REWIND-OPD: DISCREPANCY-GUIDED TRAJECTORY ROLLBACK FOR ON-POLICY DISTILLATION IN AGENTIC TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) provides dense teacher supervision on student-generated trajectories. In agentic tasks, an early divergent action can change subsequent observations, allowing teacher–student discrepancies to accumulate across turns. Existing intervention methods typically act after this drift has become apparent, when the current state may already reflect an earlier mistake. We propose REWIND-OPD, which uses the history of teacher–student discrepancies to select an earlier intervention point, then rolls back to that state and reconstructs the trajectory under teacher guidance before returning control to the student. Across three teacher–student pairs and three agentic benchmarks, REWIND-OPD achieves the best performance among six baselines. With a general 30B teacher, it improves success rate over standard OPD by 4.14 and 4.66 percentage points for 1.7B and 4B students, while reducing average output tokens by 16.0% and 8.0%, respectively. With specialized 8B teachers, the 4B student exceeds the strongest baseline and teachers in average success rate by 1.72 and 1.49 percentage points, respectively, supporting its effectiveness and generalizability.

1 INTRODUCTION

On-policy distillation (OPD) trains a student on its own rollouts using token-level supervision from a teacher (Agarwal et al., 2024). It has been used to transfer capabilities to smaller models (Yang et al., 2025), recover capabilities from earlier training stages (GLM-5 Team, 2026), and combine expertise from specialized teachers (DeepSeek-AI et al., 2026). Unlike sparse outcome-level rewards, teacher supervision at each visited prefix provides dense feedback (Lu & Thinking Machines Lab, 2025). However, this feedback can become less reliable as student-generated prefixes drift away from those induced by the teacher (Li et al., 2026b).

This issue is amplified in agentic interactions, where a divergent action changes subsequent observations and decision contexts, allowing teacher–student discrepancies to cascade across turns (Zhong et al., 2026; Wang et al., 2026). Some methods reweight training signals to control credit assignment: SOD (Zhong et al., 2026) downweights steps with large teacher–student divergence, while TurnOPD (Zhou et al., 2026) interpolates between trajectory-level and turn-level loss normalization to prevent deeper turns from being underweighted. Other methods modify rollout construction through teacher intervention. Relay-OPD (Xu et al., 2026) inserts teacher segments at discrepant reflection tokens; Guided-OPD (Li et al., 2026a) selects teacher or student generation according to a scheduled teacher probability; and DASH-OPD (Xia et al., 2026) accumulates discrepancy as drift and recovery evidence, switching control when either exceeds its threshold.

Reweighting methods change how sampled trajectories are used for training but leave suboptimal trajectories unchanged. Existing teacher-intervention methods can change the continuation, yet begin teacher generation at the selected turn without revisiting earlier student actions or their consequences. The left panels of Figure 1 illustrate both limitations: the student-only trajectory fails, while the teacher finds it hard to revise the earlier product choice from the current state. This suggests intervening at the state preceding the action responsible for the discrepancy. On the right, rollback allows the teacher to revise the action and complete the task.

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

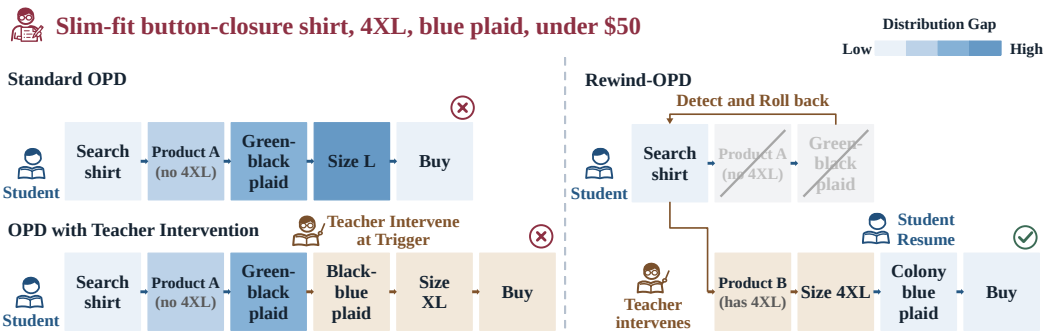

Figure 1: Comparison of rollout strategies in OPD. **Top left:** The student selects a product without the required size and fails to complete the task. **Bottom left:** Teacher intervention begins after a large discrepancy is detected. The teacher changes the color option but continues with the unsuitable product. **Right:** REWIND-OPD returns to an earlier state and produces a successful continuation with a different product.

Inspired by this observation, we propose REWIND-OPD, a discrepancy-guided online distillation method that selects an earlier intervention point and rolls back to it for teacher-guided trajectory reconstruction. The method has two components. The discrepancy-based switching controller detects when sustained teacher–student divergence calls for intervention, and online rollback with Median-Centered CUSUM Onset Localization (MCOL) uses historical student-gap scores to estimate a candidate intervention point, then lets the teacher reconstruct the trajectory until recovery returns control to the student. We evaluate the method on ALFWorld, WebShop, and ScienceWorld using a general Qwen3-30B-A3B teacher and task-specialized Qwen3-8B teachers. With the general teacher, REWIND-OPD improves average success rate over Vanilla OPD by 4.14 and 4.66 percentage points for 1.7B and 4B students, respectively, while reducing average output tokens by 16.0% and 8.0% and interaction rounds by 5.52% and 6.11%. With specialized teachers, the 4B student exceeds the strongest baseline by 1.72 percentage points in overall success rate and surpasses the teachers by 1.49 percentage points. Beyond the overall performance gains, the onset localization method MCOL outperforms the no-rewind baseline and alternative methods in standalone evaluations, supporting its effectiveness in selecting where teacher intervention should begin.

