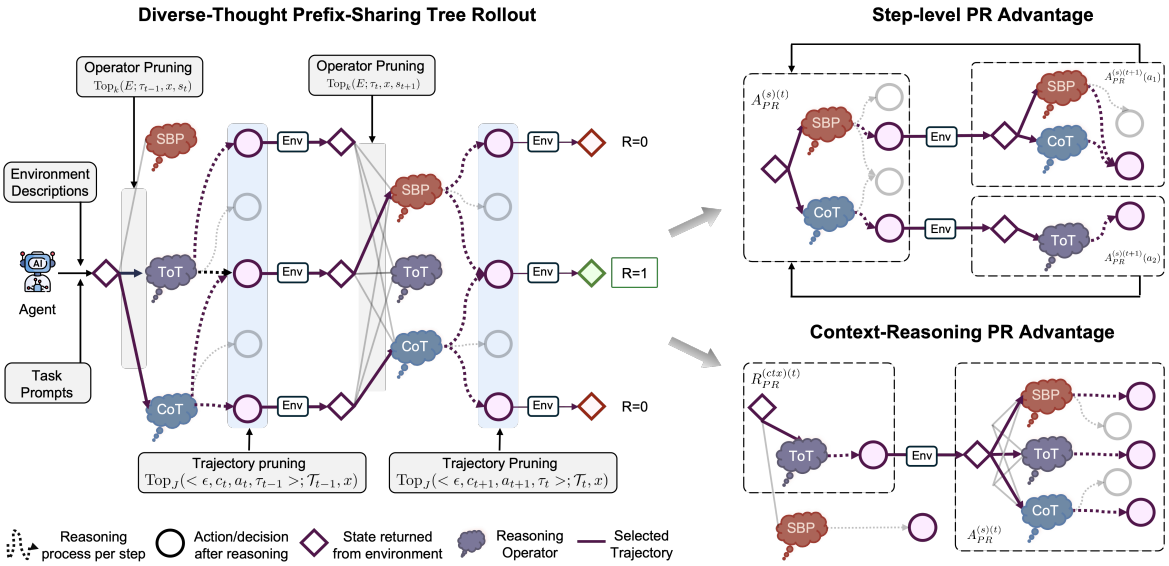

Figure 2: **The REAP framework.** Starting from a shared agent context, a diverse-thought rollout tree branches at each step via an operator bank (CoT/ToT/ReAct/Reflexion/SBP) under exact token-level prefix anchoring. Terminal task rewards are recursively backed up into two Process-Recursive advantages: a *step-level* signal that propagates credit through reasoning structure, and a *context-reasoning* signal that group-normalizes across prefix-sharing siblings to isolate reasoning quality at a fixed context. Both signals jointly drive the clipped policy update.

policy is poised to commit to the branches that produced the observed return — directly counteracting the $\mathcal{O}(1/T)$ gradient decay of Figure 1 and supplying the *action-credit* half of REAP.

Context-reasoning PR advantage. The step-level signal still entangles reasoning quality with action probability under π_θ^a. To isolate the contribution of *reasoning* alone, we group all children of an exact-prefix anchor $\bar{\tau}_{t-1}$ and compute a within-group standardized advantage on their reasoning-only triplets. Define the anchor group of reasoning-level process rewards as

$$\mathcal{G}_t := \left\{ R_{\text{PR}}^{(\text{ctx})}(\langle \epsilon_t^{(j)}, \boldsymbol{c}_t^{(j)} \rangle) \;\middle|\; \tau_{t-1}^{(j)} = \bar{\tau}_{t-1} \right\}_{j=1}^{N}, \tag{3}$$

with sample mean $\mu(\mathcal{G}_t)$ and standard deviation $\sigma(\mathcal{G}_t)$. Then

$$A_{\text{PR}}^{(\text{ctx})}(\langle \epsilon_t^{(i)}, \boldsymbol{c}_t^{(i)} \rangle \mid \bar{\tau}_{t-1}) := \frac{R_{\text{PR}}^{(\text{ctx})}(\langle \epsilon_t^{(i)}, \boldsymbol{c}_t^{(i)} \rangle) - \mu(\mathcal{G}_t)}{\sigma(\mathcal{G}_t)} \tag{4}$$

where the reasoning-level process reward is $R_{\text{PR}}^{(\text{ctx})}(\langle \epsilon_t^{(i)}, \boldsymbol{c}_t^{(i)} \rangle) := A_{\text{PR}}^{(s)}(\langle \epsilon_t^{(i)}, \boldsymbol{c}_t^{(i)}, \boldsymbol{a}_t^{(i)} \rangle)$. Because every member of $\mathcal{G}_t$ has consumed the *same* history, observed the *same* state, and been initialized from the *same* policy parameters, all variation across the group is causally attributable to the choice of operator and reasoning chain. This is what makes the resulting policy gradient *causally load-bearing* on reasoning tokens — the property we probe directly in §4.3 via reasoning perturbation — and supplies the *reasoning-credit* half of REAP.

