VH-GPO: VERIFIED HANDOFF FOR LLM AGENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) plays a key role in training large language models (LLMs) to solve complex tasks through outcome-based feedback. Group-relative RL methods such as GRPO train language-model agents by comparing the outcomes of multiple rollouts for the same task. This mechanism is effective when outcomes vary, but if every trajectory in a group fails with the same reward, relative advantages collapse toward zero and the group supplies no useful learning signal. When the trainable agent's rollouts fail to produce any successful trajectory, one possible solution is to introduce a frozen teacher policy. We refer to the trainable agent as the student. The teacher attempts recovery from a state reached by the student, and its continuation is retained only if the environment verifies success. Yet allowing the teacher to complete the remaining task makes the entire recovered suffix teacher-generated, turning supervision into off-policy imitation and weakening a key benefit of RL: learning on the state distribution induced by the student. We therefore ask: after a teacher has repaired a student failure, when should control return to the student? Our core insight is that the teacher should act only long enough to cross what the student cannot yet handle, then hand back control so that successful experience is again student-generated. We formalize this idea as *outcome-verified teacher–student handoff* and instantiate it in Verified Handoff for Group Policy Optimization (VH-GPO). Given a verified teacher recovery, VH-GPO returns control to the student at the earliest independently confirmed handoff point found within a bounded search budget. With Qwen2.5-1.5B, VH-GPO increases success rates over GRPO from 59.38% to 94.79% on ALFWorld and from 71.62% to 77.34% on WebShop. Improvements also hold with the 7B student, for which additional teacher and student output tokens amount to 1.77% and 2.61% of regular student rollout output tokens on ALFWorld and WebShop, respectively. Ablation studies validate the design of VH-GPO and demonstrate the effectiveness of its key components. The teacher need not solve every problem for the student; instead, it should provide just enough support when the student struggles to help it find a way forward and gradually learn to solve problems independently.[1]

1 INTRODUCTION

Reinforcement learning (RL) has become an important component of large language model (LLM) post-training, improving instruction following and reasoning (Ouyang et al., 2022; Guo et al., 2025). Beyond text generation, LLM agents reason over observations and issue actions to external environments, including web browsing (Nakano et al., 2022; Yao et al., 2023). RL is particularly attractive for improving such agents because the agent learns from trajectories sampled by its own evolving policy. The resulting supervision reflects the states, actions, and errors that the policy will actually encounter, rather than only those chosen by a fixed demonstrator. In long-horizon environments, however, outcome supervision is often extremely sparse: an agent may execute many locally plausible actions and still receive no terminal reward. This problem is especially acute for tasks that require several dependent subgoals, such as manipulating multiple objects in ALFWorld (Shridhar et al., 2021) or satisfying multiple product constraints in WebShop (Yao et al., 2022).

[1] Code is available at `https://anonymous.4open.science/r/VO-86EE`.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Group-relative policy optimization methods such as GRPO (Shao et al., 2024) learn by comparing the outcomes of multiple rollouts for the same task. When every rollout fails with the same terminal reward, however, the group-relative advantages collapse to zero, leaving no outcome-based policy-gradient signal to guide improvement on that task. Such groups expose failures of the current policy without providing a successful alternative to learn from. If these failures recur, the corresponding capability gaps can persist even as the policy improves on tasks with informative reward contrast. DAPO maintains informative training batches by filtering out all-success and all-failure groups through dynamic sampling (Yu et al., 2025). When zero-advantage groups are frequent, as can occur early in training on sparse-reward tasks, dynamic sampling requires additional rollouts to replace discarded groups and fill each training batch, increasing rollout cost. Yet filtering alone does not provide corrective targets for the discarded failures. Fine-grained grouping and graph-based credit assignment improve attribution within sampled experience (Feng et al., 2025; Cheng et al., 2026), yet cannot supply a successful continuation when none is present. This motivates a mechanism that turns selected zero-signal failures into verified learning opportunities.

When the current policy cannot produce a successful continuation, a natural recovery mechanism is to invoke a frozen auxiliary policy from a failure state that the policy has already reached. The continuation is retained only if the environment verifies terminal success (Xia et al., 2026a), thereby turning an otherwise uninformative failure into a grounded positive target. From this point onward, we call the trainable policy the *student* and the frozen recovery policy the *teacher*. A full teacher continuation can convert a failed trajectory into a verified success. However, the longer the teacher remains in control, the more student-generated RL experience is replaced by off-policy imitation on the teacher's state–action distribution.

This raises the question of when a recovering teacher should return control to the student. We formulate *outcome-verified teacher–student handoff* as a training principle: the teacher corrects a student failure, then hands back control once independent student continuations verify successful completion.

Based on this principle, we propose Verified Handoff for Group Policy Optimization (VH-GPO). For an all-failure, zero-signal group, VH-GPO invokes a frozen teacher from a student-reached state and retains its recovery only if the environment verifies success. It then tests states along that recovery to identify, within a bounded search budget, the earliest tested state from which student success is independently confirmed. The resulting short teacher bridge and verified student continuation provide three auxiliary training signals: state-aligned action preferences, teacher-bridge supervision, and student-resume supervision. These signals complement the unchanged root GRPO update. The teacher is used only during training; the student acts independently at inference time.

Experiments on ALFWorld and WebShop with Qwen2.5-Instruct students at 1.5B and 7B scales show that VH-GPO consistently improves the corresponding GRPO baseline. With the 1.5B student, success increases from 59.38% to 94.79% on ALFWorld and from 71.62% to 77.34% on WebShop. With the 7B student, success increases from 78.12% to 95.57% on ALFWorld and from 71.61% to 79.69% on WebShop. These gains require little additional generation: in the 7B runs, all additional teacher and student generation totals just 1.77% of the tokens generated by regular student rollouts on ALFWorld and 2.61% on WebShop.

Our contributions are threefold:

- We highlight the lack of corrective learning targets in all-failure, zero-signal groups and formulate *outcome-verified teacher–student handoff* as a training principle: the teacher should provide only the support needed to restore the student's ability to complete the task, then return control once independent completion is verified.
- We introduce VH-GPO, which implements this principle through outcome-verified teacher recovery and bounded search for the earliest tested, independently confirmed student handoff. The resulting relays supply state-aligned action preferences, teacher-bridge supervision, and student-resume supervision alongside the unchanged root GRPO objective.
- Experiments on ALFWorld and WebShop show that VH-GPO improves overall task success over GRPO and competitive group-based RL baselines at both 1.5B and 7B scales, with modest generation overhead: additional teacher and student generation totals only 1.77% and 2.61% of root rollout output tokens on the two benchmarks, respectively, in the 7B runs.