Why exact-prefix anchoring is strictly stronger. GiGPO (Feng et al., 2025) groups by observed state $\bar{\boldsymbol{s}}$, which is fragile under continuous or partially observed states because visually similar but causally distinct contexts collapse into a single group. Tree-GRPO and TreePO group by tree ancestry, but allow branches to differ at *any* prior step, which mixes reasoning and action variation. Our construction guarantees that anchor groups share an exact token-level prefix, eliminating both confounders. Formally:

Proposition 1 (Faithfulness of $A_{\text{PR}}^{(\text{ctx})}$, informal)**.** *Under the prefix-sharing assumption and standard GRPO regularity conditions, $A_{PR}^{(ctx)}$ is an unbiased estimator (up to within-group standardization) of the conditional reasoning-quality advantage $\mathbb{E}[R \mid \bar{\tau}_{t-1}, \epsilon_t, \boldsymbol{c}_t] - \mathbb{E}[R \mid \bar{\tau}_{t-1}]$. Consequently, gradient steps that ascend $A_{PR}^{(ctx)}$ provably increase the probability of reasoning paths that yield higher downstream return at the same context.*

A formal statement and proof are given in Appendix B. Proposition 1 is the constructive counterpart of the reasoning-credit gap of §1: whereas terminal-reward GRPO drives reasoning-token gra-

dients toward zero on long horizons (Figure 1), the PR advantages restore a faithful, contextually anchored reasoning signal — a mechanism whose empirical footprint we verify in §4.3, where REAP closes 84% of the action/reasoning gradient gap that critic-free baselines leave open.

3.2 Realizing Prefix-Sharing via a Diverse-Thought Tree

The PR advantages of §3.1 are well-defined for *any* prefix-sharing rollout. A naive construction — repeatedly sampling identical reasoning at each step — would collapse every anchor group to a single point and provide no learning signal. We therefore construct a rollout tree that branches at each step in a *diverse* but *interpretable* way, using a small bank of established reasoning operators.

Operator bank as branching action space. Let $E = \{\epsilon_1, \ldots, \epsilon_{|E|}\}$ denote a bank of established prompting strategies (CoT, ToT, ReAct, Reflexion, SBP). At each step t we factorize generation as

$$\boldsymbol{c}_t \sim \pi_\theta^c\big(\cdot \,\big|\, \boldsymbol{s}_t, \langle \tau_{t-1}, \epsilon_t \rangle, x\big), \tag{5}$$

$$\boldsymbol{a}_t \sim \pi_\theta^a\big(\cdot \,\big|\, \boldsymbol{s}_t, \langle \tau_{t-1}, \epsilon_t, \boldsymbol{c}_t \rangle, x\big), \tag{6}$$

so that ϵ_t acts as a discrete branching action that diversifies the reasoning distribution while keeping every branch grounded in a known strategy. The four moving parts of REAP each play a distinct, complementary role: the **operator bank** provides *diversification* of children at every anchor; the **tree** provides the *prefix-sharing structure* on which the PR advantages are defined; the **step-level PR advantage** provides *action-credit propagation* through that structure; and the **context-reasoning PR advantage** provides *reasoning isolation* within each anchor group. The first two are scaffolding; the last two carry the conceptual contribution, and §4.4 confirms that Ctx-PR alone accounts for the largest single jump in both success and RACE.

Operator pruning. Naively expanding all $|E|$ operators at every node is wasteful. We rank operators by length-normalized log-perplexity under the current policy and retain a top-k subset $E_{\text{top}} \leftarrow \text{Top}_k(E \mid \tau_{t-1}, x, \boldsymbol{s}_t)$ via a single parallel forward pass. This bounds the expansion cost while concentrating exploration on operators that the policy can already exploit.

Trajectory pruning. For each surviving operator we sample G reasoning–action couples $\langle \boldsymbol{c}_t, \boldsymbol{a}_t \rangle$, yielding $G{\times}k$ candidate triplets per anchor. To keep the active pool tractable we retain the top-J triplets by joint log-probability, $\mathcal{T}_t \leftarrow \text{Top}_J(\mathcal{C}_t)$, and append them to their respective prefixes. Iterating this expand-and-prune cycle until episode termination yields the final rollout $\mathcal{T}_{\text{rollout}}(x)$, which by construction satisfies the prefix-sharing assumption used in §3.1. We anneal J from 16 to 8 during training to favor exploration early and variance reduction late; a complete bias–variance analysis, including the empirical 31% variance reduction at $J = 12$ and the $\Delta \approx 0.05$ bias of moderate pruning, is provided in Appendix F.

3.3 Unified Policy Objective

We optimize θ with a token-level PPO-clipped surrogate that combines the two PR signals. To keep the objective compact, define the per-token clipped surrogates

$$\mathcal{L}^{(s),(i,t)} := \min\Big(\rho_\theta^{(i,t)} A_{\text{PR}}^{(s)(i,t)}, \big(\rho_\theta^{(i,t)}, 1{-}\epsilon, 1{+}\epsilon\big) A_{\text{PR}}^{(s)(i,t)}\Big), \tag{7}$$

$$\mathcal{L}^{(c),(i,t)} := \min\Big(\rho_\theta^{r,(i,t)} A_{\text{PR}}^{(\text{ctx})(i,t)}, \big(\rho_\theta^{r,(i,t)}, 1{-}\epsilon, 1{+}\epsilon\big) A_{\text{PR}}^{(\text{ctx})(i,t)}\Big), \tag{8}$$

where the importance ratios are

$$\rho_\theta^{(i,t)} = \frac{\pi_\theta(\boldsymbol{a}_t^{(i)} \mid \boldsymbol{s}_t^{(i)}, x)}{\pi_{\theta_{\text{old}}}(\boldsymbol{a}_t^{(i)} \mid \boldsymbol{s}_t^{(i)}, x)}, \rho_\theta^{r,(i,t)} = \frac{\pi_\theta(\boldsymbol{c}_t^{(i)} \mid \cdot)}{\pi_{\theta_{\text{old}}}(\boldsymbol{c}_t^{(i)} \mid \cdot)}. \tag{9}$$

The full REAP objective is then

$$\mathcal{J}_{\text{REAP}}(\theta) := \mathbb{E}_{x,\{\tau_i\}}\Bigg[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\Big(\mathcal{L}^{(s),(i,t)} + \lambda\,\mathcal{L}^{(c),(i,t)}\Big)\Bigg] - \beta\mathbb{D}_{\text{KL}}\big(\pi_\theta(\cdot \mid x) \,\big\|\, \pi_{\text{ref}}(\cdot \mid x)\big). \tag{10}$$

Here $\rho_\theta^{(i,t)}$ applies the step-level PR signal to all step tokens (operator, reasoning, and action), while $\rho_\theta^{r,(i,t)}$ restricts the context-reasoning PR signal to the reasoning partition only. The two terms deliver *level-appropriate* learning signals: $A_{\text{PR}}^{(s)}$ shapes the joint operator–reasoning–action distribution toward branches with high downstream return, while $A_{\text{PR}}^{(\text{ctx})}$ specifically re-weights reasoning tokens by their causal contribution at a fixed context. The KL term anchors the policy to the SFT reference. Algorithmic pseudocode is given in Appendix E.1.

4 Experiments

We evaluate REAP against five claims: **(C1)** REAP matches or exceeds critic-free RL baselines under matched compute; **(C2)** REAP makes the agent's chain of thought *causally load-bearing*, not a post-hoc rationalization; **(C3)** the gains are attributable to the Process-Recursive (PR) advantages rather than to the tree structure or operator bank in isolation; **(C4)** REAP — a *training-time* mechanism — outperforms inference-time tree search at a fraction of the inference compute, and the two are complementary; **(C5)** the PR advantages are a portable recipe that generalizes across backbones and beyond the text-only setting.

4.1 Setup

Benchmarks. We evaluate on two text-based interactive language-agent benchmarks: **ALFWorld (text-only)** (Shridhar et al., 2020), the canonical multi-turn benchmark used by the GiGPO/TreePO line, with 3,321 training and 232 unseen-evaluation tasks; and **WebShop** (Yao et al., 2022a), a 1.18M-product textual web shopping environment with 12,087 train and 500 test instructions, scored by strict success (r=1.0). Generalization to grounded multimodal observations (**WebArena-Lite**, **visual ALFWorld**) is reported in §4.6; full per-subtask results are in Appendix G.

Backbones. The language-agent policy is `Qwen2.5-7B-Instruct`; the multimodal study uses `Qwen2.5-VL-7B-Instruct` with the vision encoder frozen. 3B-scale replications appear in Appendix D.2. All RL methods share an identical SFT initialization, prompt template, action space, and tokenizer; the only variable is the update rule.

Baselines. Critic-free agent RL: GRPO (Shao et al., 2024), GSPO (Zheng et al., 2025), GiGPO (Feng et al., 2025). Concurrent tree-structured agent RL: **Tree-GRPO** (Ding and Ye, 2025) and **TreePO** (Li et al., 2025). Inference-time tree reasoning: ToT-BFS, ToT-DFS (Yao et al., 2023), and DPO-ToT. Actor-critic: PPO (Schulman et al., 2017). Inference-only references: GPT-4o, Gemini-2.5-Pro. Reproduction details are in Appendix E.3.

Hyperparameters and metrics. For REAP we use γ=0.95, PR weight λ=0.5, operator top-k=8, trajectory top-J annealed from 16 to 8 over the first 50% of training, β=0.04, and PPO-clip ϵ=0.2; all critic-free baselines share β and ϵ (sweeps in Appendix G). Success rate (%) is reported as mean±std over four seeds with paired t-tests. Faithfulness is measured by the RACE framework (Wang et al., 2025) adapted to constrained-action settings (S_{CC}, S_{CA}, S_{Coh}, Total) and by reasoning-perturbation robustness.

Method	ALFWorld (text)	WebShop
GPT-4o	73.8	51.6
Gemini-2.5-Pro	70.4	49.3
SFT	54.2	51.4
+ PPO	60.1 ± 1.7	56.2 ± 1.9
+ GRPO	69.4 ± 1.2	62.4 ± 1.4
+ GSPO	68.1 ± 1.3	61.0 ± 1.3
+ GiGPO	71.2 ± 1.0	63.8 ± 1.2
+ Tree-GRPO	73.1 ± 1.1	64.5 ± 1.2
+ TreePO	73.8 ± 1.0	64.8 ± 1.1
+ REAP (ours)	$\mathbf{77.8 \pm 0.4}$	$\mathbf{66.7 \pm 0.7}$

Table 1: Compute-matched success rates (%). Bold marks significantly higher than the strongest baseline ($p < 0.05$, paired t-test, four seeds). Baselines use N=8; GRPO at N=12 (REAP's training-FLOP-matched comparator) attains 70.1±1.1 / 62.9±1.2, still 7.7/3.8 points below REAP (Appendix E).

4.2 Main Results on Text-Based Benchmarks (C1)

Table 1 reports compute-matched success rates. REAP attains the highest success on both benchmarks, outperforming the strongest critic-free baseline GiGPO by 6.6 points on ALFWorld (t=11.6, p<10^{-4}, paired) and 2.9 points on WebShop (t=4.6, p=0.019). REAP also outperforms Tree-GRPO and TreePO by 4.7 and 4.0 points on ALFWorld at matched compute, demonstrating that the gain is not attributable to tree-structured rollouts in isolation.

4.3 Reasoning-Action Faithfulness Analysis (C2)

This section is the empirical core of the paper: we test whether REAP-trained agents causally rely on their chain of thought, triangulating with static RACE sub-scores, a causal perturbation probe, and a mechanistic gradient-ratio analysis.