2 Related Work

Reinforcement learning for language-model agents. RL trains language-model agents for web interaction, device control, and search (Qi et al., 2025; Bai et al., 2024; Jin et al., 2025), supported by environment and training frameworks such as AgentGym, AgentGym-RL, and Agent Lightning (Xi et al., 2025; 2026; Luo et al., 2025). PPO and ArCHer provide actor–critic approaches to sequential decision-making (Schulman et al., 2017; Zhou et al., 2024), while RAGEN examines multi-turn training instabilities, including reward-variance collapse (Wang et al., 2025). Critic-free methods include REINFORCE leave-one-out and GRPO (Ahmadian et al., 2024; Shao et al., 2024); REINFORCE++, Dr. GRPO, and GSPO refine normalization or importance weighting (Hu et al., 2025; Liu et al., 2025; Zheng et al., 2025). Group-relative estimation depends on within-task reward variation. DAPO filters out all-success and all-failure groups through dynamic sampling (Yu et al., 2025).

Fine-grained credit assignment for multi-turn agents. Process supervision supplies intermediate feedback through human or automatically generated step labels (Lightman et al., 2024; Wang et al., 2024). MCTS-based preference learning, ReST-MCTS*, and Agent Q use search to construct finer-grained training signals (Xie et al., 2024; Zhang et al., 2024; Putta et al., 2024). VinePPO estimates advantages using Monte Carlo continuations from intermediate states (Kazemnejad et al., 2025); VH-GPO uses such continuations to verify student resumability. For interactive agents, GiGPO combines episode-level advantages with same-state action groups (Feng et al., 2025), HGPO groups steps by shared historical context (He et al., 2026), and GraphGPO assigns credit through a state-transition graph and distance to a verified goal (Cheng et al., 2026).

Distillation for LLM agents. STaR, ReST, ReST^{EM}, RAFT, and ReST meets ReAct learn from successful or feedback-selected model outputs (Zelikman et al., 2022; Gulcehre et al., 2023; Singh et al., 2024; Dong et al., 2023; Aksitov et al., 2023), while self-imitation learning reinforces past good decisions through an off-policy actor–critic objective (Oh et al., 2018). Teacher supervision can supply additional targets, including rationales in Distilling Step-by-Step (Hsieh et al., 2023). Generalized Knowledge Distillation and MiniLLM reduce distribution mismatch using student-generated sequences and teacher feedback (Agarwal et al., 2024; Gu et al., 2024), but matching teacher predictions alone does not verify environment success. Outcome-Verified Comparative Self-Distillation (OVCSD) (Xia et al., 2026a) organizes failed student rollouts into a prefix tree, invokes a teacher from selected student states, and retains successful continuations for divergence-level comparison and suffix distillation. Full-suffix distillation can include segments that the student could already complete independently. VH-GPO uses an independently confirmed student handoff to determine where teacher supervision ends.

Selective intervention and recovery. DAgger and SafeDAgger study learner-state supervision and selective expert queries (Ross et al., 2011; Zhang & Cho, 2017). HG-DAgger and ThriftyDAgger regulate recovery interventions (Kelly et al., 2019; Hoque et al., 2022). Go-Explore and Jump-Start RL exploit intermediate states (Ecoffet et al., 2021; Uchendu et al., 2023); Reflexion uses reflection without parameter updates (Shinn et al., 2023). ActGuide-RL (Ji et al., 2026) minimizes action-plan guidance after all-failure rollouts and trains guided and unguided trajectories jointly. DASH-OPD (Xia et al., 2026b) uses discrepancy-based hysteresis for teacher–student switching in on-policy distillation. VH-GPO instead verifies student handoff through independently sampled successful continuations to learn from all-failure groups in group-relative RL.

3 Method

3.1 Problem Setup

We consider episodic agent tasks with environment-verifiable outcomes. In our training setup, each episode is capped at H environment interactions. After t interactions, the remaining interaction budget is $b_t = H - t$. This cap is set by the training protocol; b_t measures the remaining execution allowance and does not estimate the number of steps needed to reach success.

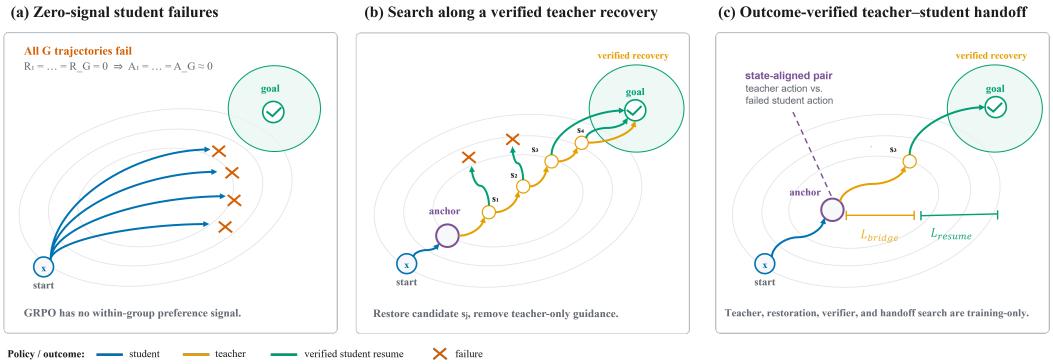

Figure 1: Overview of VH-GPO. (a) When every student trajectory fails, group-relative advantages provide no useful within-group preference. (b) A frozen teacher recovers from a restorable student state; candidate states along this verified recovery are restored and tested by the frozen student. (c) The earliest state from which student success is independently confirmed defines the handoff frontier. The retained trajectory consists of the original student prefix, a short teacher bridge, and a verified student resume; a state-aligned teacher–student divergence supplies the pairwise target.

For a task instruction x, we freeze the current trainable policy as π_{old} at the beginning of each update and sample G independent interaction trajectories. Let τ_i be the i-th trajectory and R_i its final environment reward. The rollout group is

$$\mathcal{G}_x = \{\tau_1, \ldots, \tau_G\}, \qquad \tau_i \sim \pi_{\text{old}}. \tag{1}$$

The trainable policy is denoted by π_θ. GRPO compares rewards within $\mathcal{G}_x$ and assigns each trajectory a group-relative advantage A_i. Reward variation therefore identifies which sampled behaviors should become more or less likely. If every trajectory fails with the same reward, however,

$$R_1 = \cdots = R_G \quad \Longrightarrow \quad A_1 = \cdots = A_G \approx 0. \tag{2}$$

and the group provides little or no policy-gradient signal. We call this an *all-failure, zero-signal group*.

Groups with meaningful reward contrast follow the standard GRPO update. VH-GPO intervenes only when every trajectory fails, the rewards are equal up to a small tolerance, and the resulting advantages are numerically zero. These groups define the failure cases addressed by our method.

3.2 VH-GPO Overview

VH-GPO constructs a verified training relay from an all-failure, zero-signal group (Figure 1). A frozen teacher first recovers from a restorable student state. Within a bounded search budget, we then identify the earliest tested state from which the frozen student's success is independently confirmed. The relay retains the original student prefix, the teacher bridge before handoff, and the verified student continuation:

$$\boxed{\text{student prefix} \longrightarrow \text{teacher bridge} \longrightarrow \text{student resume}}. \tag{3}$$

Only candidates that pass state restoration, teacher recovery, independent student confirmation, and full-trajectory replay contribute auxiliary targets.