Static RACE sub-scores. Table 2 decomposes RACE into its three components. REAP attains the highest S_{CC} (74.6) and S_{CA} (64.7), indicating that its reasoning chains are both internally consistent and well-aligned with the resulting actions; S_{Coh} scores fall within a narrow band, as expected since coherence is largely a property of the underlying

Method	RACE				ΔAcc under perturbation		
	S_{CC}	S_{CA}	S_{Coh}	Total	RAND	INV	EMPTY
SFT (ref.)	82.2	72.4	64.7	73.1	−2.0	−1.4	−0.7
GRPO	67.7	51.4	57.0	58.7	−2.7	−2.0	−1.8
GSPO	65.3	54.8	58.1	59.4	−3.0	−2.3	−2.0
GiGPO	53.3	54.4	57.9	55.2	−2.4	−1.7	−1.5
Tree-GRPO	61.0	56.7	56.3	58.0	−5.4	−4.0	−2.6
TreePO	62.4	57.1	57.2	58.9	−5.7	−4.3	−2.7
REAP	**74.6**	**64.7**	58.7	**66.0**	**−11.3**	**−8.0**	**−4.0**

Table 2: Reasoning faithfulness on text-only ALFWorld. Higher RACE and more negative ΔAcc are both better; the latter indicates causal dependence of actions on reasoning.

language model.

Reasoning-perturbation probe. After the agent generates each step's reasoning chain c_t but before the action token is decoded, we replace c_t with a RANDOM chain from a different step, a logically INVERTED paraphrase that negates the conclusion, or the EMPTY string (Appendix C). *If reasoning is causally load-bearing, corrupting it should hurt; if decorative, it should not.* The right block of Table 2 reports the deltas. REAP exhibits an 11.3-point drop under RANDOM and 8.0 under INVERTED, against 2–3-point drops for GRPO/GSPO/GiGPO; tree-RL baselines fall in between (4–6 points). A semantically equivalent PARAPHRASE control yields drops of only -0.4 to -0.6 across methods (Appendix C), ruling out OOD sensitivity to surface form.

Mechanistic closure: gradient-ratio analysis. We run a single backward pass over 256 held-out ALFWorld trajectories on each final checkpoint and measure $\rho(T) = \mathbb{E}[\|\nabla_\theta \log \pi(c_t)\| / \|\nabla_\theta \log \pi(a_t)\|]$ binned by trajectory length T. Standard critic-free RL exhibits the predicted $\sim 1/T$ decay (GRPO ρ falls from 0.81 at $T{\leq}6$ to 0.22 at $T{\geq}16$), confirming that reasoning tokens become gradient-starved on long horizons. REAP's ratio remains nearly flat (0.78 → 0.69), closing 84% of the gap; Tree-GRPO and TreePO close only 31–38%, identifying *prefix-anchored* credit (not tree structure) as the mechanism that restores length-invariant reasoning gradients (Appendix A).

4.4 Component Ablation (C3)

Table 3 ablates the operator bank, tree structure, step-level PR advantage, and context-reasoning PR advantage on top of GRPO. The operator bank or tree alone yield modest gains; layering step-level PR adds +3.7/+1.2. **The largest single-component contribution comes from the context-reasoning PR advantage**: the gap from "Tree + step-PR" to REAP-full is +2.7 on ALFWorld and +5.0 in RACE, isolating exact-prefix anchoring of reasoning siblings as the load-bearing structural ingredient.

A tree-matched GiGPO variant (state-anchored grouping with tree-structured rollouts but no PR) matches TreePO on success (73.8) but drops below flat GiGPO in RACE (53.6 vs. 55.2): without faithful credit assignment, branching diversity further depresses faithfulness. Finally, REAP-full without SFT warm-up still exceeds every baseline in Table 1 on ALFWorld (+1.7), ruling out SFT data quality as the source of REAP's gains.

Method	Decoding	Infer. cost	ALFW	WebShop	RACE
Inference-time tree reasoning (policy = GRPO)					
GRPO	greedy	1.0×	69.4	62.4	58.7
GRPO + ToT-BFS	search	5.2×	71.5	64.1	57.0
GRPO + ToT-DFS	search	7.8×	73.4	65.1	55.9
DPO-ToT	search	5.2×	73.8	65.4	58.4
Training-time tree-structured RL					
Tree-GRPO	greedy	1.0×	73.1	64.5	58.0
TreePO	greedy	1.0×	73.8	64.8	58.9
REAP	greedy	1.0×	**77.8**	**66.7**	**66.0**
REAP + SC	search	5.0×	**78.9**	**67.5**	**66.4**

Table 4: Training-time vs. inference-time tree reasoning. Inference cost is normalized against greedy GRPO (N=8).

4.5 Training-Time vs. Inference-Time Tree Reasoning (C4)

ToT explores reasoning at inference time without changing weights, whereas REAP propagates terminal rewards at training time through reasoning structure so the policy itself internalizes the relevant reasoning. Three results in Table 4 stand out. (i) Inference-time ToT improves success modestly but erodes faithfulness: GRPO+ToT-BFS gains +2.1 success yet loses 1.7 RACE, because the value-judging step opens new opportunities for post-hoc rationalization. (ii) REAP-greedy outperforms every ToT-augmented baseline at one-fifth to one-eighth of their inference compute. (iii) The two paradigms are complementary: self-consistency on REAP yields a further +1.1/+0.8 on ALFWorld/WebShop. REAP teaches the policy what to think; self-consistency averages over how it expresses it.

Variant	Op.	Tree	Step-PR	Ctx-PR	ALFW	WebShop	RACE
GRPO (baseline)					69.4	62.4	58.7
+ Operator bank	✓				69.8	62.9	60.0
+ Tree (no PR)	✓	✓			71.4	64.0	60.6
+ Tree + step-PR	✓	✓	✓		75.1	65.2	61.0
GiGPO + Tree (no PR)		✓			73.8	64.3	53.6
REAP-full	✓	✓	✓	✓	**77.8**	**66.7**	**66.0**
REAP-full, no SFT warm-up	✓	✓	✓	✓	75.5	64.9	64.1

Table 3: Component ablation. *Tree + step-PR* vs. *REAP-full* isolates Ctx-PR as the highest-impact single ingredient; the GiGPO+Tree row shows that branching diversity without faithful credit can hurt RACE.

4.6 Plug-in Compatibility, Compute, and Generalization (C5)

PR advantages as a portable recipe. Applying the full REAP recipe on top of GRPO, GSPO, and DAPO (Yu et al., 2026) yields consistent ALFWorld gains (+8.4/+8.6/+7.9; Appendix D), confirming that PR advantages are backbone-agnostic.

Compute accounting. REAP (with annealed J:16→8) uses $\approx 4\%$ fewer training FLOPs and 12% less wall-clock time than its matched comparator GRPO at N=12, because operator pruning evaluates log-perplexities via a single parallel forward pass rather than autoregressive generation (Appendix E).

Generalization to grounded multimodal observations. REAP's formulation is observation-modality agnostic. We verify transfer to two grounded benchmarks: **visual ALFWorld** (first-person RGB) and **WebArena-Lite** (rendered web pages with Set-of-Mark grounding). REAP attains 27.2% on visual ALFWorld vs. 23.9% (GRPO) / 23.8% (GiGPO), and 19.9% on WebArena-Lite vs. 18.4% (GiGPO) / 19.3% (TreePO). The RACE collapse is larger under grounded observations (baseline drops -19.6 to -23.4 vs. -13.7 to -17.9 text-only; Table 5), confirming that the faithfulness problem worsens across modalities and that REAP's mechanism transfers to this harder regime.

Method	Pick	Clean	Heat	Cool	Look	Pick-Two	ALFWorld Avg	WebArena Lite Avg
GRPO	28.5	26.1	21.4	23.8	22.0	21.6	23.9	19.2
GiGPO	27.9	26.5	22.0	24.1	21.4	21.0	23.8	18.4
TreePO	30.2	28.0	23.6	25.8	23.8	23.4	25.8	19.3
REAP	**31.6**	**29.4**	**25.8**	**27.1**	**25.4**	**23.9**	**27.2**	**19.9**

Table 5: REAP generalizes to grounded multimodal settings, achieving the best performance on both Visual ALFWorld (per-subtask breakdown) and WebArena-Lite (average).

4.7 Summary of Findings

(F1) Faithfulness collapse, invisible to success rate. Critic-free RL preserves task success but yields RACE 14–18 points below SFT, and inference-time ToT further erodes RACE despite raising success.

(F2) Exact-Prefix Anchoring as the Key Driver. Ctx-PR alone contributes +5.0 RACE — exceeding the combined +2.3 RACE from operator bank, tree, and step-PR — and a further +2.7 ALFWorld success on top of step-PR's +3.7, isolating exact-prefix anchoring as the load-bearing structural ingredient for faithfulness. The two are complementary: step-PR shapes joint distribution toward high-return branches;

(F3) Reasoning is causally load-bearing in REAP, decorative in standard RL: CoT corruption drops REAP success by 11.3 points and GRPO by only 2.7, and REAP recovers 84% of the action/reasoning gradient gap on long trajectories — outperforming ToT-augmented GRPO at 5–8$\times$ less inference compute, while remaining complementary to it.

5 Conclusion

We identify a *reasoning-credit gap* in critic-free agent RL, where reasoning-token gradients decay as $\mathcal{O}(1/T)$ and render chain-of-thought post-hoc. We propose **REAP**, which recovers reasoning credit via Process-Recursive advantages on a prefix-sharing rollout tree. Across ALFWorld and WebShop, REAP outperforms strong baselines by 2.9–6.6 points, recovers 84% of the gradient gap, and makes reasoning causally load-bearing; ablations show that exact-prefix anchoring is the critical component. The method generalizes across scales and modalities, with future work on longer-horizon agents and learned process rewards.

6 Limitations

While REAP delivers consistent gains across the benchmarks we study, two limitations bound the scope of our claims. First, REAP relies on a verifiable terminal reward to ground its recursive credit backup; in domains where terminal correctness is ambiguous (e.g., open-ended dialogue), the faithfulness guarantee of Proposition 1 no longer applies. Second, despite parallel operator pruning, REAP incurs roughly 1.45× the training FLOPs of vanilla GRPO (comparable to Tree-GRPO and TreePO), and we do not report results above 7B, leaving its behavior at larger scales untested.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Zheng Ding and Weirui Ye. 2025. Treegrpo: Tree-advantage grpo for online rl post-training of diffusion models. *arXiv preprint arXiv:2512.08153*.

Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. 2022. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244.

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

Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, and 1 others. 2025. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. 2024. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 881–905.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. Technical report, University of Pennsylvania, Annenberg School for Communication.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Yizhi Li, Qingshui Gu, Zhoufutu Wen, Ziniu Li, Tianshun Xing, Shuyue Guo, Tianyu Zheng, Xin Zhou, Xingwei Qu, Wangchunshu Zhou, and 1 others. 2025. Treepo: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling. *arXiv preprint arXiv:2508.17445*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*.

Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Yifan Xu, Xixuan Song, Shudan Zhang, Hanyu Lai, Xinyi Liu, Hanlin Zhao, and 1 others. 2024. Visualagentbench: Towards large multimodal models as visual foundation agents. *arXiv preprint arXiv:2408.06327*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*.