3.3 Verified Teacher–Student Handoff and Relay Construction

Obtaining a verified recovery.

For each selected all-failure, zero-signal group, we choose one failed student trajectory as the recovery *backbone*. To reserve interaction budget for recovery, we retain collected candidate states with $b_t \geq b_{\min}$, where $b_{\min}$ is a configured minimum budget. We select the failed trajectory with the deepest candidate satisfying the budget condition. An unsuccessful episode may still contain useful

task progress. By first attempting recovery from a late student-visited state, we aim to build on this progress and obtain corrective supervision for the situation created by the student's own actions, reducing the need for the teacher to demonstrate earlier portions of the task again. The group contributes no auxiliary data if no eligible trajectory exists or the selected backbone lacks an initial-state candidate.

We call the state where the frozen teacher π_T begins acting the *anchor*. Starting from the deepest eligible candidate on the selected backbone, we construct a bounded schedule of progressively earlier candidates using a fixed backtracking stride, including the initial state as the final fallback. At each anchor, we restore and verify the state, then let the teacher attempt recovery using the task and visible interaction history within the remaining interaction budget. We retain the first anchor yielding a verified successful recovery; if all scheduled anchors fail, we discard the group.

If multiple attempts from the same anchor succeed, we keep the shorter recovery. A retained recovery with L teacher actions is written as

$$\tau_T = (a_1^T, a_2^T, \ldots, a_L^T), \tag{4}$$

where the superscript T marks teacher actions.

Handoff selection and relay verification.

The verified recovery shows that the task can be completed from the anchor, but the teacher may not need to execute all L actions. Let s_j be the state after its first j actions, with s_0 denoting the anchor and s_L the terminal state. At a tested nonterminal state s_j, we return control to the frozen rollout policy π_{old} and let it attempt the remainder of the task without teacher assistance, subject to the original episode's remaining interaction budget.

Within a fixed search budget, we test candidate handoff states and select the earliest one with a successful student continuation. We then independently validate this candidate using newly sampled continuations from the same state. Search details are given in Appendix A. Let $\mathcal{J}_{\text{test}}$ be the set of indices examined within this search budget. We select

$$j^\star = \min \{ j \in \mathcal{J}_{\text{test}} : \text{success from } s_j \text{ is independently confirmed} \}. \tag{5}$$

We require $j^\star < L$ so that control returns before the teacher completes the task.

The final relay concatenates the original student trajectory before the anchor, the teacher actions $(a_1^T, \ldots, a_{j^\star}^T)$, and the independently confirmed student continuation from $s_{j^\star}$.

3.4 Relay-based Training Objectives

Given a verified relay, we derive up to three complementary signals: an action preference, teacher-bridge completion, and student-resume completion. The three objectives are normalized independently. We write $s_\theta(y \mid h)$ for the student's mean token log probability of target text y given student-visible context h.

Action Preference Objective.

Let h_d be the earliest verified prompt at which the teacher bridge and the failed student trajectory choose different executable actions. At this prompt, a_{chosen} is the teacher action from the successful relay and a_{rejected} is the student action from the failed trajectory. The current student's chosen-over-rejected score, computed over executable action tokens, is

$$p_\theta = s_\theta(a_{\text{chosen}} \mid h_d) - s_\theta(a_{\text{rejected}} \mid h_d). \tag{6}$$

Let p_{ref} be the same score computed by a fixed reference policy initialized from the student's starting checkpoint. Let σ be the logistic sigmoid and let the positive coefficient β control preference strength. We adapt the reference-anchored preference objective of DPO (Rafailov et al., 2023) to state-aligned executable actions using the mean-token scores above. The loss for one action pair is

$$\ell_{\text{pair}} = -\log \sigma(\beta(p_\theta - p_{\text{ref}})). \tag{7}$$

We denote the average over valid pairs by $\mathcal{L}_{\text{pair}}$. This objective teaches the student to prefer the teacher action from a verified successful relay over the action observed in a failed student trajectory at the same state, providing a localized correction at their first action divergence.

Teacher Bridge Objective.

Let $\mathcal{B}$ contain the prompt–response pairs from all verified teacher bridges in the update. For each pair (h, y), h is the student-visible prompt and y is the complete teacher response for that turn. The teacher-bridge objective is

$$\mathcal{L}_{\text{bridge}} = -\frac{1}{|\mathcal{B}|} \sum_{(h,y)\in\mathcal{B}} s_\theta(y \mid h). \tag{8}$$

This objective teaches the minimal corrective behavior needed to restore student solvability. It is omitted when $\mathcal{B}$ is empty.

Student Resume Objective.

Let $\mathcal{C}$ contain the prompt–response pairs from all independently confirmed student resumes in the update, and let $|\mathcal{C}|$ be their number. The student-resume objective is

$$\mathcal{L}_{\text{resume}} = -\frac{1}{|\mathcal{C}|} \sum_{(h,y)\in\mathcal{C}} s_\theta(y \mid h). \tag{9}$$

This objective reinforces student-generated actions that are verified to lead to success. It is an outcome-filtered positive completion objective rather than an unbiased policy-gradient estimator, and it is omitted when $\mathcal{C}$ is empty. It shares the motivation of learning from the agent's own good experience with self-imitation learning (Oh et al., 2018). Unlike SIL's off-policy actor–critic formulation, we use a positive completion loss on independently verified student continuations following teacher intervention.

Overall Objective.

Let the non-negative coefficients λ_{pair}, λ_{bridge}, and λ_{resume} weight the three relay-based losses. Their combined auxiliary objective is

$$\mathcal{L}_{\text{aux}} = \lambda_{\text{pair}}\mathcal{L}_{\text{pair}} + \lambda_{\text{bridge}}\mathcal{L}_{\text{bridge}} + \lambda_{\text{resume}}\mathcal{L}_{\text{resume}}. \tag{10}$$

Let $\mathcal{L}_{\text{GRPO}}$ denote the unchanged standard GRPO objective and let $\mathcal{L}_{\text{KL}}$ denote the reference-policy regularizer, weighted by η. The optimizer combines the available terms as

$$\mathcal{L} = \mathcal{L}_{\text{GRPO}} + \eta\mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{aux}}. \tag{11}$$

These terms are generally contributed by different rollout groups within the same training batch. Groups with valid reward contrast contribute $\mathcal{L}_{\text{GRPO}}$. Zero-signal all-failure groups contribute $\mathcal{L}_{\text{aux}}$ only if recovery and relay verification succeed. Thus, the total loss is combined at the batch or optimizer-update level; a failed trajectory is not simultaneously treated as an ordinary GRPO positive sample and a teacher-supervised positive target. The root GRPO and auxiliary terms are normalized separately before combination. Concrete coefficient values are specified in the experimental setup.

The complete training algorithm and deployment details are provided in Appendix A.