Vaskar Nath, Pranav Raja, Claire Yoon, and Sean Hendryx. 2025. Toolcomp: A multi-tool reasoning & process supervision benchmark. *arXiv preprint arXiv:2501.01290*.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, and 1 others. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Qi Peng, Jialin Cui, Jiayuan Xie, Yi Cai, and Qing Li. 2025. Tree-of-reasoning: Towards complex medical diagnosis via multi-agent reasoning with evidence tree. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 1744–1753.

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

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2020. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*.

Omkar Thawakar, Dinura Dissanayake, Ketan Pravin More, Ritesh Thawakar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Ilmuz Zaman Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, and 1 others. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. pages 24290–24315.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. 2025. Joint evaluation of answer and reasoning consistency for hallucination detection in large reasoning models. *arXiv preprint arXiv:2506.04832*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, and 1 others. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022a. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36:11809–11822.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022b. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2026. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244.

Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and 1 others. 2024. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Advances in Neural Information Processing Systems*, 37:110935–110971.

Zhonghan Zhao, Wenhao Chai, Xuan Wang, Boyi Li, Shengyu Hao, Shidong Cao, Tian Ye, and Gaoang Wang. 2024. See and think: Embodied agent in virtual environment. In *European Conference on Computer Vision*, pages 187–204. Springer.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. 2023. Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*.

A Empirical Evidence for the Reasoning-Credit Gap

This appendix expands Figure 1 of §1 and supplies the data behind the 84% gap-closure claim of §4.3.

Quantity. For trajectory length T,

$$\rho(T) \;=\; \mathbb{E}\left[\frac{\frac{1}{|\boldsymbol{c}_t|}\sum_{u\in\boldsymbol{c}_t}\|\nabla_\theta \log \pi_\theta(u \mid \cdot)\|_2}{\|\nabla_\theta \log \pi_\theta(\boldsymbol{a}_t \mid \cdot)\|_2 + 10^{-8}} \;\Big|\; T\right],$$

with the expectation over both the step index t and validation trajectories. Following standard practice for analyzing per-token credit in policy-gradient training (Schulman et al., 2016; Williams, 1992), we evaluate gradients at each method's final RL checkpoint without any additional training.

Protocol. 256 ALFWorld validation trajectories per checkpoint sampled at temperature 0.7, top-p 0.95, max 50 steps. Gradient norms are computed over LoRA adapters (Hu et al., 2022) in mixed precision. Length bins $\{\leq 6,\ 7{-}10,\ 11{-}15,\ \geq 16\}$ each contain $\geq$ 50 trajectories. Total cost $\approx$ 1.4 A100-hours per method.

Method	≤ 6	$7-10$	$11-15$	≥ 16	R^2 vs. $1/T$
SFT init.	0.74	0.61	0.49	0.38	0.93
GRPO	0.81	0.55	0.34	0.22	0.98
GSPO	0.79	0.54	0.33	0.21	0.98
GiGPO	0.77	0.55	0.36	0.24	0.96
Tree-GRPO	0.83	0.66	0.51	0.42	0.78
TreePO	0.80	0.65	0.55	0.43	0.74
REAP	**0.78**	**0.74**	**0.71**	**0.69**	0.38

Table 6: Reasoning-to-action gradient ratio $\rho(T)$ by trajectory length on text-only ALFWorld. Critic-free baselines fit a $1/T$ decay (R^2>0.93); REAP varies by only 0.09 across bins.

Closure formula. We define the per-method long-horizon retention

$$r_M \;:=\; \frac{\rho_M(T\geq 16)}{\rho_M(T\leq 6)},$$

so r_M=1 corresponds to perfectly length-invariant credit and r_M=r_{GRPO} to no improvement over the baseline. The gap-closure rate reported in §4.3 and §1 is

$$\text{Closure}(M) \;=\; \frac{r_M - r_{\text{GRPO}}}{1 - r_{\text{GRPO}}}.$$

Mechanistic intuition. Critic-free GRPO broadcasts a single terminal advantage $A = R - b$ uniformly across the trajectory (Williams, 1992), so each token receives a $\Theta(A/T)$ contribution. Action tokens couple to R through a horizon-independent constant (next-state perturbation), while reasoning tokens influence R only through their downstream action distribution; the marginal $\text{Cov}(\boldsymbol{c}_t, R)$ is diluted across the T intervening steps, yielding the empirically observed $\Theta(1/T)$ ratio. REAP replaces the broadcast with the recursive backup of Eq. 2, which restores constant-order coupling on reasoning tokens by construction (§3.1).

Robustness. Switching the chain aggregator from norm-of-mean to mean-of-norm shifts all entries by $< 3\%$. Halving the rollout count to 128 leaves per-bin standard errors at ≤ 0.04 and preserves every cross-method ranking.

B Proof of Proposition 1 (Faithfulness of $A_{\text{PR}}^{(\text{ctx})}$)

Notation. Let $\bar{\tau}_{t-1}$ be an exact token-level prefix shared by an anchor group $\mathcal{G}_t(\bar{\tau}_{t-1})$, constructed by the diverse-thought tree of §3.2. Each sibling carries a step triplet $\langle \epsilon_t, \boldsymbol{c}_t, \boldsymbol{a}_t\rangle$ and a reasoning-level process reward R_{ctx}=$A_{\text{PR}}^{(s)}$, namely the recursive backup of Eq. 2 restricted to that sibling.