4 Experiments

We evaluate VH-GPO with two questions in mind. First, does outcome-verified teacher–student handoff improve the task performance of LLM agents? Second, which parts of the method are responsible for the improvement? We examine these questions on two long-horizon interactive benchmarks at two model scales, complemented by controlled ablations and relay diagnostics.

4.1 Experimental Setup

Benchmarks and metrics. We evaluate on ALFWorld (Shridhar et al., 2021), a text environment with six embodied task families, and WebShop (Yao et al., 2022), where agents purchase products satisfying natural-language requests. We report per-family and overall success on ALFWorld, and attribute-based partial-credit score and exact success on WebShop. Main and ablation results are reported as mean $\pm$ standard deviation in percent over three independent runs with different random seeds. Reward definitions are given in Appendix B.

Table 1: Final performance on ALFWorld and WebShop (%), reported as mean ± standard deviation over three seeds. Bold denotes the best mean within a model-size block.

		ALFWorld							WebShop	
Model	Method	Pick	Clean	Cool	Look	Heat	Pick2	All	Score	Succ.
1.5B	GRPO	75.00±2.91	74.73±2.58	63.06±3.80	50.00±0.00	38.46±12.56	35.01±4.99	59.38±1.28	85.98±2.20	71.62±1.48
1.5B	GiGPO	94.05±1.68	**98.96**±1.47	69.12±6.31	77.78±7.85	**100.00**±0.00	92.64±8.02	90.63±1.11	84.83±2.76	71.09±1.28
1.5B	GraphGPO	94.05±3.37	96.84±0.05	**84.56**±2.55	83.33±0.00	**100.00**±0.00	91.58±1.62	92.71±0.74	86.55±0.97	75.52±3.21
1.5B	HGPO ($K = 4$)	97.62±1.68	93.71±2.51	76.91±0.51	77.78±7.85	94.87±3.63	74.69±0.44	86.98±0.97	88.23±1.13	75.52±2.42
1.5B	VH-GPO	**98.81**±1.68	95.80±1.45	83.04±2.37	**94.44**±7.86	97.44±3.63	**97.58**±1.71	**94.79**±0.97	**88.71**±1.45	**77.34**±3.37
7B	GRPO	82.14±2.91	84.21±0.24	61.54±5.57	77.78±7.85	89.75±3.63	74.74±2.60	78.12±0.64	83.35±2.84	71.61±4.70
7B	GiGPO	96.43±0.00	88.44±1.33	86.14±0.31	**100.00**±0.00	**100.00**±0.00	**94.01**±3.34	92.71±0.37	86.00±1.17	75.78±0.64
7B	GraphGPO	98.81±1.68	95.83±2.95	89.32±4.19	88.89±7.86	94.87±3.63	93.96±1.75	94.53±0.64	88.18±1.62	78.38±3.15
7B	HGPO ($K = 4$)	97.62±1.68	91.63±3.85	90.91±7.42	94.44±7.86	**100.00**±0.00	**94.01**±3.34	94.27±1.33	88.59±0.68	78.12±3.37
7B	VH-GPO	**100.00**±0.00	**97.88**±1.50	**98.48**±2.14	**100.00**±0.00	97.44±3.63	84.26±6.34	**95.57**±0.98	**88.81**±0.56	**79.69**±2.92

Implementation Details. We train Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct students (Yang et al., 2024) from their instruction-tuned checkpoints for 150 updates using AdamW with learning rate 10^{-6}, betas $(0.9, 0.999)$, and weight decay 0.01. Each update uses 16 tasks with eight root rollouts each. GRPO uses clipping parameter 0.2; the KL coefficient is $\eta = 0$ on ALFWorld and $\eta = 0.01$ on WebShop. Interaction horizons for training and evaluation are 50 steps on ALFWorld and 15 on WebShop. Student responses are capped at 512 tokens. We set the pairwise $\beta = 0.1$ and $(\lambda_{\text{pair}}, \lambda_{\text{bridge}}, \lambda_{\text{resume}}) = (0.5, 0.25, 0.1)$, processing at most two eligible zero-signal groups per update. The teacher is a frozen, GRPO-trained Qwen2.5-1.5B-Instruct policy used only during training.

Baselines. We compare VH-GPO with representative and competitive RL training methods: GRPO (Shao et al., 2024), GiGPO (Feng et al., 2025), GraphGPO (Cheng et al., 2026), and HGPO (He et al., 2026). HGPO uses $K = 4$ at both model scales. Our reimplementation of OVCSD (Xia et al., 2026a) and its results under our experimental setting are reported in Appendix C.

4.2 Main Results

Table 1 shows that VH-GPO achieves the highest overall ALFWorld success and both WebShop metrics among the compared methods at both model scales. Compared with credit-assignment baselines, VH-GPO addresses not only where to assign credit but also what successful behavior is available to learn. For all-failure groups, it constructs a verified alternative from a student-visited state, turning otherwise uninformative outcomes into corrective action preferences and successful completion targets.

The 7B results further suggest that useful assistance need not come from a larger teacher. Verification allows VH-GPO to select locally effective corrections from a frozen 1.5B teacher while leaving the remaining task to the 7B student. This division of labor exploits teacher recoveries without replacing the student's own completion behavior as a training target.

4.3 Training Dynamics

To complement the final evaluation in Table 1, Figure 2(a) visualizes the training dynamics on ALFWorld for the 1.5B student. VH-GPO enters the leading group early and maintains a higher smoothed training success rate through most of the run. Panels (b)–(c) show that selected zero-signal groups become less frequent and cumulative additional output tokens grow more slowly later in training.

4.4 Ablation Study

We ablate VH-GPO on ALFWorld with the 1.5B student using the same 150-update budget for all variants. We separately examine the training objectives, recovery and handoff mechanisms, and additional student rollout budget. All ablation tables report success as mean ± standard deviation in percent and the mean change Δ in percentage points relative to full VH-GPO. Detailed ablation protocols are provided in Appendix D.

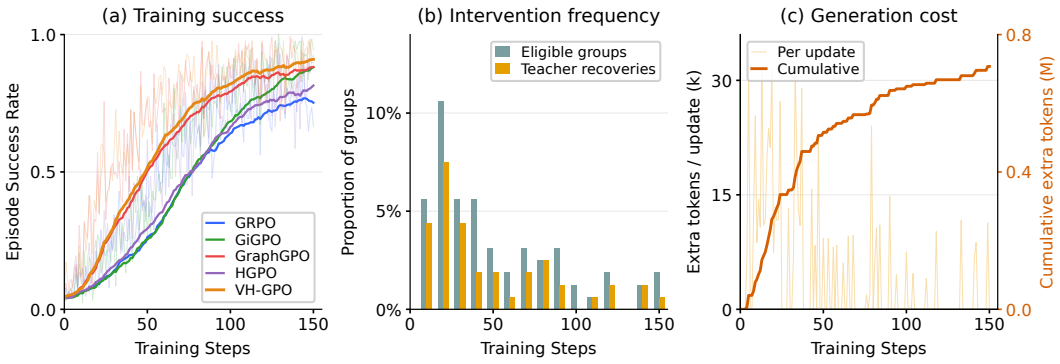

Figure 2: Training dynamics on ALFWorld with Qwen2.5-1.5B-Instruct. (a) Training episode success rate: light curves show per-update values; dark curves show EMA trends with decay $\alpha = 0.95$. (b) VH-GPO intervention frequency: eligible zero-signal groups selected for recovery and verified teacher recoveries, as proportions of all training groups in ten-update windows (16 groups per update). (c) VH-GPO additional generation: per-update tokens (left axis, thousands) and cumulative tokens (right axis, millions), including teacher recovery, student search, confirmation, and subsequently discarded attempts.