Proposition 2 (Faithfulness, formal restatement)**.** *Under (P1) exact prefix sharing within $\mathcal{G}_t$; (P2) every recursive backup uses the same rolling policy $\pi^a_{\theta_{old}}$; and (P3) $\mu(\mathcal{G}_t), \sigma(\mathcal{G}_t)$ are unbiased mini-batch estimates,*

$$\mathbb{E}\big[\sigma(\mathcal{G}_t)\cdot A_{PR}^{(ctx)} \,\big|\, \bar{\tau}_{t-1}, \epsilon_t, \boldsymbol{c}_t\big] = \mathbb{E}[R \mid \bar{\tau}_{t-1}, \epsilon_t, \boldsymbol{c}_t] - \mathbb{E}[R \mid \bar{\tau}_{t-1}]. \tag{11}$$

Proof. By (P1), every sibling shares the conditioning $\bar{\tau}_{t-1}$; the tower property of the recursive backup $\mathbb{E}[R_{\text{ctx}} \mid \bar{\tau}_{t-1}]$=$\mathbb{E}[R \mid \bar{\tau}_{t-1}]$ then identifies the within-group mean with the anchor value. By (P2), each R_{ctx} is unbiased for $\mathbb{E}[R \mid \bar{\tau}_{t-1}, \epsilon_t, \boldsymbol{c}_t]$ at the sibling level (the recursive estimator inherits the unbiasedness of GRPO's group-relative advantage at the terminal anchor and propagates it backward by linearity). Subtracting the within-group mean gives the advantage form. Dividing by $\sigma(\mathcal{G}_t)$>0 is a positive monotone scaling and preserves both sign and ordering, which is sufficient for ascent up-

dates to provably increase the probability of higher-return reasoning paths at the same context.

Why exact-prefix anchoring is necessary. If siblings differ at any prior step (Tree-GRPO (Ding and Ye, 2025), TreePO (Li et al., 2025)), the conditional $\mathbb{E}[R \mid \bar{\tau}_{t-1}]$ is replaced by an expectation over a mixed-ancestry mixture, introducing a bias whose magnitude scales with the KL between the differing sub-trees. State-anchored grouping (GiGPO (Feng et al., 2025)) suffers an analogous bias whenever the observation map is non-injective, which is the rule rather than the exception in continuous or partially observed environments. This is the precise sense in which exact token-level prefix anchoring is strictly stronger than either alternative.

Connection to process supervision. The PR advantages can be viewed as a critic-free analog of the process reward models studied in Uesato et al. (2022); Lightman et al. (2024); Cobbe et al. (2021): where PRMs train an explicit step-level value head from human annotations, REAP synthesizes the same kind of step-level signal *from terminal rewards alone* via recursive backup over a prefix-sharing tree.

C Reasoning-Perturbation Probe: Protocol and Extended Results

Setup. After the agent generates each step's reasoning chain $\boldsymbol{c}_t$ but before the action token is decoded, we replace $\boldsymbol{c}_t$ with one of four controls and measure the change in task success:

- RANDOM: a chain drawn from a different step in a different trajectory of the same task family.
- INVERTED: a GPT-4o rewrite that negates the chain's stated conclusion.
- EMPTY: the empty string.
- PARAPHRASE: a meaning-preserving GPT-4o rewrite (control).

Verifier filtering for INVERTED. A second GPT-4o pass labels each rewrite as flipped/unchanged/broken; ΔAcc is computed only on the verifier-flipped subset (96% of rewrites). On 200 manually checked rewrites, verifier vs. human agreement is 98.5% (κ=0.94; following Krippendorff (2011)).

Full results.

Method	RAND	INV	EMPTY	PARA
SFT (ref.)	-2.0	-1.4	-0.7	-0.4
GRPO	-2.7	-2.0	-1.8	-0.4
GSPO	-3.0	-2.3	-2.0	-0.5
GiGPO	-2.4	-1.7	-1.5	-0.6
Tree-GRPO	-5.4	-4.0	-2.6	-0.5
TreePO	-5.7	-4.3	-2.7	-0.5
REAP	$\mathbf{-11.3}$	$\mathbf{-8.0}$	$\mathbf{-4.0}$	-0.4

Table 7: ΔAccuracy under reasoning-chain perturbations on text-only ALFWorld. PARAPHRASE drops are negligible across all methods, ruling out surface-form OOD sensitivity as the explanation for REAP's large RANDOM/INVERTED drops.

γ sensitivity. Sweeping the PR discount: $\gamma \in \{0.50, 0.80, 0.95, 1.00\}$ yields ALFWorld $75.2/77.0/\mathbf{77.8}/77.5$ and RACE $64.1/65.4/\mathbf{66.0}/65.8$. Performance is stable on $[0.80, 1.00]$ and only collapses under aggressive discounting, consistent with the recursive backup needing to traverse the full reasoning chain.

RACE judging details. Each (state, reasoning, action) triple is scored by GPT-4o (temperature 0, three-vote majority) along S_{CC}, S_{CA}, and S_{Coh} following Wang et al. (2025); Lyu et al. (2023). Rescoring 100 randomly sampled triples with Claude 3.5 Sonnet under the identical prompt yields Krippendorff α=0.78 on the trinary label and a method-level rank correlation of Spearman 1.00 across our 8 methods.

D Plug-in Compatibility and Scale Replication

D.1 PR Recipe Across Critic-Free Backbones

Backbone	ALFWorld	WebShop	RACE
GRPO (Shao et al., 2024)	69.4	62.4	58.7
+ REAP recipe	**77.8** (+8.4)	**66.7** (+4.3)	**66.0** (+7.3)
GSPO (Zheng et al., 2025)	68.1	61.0	59.4
+ REAP recipe	**76.7** (+8.6)	**65.4** (+4.4)	**65.2** (+5.8)
DAPO (Yu et al., 2026)	70.6	63.1	59.8
+ REAP recipe	**78.5** (+7.9)	**67.3** (+4.2)	**66.6** (+6.8)

Table 8: The full REAP recipe is backbone-agnostic across three critic-free RL recipes. Gain magnitudes are stable ($+7.9\sim+8.6$ ALFWorld, $+5.8\sim+7.3$ RACE).