4.4.1 Training Objectives

Table 2: Training-objective ablations on ALFWorld with the 1.5B student.

Method	Success	Δ
VH-GPO	**94.79**$\pm$0.97	–
w/o action preference	88.28$\pm$1.68	-6.51
w/o student resume loss	90.62$\pm$0.64	-4.17
w/o teacher bridge loss	80.47$\pm$1.69	-14.32

The variants in Table 2 each remove one auxiliary loss while retaining relay construction and the other two objectives. Removing the teacher bridge loss yields the largest decrease, from 94.79% to 80.47%. Equal failure rewards in these zero-signal groups cannot distinguish useful corrective behavior. The bridge loss converts a verified recovery into stepwise supervision, increasing the likelihood of teacher responses that lead toward a student-solvable state. It thus trains the sequence of decisions needed to escape the observed failure. Removing this loss eliminates direct imitation of that sequence; the performance drop supports the value of explicitly learning how to recover.

Removing action preference reduces success by 6.51 percentage points. Bridge imitation still rewards the teacher's responses, but no longer explicitly contrasts the chosen action with the failed student's action at the same state. The drop suggests that this localized contrast adds useful decision-level supervision beyond imitating positive targets alone.

Removing the student resume loss reduces success by 4.17 points. Verified student continuations are still collected, but are no longer reinforced as completion targets. This suggests that being able to sample a successful continuation is not equivalent to consolidating that behavior through training: learning the correction and learning to finish independently provide complementary benefits.

4.4.2 Recovery and Handoff Mechanisms

Table 3 examines where teacher intervention starts and ends. *Random Failure Anchor* shuffles the non-root anchor order on the same selected backbone, retaining the candidate set, attempt limits, and root fallback; success falls by 3.90 points. *Random Verified Handoff* retains the discovery search and budget but uniformly selects an alternative discovery-success state after excluding the earliest one. Independent confirmation and full-trajectory replay remain required, and success falls by 4.69 points. *Terminal Teacher* instead supervises the full successful teacher suffix, retains valid action preference pairs, and omits handoff search and student-resume supervision while still requiring full-

Table 3: Recovery and handoff ablations on ALFWorld with the 1.5B student.

Method	Success	Δ
VH-GPO	**94.79**$\pm$0.97	–
Random Failure Anchor	90.89$\pm$0.74	−3.90
Random Verified Handoff	90.10$\pm$0.98	−4.69
Terminal Teacher	86.98$\pm$1.33	−7.81

trajectory replay. Its 7.81-point drop supports combining a short teacher correction with a verified student continuation rather than relying on full-suffix teacher supervision. Together, these comparisons support the designed anchor order and early handoff selection.

4.4.3 Additional Rollout Budget

Table 4: Additional-rollout comparison on ALFWorld with the 1.5B student.

Method	Success	Δ
VH-GPO	**94.79**$\pm$0.97	–
GRPO	59.38$\pm$1.28	−35.41
Extra-rollout GRPO	64.32$\pm$2.88	−30.47

Extra-rollout GRPO generates four additional complete student trajectories for each selected zero-signal group and trains them with ordinary GRPO, without a teacher; the budget rationale is given in Appendix D. It improves success over GRPO by 4.94 percentage points to 64.32%, but remains 30.47 points below VH-GPO (Table 4), indicating that this additional-sampling configuration does not reproduce VH-GPO's gains.

4.5 Generation Overhead

Table 5 reports additional output tokens summed over all 150 updates of the 7B runs, normalized by the corresponding root rollout output tokens. The count includes teacher recovery, student handoff search, and independent confirmation, including failed attempts and ultimately discarded candidates. With recovery limited to at most two selected zero-signal groups per update, additional generation amounts to 1.77% on ALFWorld and 2.61% on WebShop, with teacher recovery accounting for most of the overhead. Appendix E reports the corresponding recovery yields and retained-relay statistics.

Table 5: Additional generation in the 7B training runs. Each entry is a percentage of root rollout output tokens, summed over 150 updates. Search and Confirmation denote student continuations.

Benchmark	Teacher recovery	Search	Confirmation	Total
ALFWorld	1.04	0.47	0.26	1.77
WebShop	2.07	0.41	0.13	2.61

5 Conclusions

VH-GPO learns from all-failure, zero-signal groups through verified teacher bridges and student continuations, returning control at the earliest independently confirmed handoff found within a bounded search budget. Experiments show improved agent performance with modest additional output-token overhead and no teacher at inference. Ablations suggest complementary benefits from local correction and student-generated completion.

Limitations. VH-GPO requires restorable environment states and verifiable task outcomes. Its budget-limited handoff search incurs additional rollouts and may miss earlier feasible handoffs. More efficient search strategies remain future work.

REPRODUCIBILITY STATEMENT

We provide implementation details and training hyperparameters in Section 4.1, the complete training algorithm and recovery-search configuration in Appendix A, reward definitions and evaluation metrics in Appendix B, details of the OVCSD reimplementation in Appendix C, ablation protocols in Appendix D, and relay diagnostics in Appendix E. Anonymous source code is provided with the submission.

AI USE STATEMENT

We used generative AI tools to assist with language polishing and manuscript formatting. AI tools were not used to generate the core scientific content of this work, including the research ideas, proposed method, experimental design, or scientific conclusions. The authors take full responsibility for the accuracy and integrity of the final manuscript.

REFERENCES

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

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Renat Aksitov, Sobhan Miryoosefi, Zonglin Li, Daliang Li, Sheila Babayan, Kavya Kopparapu, Zachary Fisher, Ruiqi Guo, Sushant Prakash, Pranesh Srinivasan, Manzil Zaheer, Felix Yu, and Sanjiv Kumar. ReST meets ReAct: Self-Improvement for Multi-Step Reasoning LLM Agent. *arXiv preprint arXiv:2312.10003*, 2023.

Hao Bai, Yifei Zhou, Mert Cemri, Jiayi Pan, Alane Suhr, Sergey Levine, and Aviral Kumar. DigiRL: Training In-The-Wild Device-Control Agents with Autonomous Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2024.

Xin Cheng, Shuo He, Lang Feng, HaiYang Xu, Ming Yan, Lei Feng, and Bo An. Beyond trajectory-level attribution: Graph-based credit assignment for agentic reinforcement learning. In *International Conference on Machine Learning*, 2026.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun SHUM, and Tong Zhang. RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*, 2023.

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. First return, then explore. *Nature*, 2021.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Advances in Neural Information Processing Systems*, 2025.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *International Conference on Learning Representations*, 2024.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced Self-Training (ReST) for Language Modeling. *arXiv preprint arXiv:2308.08998*, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 2025.

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

Ryan Hoque, Ashwin Balakrishna, Ellen Novoseller, Albert Wilcox, Daniel S. Brown, and Ken Goldberg. ThriftyDAgger: Budget-Aware Novelty and Risk Gating for Interactive Imitation Learning. In *Proceedings of the 5th Conference on Robot Learning*, 2022.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. REINFORCE++: Stabilizing Critic-Free Policy Optimization with Global Advantage Normalization. *arXiv preprint arXiv:2501.03262*, 2025.

Yuxiang Ji, Zengbin Wang, Yong Wang, Shidong Yang, Ziyu Ma, Guanhua Chen, Zonghua Sun, Liaoni Wu, and Xiangxiang Chu. Learning Agentic Policy from Action Guidance. *arXiv preprint arXiv:2605.12004*, 2026.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*, 2025.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining Credit Assignment in RL Training of LLMs. In *Forty-second International Conference on Machine Learning*, 2025.

Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. HG-DAgger: Interactive imitation learning with human experts. In *IEEE International Conference on Robotics and Automation*, 2019.

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

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Conference on Language Modeling (COLM)*, 2025.

Xufang Luo, Yuge Zhang, Zhiyuan He, Zilong Wang, Siyun Zhao, Dongsheng Li, Luna K. Qiu, and Yuqing Yang. Agent Lightning: Train ANY AI Agents with Reinforcement Learning. *arXiv preprint arXiv:2508.03680*, 2025.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2022.

Junhyuk Oh, Yijie Guo, Satinder Singh, and Honglak Lee. Self-imitation learning. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

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

Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent Q: Advanced Reasoning and Learning for Autonomous AI Agents. *arXiv preprint arXiv:2408.07199*, 2024.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Jiadai Sun, Xinyue Yang, Yu Yang, Shuntian Yao, Wei Xu, Jie Tang, and Yuxiao Dong. WebRL: Training LLM Web Agents via Self-Evolving Online Curriculum Reinforcement Learning. In *International Conference on Learning Representations*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011.

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

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

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2023.

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

Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron T Parisi, Abhishek Kumar, Alexander A Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Fathy Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura A Culp, Lechao Xiao, Maxwell Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond Human Data: Scaling Self-Training for Problem-Solving with Language Models. *Transactions on Machine Learning Research*, 2024.

Ikechukwu Uchendu, Ted Xiao, Yao Lu, Banghua Zhu, Mengyuan Yan, Joséphine Simon, Matthew Bennice, Chuyuan Fu, Cong Ma, Jiantao Jiao, Sergey Levine, and Karol Hausman. Jump-Start Reinforcement Learning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Fei-Fei Li, Lijuan Wang, Yejin Choi, and Manling Li. RAGEN: Understanding self-evolution in LLM agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025.

Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Xin Guo, Dingwen Yang, Chenyang Liao, Wei He, Songyang Gao, Lu Chen, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. AgentGym: Evaluating and training large language model-based agents across diverse environments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Jiaqi Liu, Honglin Guo, yajie yang, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, Wei He, Yiwen Ding, Guanyu Li, Zehui

Chen, Zhengyin Du, Xuesong Yao, Yufei Xu, Jiecao Chen, Tao Gui, Zuxuan Wu, Qi Zhang, Xuanjing Huang, and Yu-Gang Jiang. AgentGym-RL: An open-source framework to train LLM agents for long-horizon decision making via multi-turn RL. In *International Conference on Learning Representations*, 2026.

Xu Xia, Jinghua Piao, Min Yang, Xiaochong Lan, Jiaju Chen, and Yong Li. From scoring to acting: Outcome-verified comparative self-distillation for LLM agents. *arXiv preprint arXiv:2607.27937*, 2026a.

Yuchen Xia, Qianguo Sun, Chao Song, Junlong Wu, Yiyan Qi, and Yunjian Xu. DASH-OPD: Discrepancy-Aware Switching with Hysteresis for On-Policy Distillation. *arXiv preprint arXiv:2607.29078*, 2026b.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P. Lillicrap, Kenji Kawaguchi, and Michael Shieh. Monte Carlo Tree Search Boosts Reasoning via Iterative Preference Learning. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping Reasoning With Reasoning. In *Advances in Neural Information Processing Systems*, 2022.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. ReST-MCTS*: LLM Self-Training via Process Reward Guided Tree Search. In *Advances in Neural Information Processing Systems*, 2024.

Jiakai Zhang and Kyunghyun Cho. Query-Efficient Imitation Learning for End-to-End Autonomous Driving. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group Sequence Policy Optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

A Training Algorithm and Deployment

Algorithm 1: VH-GPO training

Input: task batch $\mathcal{X}$, student π_θ, frozen teacher π_T, intervention cap M
Output: updated student parameters θ
Freeze the current student as π_{old}
Initialize the GRPO and auxiliary loss accumulators
Initialize the number of selected zero-signal groups $m \leftarrow 0$
for each task $x \in \mathcal{X}$ **do**
  Collect rollout group $\mathcal{G}_x$ using π_{old}
  if $\mathcal{G}_x$ has useful reward contrast **then**
    Accumulate its standard GRPO objective
    continue
  end if
  if $\mathcal{G}_x$ is not an all-failure, zero-signal group **then continue**
  if $m = M$ **then continue**; otherwise set $m \leftarrow m + 1$
  Select a backbone by the deepest candidate satisfying the budget condition
  if the backbone or its initial-state candidate is missing **then continue**
  Schedule anchors on this backbone from deep to shallow, ending at the root
  Mark the teacher recovery as unavailable
  for each scheduled anchor **do**
    Restore the anchor and verify the reconstructed state
    if restoration fails **then continue**
    Sample teacher recovery attempts from the restored anchor
    if the environment verifies any success **then**
      Retain the shortest successful recovery and its anchor; **break**
    end if
  end for
  if no teacher recovery was retained **then continue**
  Find the earliest independently confirmed student handoff point
  if no such point exists **then continue**
  Build and replay the prefix–bridge–resume relay
  if replay or information-leakage checks fail **then continue**
  Accumulate its available pair, bridge, and resume objectives
end for
Normalize the GRPO and auxiliary objectives separately
Update θ using their combined objective and **return** θ

Any failure in restoration, teacher recovery, student confirmation, relay replay, or auxiliary-example construction causes that candidate to contribute no teacher-derived gradient; other groups continue through their normal paths. At deployment, only the trained student $\pi_\theta(a_t \mid h_t)$ remains. The teacher, state restoration, verifier, relay search, and handoff mechanism are all removed. VH-GPO is therefore a training-only procedure and introduces no teacher dependency at inference time.