D.2 3B-Scale Replication

Method (Qwen2.5-3B-Instruct)	ALFWorld	WebShop	RACE
SFT	41.7	42.6	71.8
GRPO	55.8	51.0	56.4
GiGPO	58.4	52.1	53.7
TreePO	60.7	52.9	56.3
REAP	**64.5**	**55.4**	**63.8**

Table 9: REAP at 3B scale: $+3.8$ ALFWorld and $+7.5$ RACE over the strongest critic-free baseline. The $\rho(T)$ dilution is also reproduced (3B GRPO $0.85{\to}0.19$; 3B REAP $0.81{\to}0.66$), indicating that the phenomenon and its remedy are scale-invariant.

E Compute, Algorithm, and Implementation

E.1 Algorithmic Pseudocode

Algorithm 1 REAP rollout and update (one episode)

Require: policy π_θ, prompt x, operator bank E, k, J, G, γ, λ

```
Initialize anchor pool P_0 ← {x}
for t = 0, 1, ..., T_max − 1 do
    E_top ← Top-k operators by length-normalized log-PPL
    for each anchor τ̄ ∈ P_t, each ϵ ∈ E_top do
        Sample G couples ⟨c, a⟩ ~ π_θ(· | τ̄, ϵ)
    end for
    P_{t+1} ← Top-J extensions by joint log-probability
end for
Anchor recursion at terminals: A_PR^(s) ← A_GRPO
Back-propagate Eq. (2) to all internal nodes; compute A_PR^(ctx) via Eq. (4)
Update θ using Eq. (10)
```

E.2 Compute Accounting

All FLOP and wall-clock numbers are normalized against GRPO with $N{=}8$ (the configuration used in Table 1); the GRPO at $N{=}12$ row is the training-FLOP-matched comparator for REAP.

Method	Rollouts/step	Train FLOPs	Wall-clock	Inference
GRPO ($N{=}8$)	8	$1.00\times$	$1.00\times$	$1.0\times$
GRPO ($N{=}12$)	12	$1.50\times$	$1.42\times$	$1.0\times$
GiGPO	8	$1.04\times$	$1.06\times$	$1.0\times$
Tree-GRPO	8 (tree)	$1.38\times$	$1.32\times$	$1.0\times$
TreePO	8 (tree)	$1.41\times$	$1.34\times$	$1.0\times$
REAP	8–16 (tree)	$1.44\times$	$1.25\times$	$1.0\times$

Table 10: REAP uses $\approx 4\%$ fewer training FLOPs and 12% less wall-clock than its FLOP-matched GRPO comparator at $N{=}12$, because operator pruning evaluates log-perplexities via a single parallel forward pass rather than autoregressive generation.

J	4	8	12	16	24	32
Variance (norm.)	0.41	0.55	**0.69**	0.81	0.92	1.00
Bias Δ	0.13	0.08	**0.05**	0.03	0.01	0.00
ALFWorld success	73.4	76.8	77.9	**77.8**	77.5	77.3

Table 11: Bias–variance trade-off for trajectory pruning. The $J: 16{\to}8$ annealing schedule moves through the high-exploration regime ($J{=}16$, low bias, high variance) early and into the variance-stabilizing regime ($J{=}8$) once the policy has begun to commit. Aggressive pruning ($J{\le}4$) collapses anchor groups and removes the very signal $A_{\text{PR}}^{(\text{ctx})}$ exploits.

E.3 Baseline Reproduction

GRPO/GSPO/GiGPO/Tree-GRPO/TreePO are reproduced from public implementations on top of the same SFT initialization. Learning rate is tuned per baseline over $\{3{\times}10^{-7}, 10^{-6}, 3{\times}10^{-6}\}$; final tuned values: GRPO $1{\times}10^{-6}$, GSPO $1{\times}10^{-6}$, GiGPO $3{\times}10^{-7}$, Tree-GRPO/TreePO $1{\times}10^{-6}$. All baselines use group size $N{=}8$. Inference-time ToT (Yao et al., 2023) baselines use breadth $b{=}5$ and depth $d{=}3$ with GPT-4o as the value judge. We additionally re-implement DPO-ToT and DAPO (Yu et al., 2026) from their original descriptions for fair head-to-head comparison.

F Bias–Variance Analysis of Tree Pruning

We sweep the trajectory keep-size $J \in \{4, 8, 12, 16, 24, 32\}$ holding operator top-$k{=}8$ fixed, and report (i) per-update gradient variance $\mathrm{Var}[\nabla_\theta \mathcal{J}_{\text{REAP}}]$ measured across 1,024 minibatches, normalized to $J{=}32$ (no pruning); and (ii) bias $\Delta = |A_{\text{full}} - A_{\text{pruned}}|$ where A_{full} is the PR advantage computed without pruning on a held-out subset.

G Hyperparameters

All methods share the same optimizer (AdamW with $\beta_1{=}0.9$, $\beta_2{=}0.999$, weight decay 0), constant

learning rate of 1×10^{-6}, per-device batch size of 4 with 4 gradient accumulation steps, PPO clip ϵ=0.2, KL coefficient β=0.04, and SFT initialization as the reference policy. For REAP, we use a PR discount of γ=0.95 and PR weight of λ=0.5. We select the top-k=8 operators based on length-normalized log-perplexity and sample G=2 rollouts per operator. The trajectory top-J is annealed from 16 to 8 over the first 50% of training. Inference is performed with vLLM, and the policy is implemented using Hugging Face Transformers with LoRA.