Anchor selection and recovery budgets. Table 6 lists the recovery-search settings used at both student model scales. Starting from the deepest eligible anchor on the selected backbone, each backtrack selects the deepest available candidate whose depth is at least one stride earlier than the current anchor. The schedule contains the initial anchor and at most four backtracked anchors. If the root is not already included, it is appended as the final fallback, giving at most six anchors per group. The actual number can be smaller when fewer candidates are available. Anchor search stops at the first anchor yielding a verified successful teacher recovery. Thus, the upper bound is 12 teacher recovery attempts per selected group; this bound excludes student handoff-search and confirmation rollouts.

Handoff search. Discovery begins near the end of the verified teacher recovery and moves backward with exponentially increasing jumps. After the first tested failure, the search tests intermediate states in increasing order between that state and the earliest later discovery-successful state, or the

Table 6: Anchor-selection and teacher-recovery parameters for both the 1.5B and 7B students. Step counts refer to environment interactions.

Parameter	ALFWorld	WebShop
Minimum remaining budget $b_{\min}$ (steps)	16	6
Anchor backtracking stride (steps)	8	2
Maximum backtracks	4	4
Maximum anchors per group, including root	6	6
Teacher recovery attempts per anchor	2	2

terminal state if no discovery success has been found. This refinement stops at the first success or when the search budget is exhausted. The earliest discovery-successful state among all tested candidates is selected for independent confirmation.

B Reward Design and Evaluation Metrics

ALFWorld. The environment checks whether the final state satisfies all conditions of the task, including the required object, its state, and its destination. For the text-only setting used here, the task reward is $r_{\text{task}} = 10\,\mathbf{1}[\text{won}]$: successful episodes receive 10 and unsuccessful episodes receive 0. Intermediate task progress does not provide a separate positive task reward. We report the percentage of episodes with $\text{won} = 1$ for each task family and overall.

WebShop. The environment assigns a partial-credit score when the agent purchases a product (Yao et al., 2022). Let N_{attr} and N_{opt} be the numbers of requested attributes and options, M_{attr} and M_{opt} the numbers matched by the purchased product, and P indicate that its price meets the upper bound. The original task score is

$$s = r_{\text{type}} \frac{M_{\text{attr}} + M_{\text{opt}} + P}{N_{\text{attr}} + N_{\text{opt}} + 1}, \tag{12}$$

where $r_{\text{type}} \in [0, 1]$ measures whether the product type matches the request. The score is zero before purchase and lies in $[0, 1]$ at purchase. Our training environment retains s for evaluation but uses the binary task reward $r_{\text{task}} = 10\,\mathbf{1}[\text{purchase completed} \wedge s = 1]$. Thus, partial credit does not become a positive task reward during training. The reported WebShop *Score* is 100 times the mean s, whereas *Succ.* is 100 times the fraction of episodes with $s = 1$.

In both environments, the training configuration additionally subtracts 0.1 for each invalidly formatted action. This penalty is separate from the task reward and is not included in the success-rate or WebShop Score metrics above.

C Comparison with an OVCSD Reimplementation

An official implementation of OVCSD (Xia et al., 2026a) was not publicly available when we conducted these experiments. To compare the methods under the same experimental setting, we reimplemented OVCSD following the description in its paper and evaluated it using the setup in Section 4.1. Table 7 reports our reimplementation results, not the scores reported by the original authors. Differences in implementation and experimental settings may contribute to the gap from the original report; these results should therefore be interpreted as a comparison under our evaluation protocol.

Under this protocol, VH-GPO achieves higher overall ALFWorld success and both WebShop metrics than our OVCSD reimplementation at both scales. Whereas OVCSD supervises verified teacher suffixes, VH-GPO concentrates teacher imitation on a corrective bridge and reinforces the student's own successful continuation. VH-GPO returns control after verifying that the student can complete the remaining task independently. It limits teacher imitation to the corrective actions before this handoff. These results suggest that the value of teacher assistance depends not only on obtaining a successful recovery, but also on using it to teach local correction while reinforcing independent student completion.

Table 7: OVCSD reimplementation results under our experimental setting, reported as mean ± standard deviation over three seeds (%). VH-GPO results from Table 1 are repeated for reference. Bold denotes the best mean within each model-size block in this table.

		ALFWorld							WebShop	
Model	Method	Pick	Clean	Cool	Look	Heat	Pick2	All	Score	Succ.
1.5B	OVCSD	84.52±1.68	77.86±4.62	63.06±3.80	38.89±7.86	84.62±0.00	57.81±4.66	71.36±1.33	83.85±2.61	72.40±1.33
1.5B	VH-GPO	**98.81**±1.68	**95.80**±1.45	**83.04**±2.37	**94.44**±7.86	**97.44**±3.63	**97.58**±1.71	**94.79**±0.97	**88.71**±1.45	**77.34**±3.37
7B	OVCSD	**100.00**±0.00	**100.00**±0.00	70.85±3.78	77.78±7.85	94.87±3.63	77.12±1.53	88.54±1.33	82.91±1.27	76.04±2.05
7B	VH-GPO	**100.00**±0.00	97.88±1.50	**98.48**±2.14	**100.00**±0.00	**97.44**±3.63	**84.26**±6.34	**95.57**±0.98	**88.81**±0.56	**79.69**±2.92

D Ablation Protocols

Random Verified Handoff. This variant uses the same discovery search procedure and rollout budget as full VH-GPO. Each tested nonterminal state receives two student continuations; a state enters the discovery-success set S if at least one continuation succeeds. Discovery uses at most eight rollouts per teacher recovery. Let $j_{\text{early}} = \min S$ denote the earliest discovery-success index, which full VH-GPO selects for independent confirmation. When alternative discovery-success states are available, Random Verified Handoff samples uniformly from $S \setminus \{j_{\text{early}}\}$. Thus, the discovery procedure, tested candidate set, and budget limit are unchanged, but the final sampling set explicitly excludes the earliest successful candidate. The selected state receives four additional confirmation continuations using seeds distinct from discovery. At least one must succeed before the complete student-prefix–teacher-bridge–student-resume action sequence is replayed from the initial state. Only a successful replay qualifies the relay for auxiliary training.

Random Failure Anchor. This variant first selects the same failed backbone as full VH-GPO: the trajectory with the deepest budget-eligible candidate, with trajectory ID used to break depth ties. It then constructs the same anchor list using the remaining-interaction-budget constraint, backtracking stride, and limit on anchor attempts. Only the non-root anchors are randomly permuted; the root anchor remains last. After all teacher attempts at one anchor fail, the procedure advances to the next anchor in this order and stops at the first verified recovery. If all non-root anchors fail, it still attempts recovery from the root. The candidate set and attempt limits are unchanged, but the next anchor need not be shallower than the previous one. Actual resource use depends on when the first recovery succeeds. Handoff search and confirmation follow the full method.

Extra-rollout GRPO. We choose the number of extra student trajectories by expressing VH-GPO's configured token reservation for confirmation continuations in units of complete student rollouts. On ALFWorld, the theoretical maximum output-token budget for one complete trajectory is

$$B_{\text{trajectory}} = 50 \times 512 = 25{,}600, \tag{13}$$

using the environment-step cap and maximum response length. The configuration reserves four such budgets, corresponding to the four parallel student continuations used for independent confirmation:

$$B_{\text{reserve}} = 4B_{\text{trajectory}} = 102{,}400, \qquad N_{\text{extra}} = \frac{B_{\text{reserve}}}{B_{\text{trajectory}}} = 4. \tag{14}$$

Thus, four extra trajectories are chosen because their combined maximum output-token budget equals the configured reservation for VH-GPO's four parallel confirmation continuations. This conversion provides a configuration-based budget rationale for allocating additional full student rollouts to each selected zero-signal group. Training these rollouts with ordinary GRPO tests whether spending this allocation on independent student exploration can reproduce the gains from verified teacher recovery and student resumption.

E Handoff Behavior

Table 8 summarizes recovery and relay construction during the 7B training runs. Eligible groups are all-failure groups that pass the zero-signal checks and are selected for recovery attempts. A recovery

is counted when the environment verifies the teacher's terminal success. A retained relay additionally requires independently confirmed nonterminal student handoff, successful full-trajectory replay, and prompt-sanitization checks. Successful replay of a full teacher recovery alone does not qualify it for auxiliary training.

Table 8: Relay diagnostics for the 7B training runs. Parentheses report percentages relative to eligible groups for Recoveries and verified recoveries for Retained relays. Only retained relays contribute auxiliary training samples.

Benchmark	Eligible	Recoveries	Retained relays
ALFWorld	50	25 (50.0)	18 (72.0)
WebShop	150	29 (19.3)	10 (34.5)

The teacher successfully recovers 25 of 50 selected groups on ALFWorld and 29 of 150 on WebShop. All these recoveries pass full-trajectory replay, but only 18 and 10, respectively, also yield an independently confirmed nonterminal handoff and are retained for auxiliary training. These represent 72.0% and 34.5% of verified recoveries. The remaining 7 ALFWorld and 19 WebShop recoveries lack a confirmed handoff and contribute no auxiliary training samples, despite successful replay of the teacher-completed trajectory. This distinction shows that teacher success does not by itself establish student resumability. Every retained relay in these runs also provides a state-aligned action preference pair. No state-restoration failures or privileged-information leaks are recorded in the selected groups.

The temporal distribution of these intervention events is shown in Figure 2(b). Bars aggregate ten training updates at a time for the 1.5B ALFWorld run. They report the proportion of all training groups selected as eligible zero-signal groups and the proportion that produce a verified teacher recovery; they are process diagnostics rather than final benchmark metrics.

E.1 Intervention Length

Sparse intervention and short intervention are distinct properties. To measure the latter directly, we analyze the same 7B training runs over all 150 updates. For each retained relay, let L be the number of actions in the verified full teacher recovery from the selected anchor, $j^\star$ the number retained in the teacher bridge, and R the number in the selected successful student resume suffix. Lengths count executed environment actions, not tokens, and exclude the student prefix before the anchor. We recover R from the successful independent-confirmation continuation selected by the training procedure: shortest action length, with output tokens and seed as tie-breakers. The retained teacher fraction is $j^\star/L$, and $L - j^\star$ counts teacher actions omitted from bridge supervision.

Table 9 and Figure 3 report these paired measurements for the 18 ALFWorld and 10 WebShop relays in Table 8. Recoveries without a valid retained relay are excluded, rather than treated as zero-length interventions. For context, across all 25 and 29 verified recoveries, respectively, the mean full teacher length is 13.44 and 4.48 actions (medians: 12 and 5).

Table 9: Intervention lengths for retained relays in the 7B training runs. Paired entries are mean / median over relays; lengths are action counts. The pooled omitted fraction weights each relay by its full teacher length.

Metric	ALFWorld	WebShop
Retained relays	18	10
Full teacher recovery L	14.78 / 14.00	5.20 / 5.00
Retained teacher bridge $j^\star$	9.39 / 9.50	1.80 / 2.00
Student resume suffix R	4.33 / 3.00	4.10 / 4.00
Retained teacher fraction $j^\star/L$ (%)	60.68 / 65.94	35.33 / 33.33
Omitted teacher actions $L - j^\star$	5.39 / 4.50	3.40 / 4.00
Total omitted / full teacher actions	97 / 266	34 / 52
Pooled omitted fraction (%)	36.47	65.38

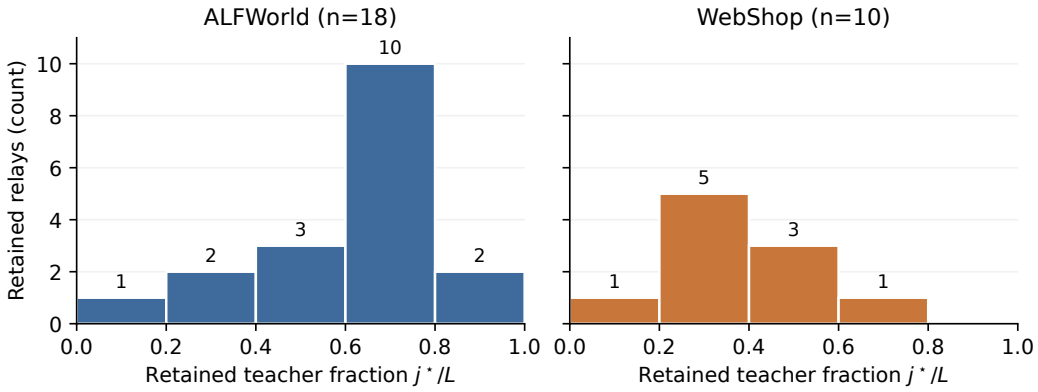

Figure 3: Distribution of the retained teacher fraction $j^\star/L$ for valid retained relays in the 7B runs. Each relay contributes one observation; bins have width 0.2 and include the left but not the right boundary (the final bin includes 1). Smaller values indicate earlier handoff relative to the verified teacher recovery.

Across retained relays, handoff removes 97 of 266 teacher actions on ALFWorld and 34 of 52 on WebShop from bridge supervision, corresponding to pooled reductions of 36.47% and 65.38%. The mean per-relay reductions, $1 - \mathrm{mean}(j^\star/L)$, are 39.32% and 64.67%, respectively. Thus, teacher supervision is shortened within individual relays, beyond the sparsity of intervention events. The distribution also shows that the bridge is not uniformly short: its required length varies across tasks and benchmarks. Here, *minimal* means the earliest independently confirmed handoff found within the finite search budget, not a globally shortest possible intervention. The omitted actions were already generated to verify the teacher recovery; these reductions measure teacher supervision removed, not generation saved. Generation costs, including discarded continuations, remain accounted for separately in Table 5.