Our contributions are threefold:

- We identify the state at which teacher intervention begins as a key factor in trajectory construction and show that rolling back to an earlier state can improve training data quality.
- We propose REWIND-OPD, an online distillation framework that estimates a candidate intervention point from discrepancy history and rolls back for teacher-guided reconstruction.
- Across three agentic benchmarks and three teacher–student settings, REWIND-OPD consistently improves task performance while reducing the average number of interaction rounds and generated response tokens.

2 RELATED WORK

2.1 KNOWLEDGE DISTILLATION AND ON-POLICY DISTILLATION

Knowledge distillation transfers teacher capabilities to a student by matching the student's distribution to that of the teacher (Hinton et al., 2015). Although effective at transferring capabilities to smaller models, off-policy distillation can cause a mismatch between the training-data distribution and the student's own state distribution (Bengio et al., 2015). On-policy distillation (OPD) methods query the teacher on student-generated rollouts and provide dense token-level supervision through reverse KL divergence (Agarwal et al., 2024). Recent work further extends this paradigm to self-distillation, where the student conditioned on privileged information serves as a teacher branch to guide the same model toward self-improvement (Hübotter et al., 2026; Zhao et al., 2026).

2.2 AGENTIC ON-POLICY DISTILLATION

In agentic tasks, actions change subsequent observations and thus the contexts in which supervision is provided. This causes the teacher–student distribution mismatch to accumulate over the trajectory, making reliable teacher guidance increasingly difficult (Zhong et al., 2026). One line of work combines on-policy self-distillation with GRPO-style reinforcement learning: SDAR (Lu et al., 2026) gates privileged-context distillation as an auxiliary objective, while SEED (Wu et al., 2026) distills action-probability shifts induced by evolving hindsight skills. Another line directly optimizes multi-turn OPD through temporal curricula, divergence-based weighting, or turn-level rollout and loss allocation (Wang et al., 2026; Zhong et al., 2026; Zhou et al., 2026). Unlike these methods, REWIND-OPD modifies the interaction trajectory on which subsequent supervision is collected, thereby improving the quality of the training data itself.

2.3 TEACHER-INTERVENED TRAJECTORY GENERATION

Teacher intervention can narrow teacher–student distribution gaps, thereby improving both trajectory quality and training signals. Heuristic schedules include TCOD-B2F (Wang et al., 2026), which progressively shortens successful teacher prefixes, and Guided-OPD (Li et al., 2026a), which anneals the probability of teacher-generated turns. Among discrepancy-driven methods, FutureBridge-OPD (Chen et al., 2026) improves collected trajectories offline by locating the turn with the largest teacher–student discrepancy then insert teacher segments. Relay-OPD (Xu et al., 2026) triggers teacher takeover during generation at the discrepancy of reflection token. DASH-OPD (Xia et al., 2026) accumulates drift and recovery evidence and uses hysteresis for bidirectional switching between student and teacher. REWIND-OPD combines online switching with retrospective onset localization: once intervention is triggered, it uses the preceding student-gap history to select an earlier state from which teacher guidance begins.

3 PRELIMINARIES

3.1 ON-POLICY DISTILLATION ON AGENTIC TASKS

Multi-turn Agentic Tasks. For agentic tasks, we model interaction as a finite-horizon Partially Observable Markov Decision Process (Kaelbling et al., 1998) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P, \Omega, \rho_0, H)$, where $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{O}$ are the state, action, and observation spaces; P and Ω are the transition and observation kernels; ρ_0 is the initial-state distribution; and H is the maximum number of interaction turns.

In text-based LLM agent settings, the policy conditions on the instruction I and interaction history, $h_t = (I, o_0, y_0, o_1, \ldots, y_{t-1}, o_t)$. At turn t, the student generates a response $y_t = (r_t, a_t) \sim \pi_\theta(\cdot \mid h_t)$ containing reasoning r_t and an action command, which a parser extracts as $a_t = \mathrm{Parse}(y_t) \in \mathcal{A}$. The environment then transitions according to $s_{t+1} \sim P(\cdot \mid s_t, a_t)$ and returns $o_{t+1} \sim \Omega(\cdot \mid s_{t+1})$. The trajectory $\tau_{student}$ is written as $(I, o_0, r_0, a_0, \ldots, r_{H_\tau-1}, a_{H_\tau-1}, o_{H_\tau})$, where $H_\tau \leq H$ allows for early termination.

Agentic OPD. Agentic OPD extends standard OPD to each turn of a trajectory. At each turn, the student and teacher are conditioned on the same interaction history h_t, which incorporates previous student-generated responses and the resulting environment observations. The objective is written as Eq. (1).

$$\mathcal{L}_{\mathrm{Agentic-OPD}}(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{H_\tau - 1} D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid h_t) \| \pi_T(\cdot \mid h_t)\big)\right] \tag{1}$$

3.2 INTERLEAVED GENERATION WITH A TEACHER MODEL

Interleaved Trajectory. Let $\sigma_t \in \{S, T\}$ indicate whether the student or teacher generates turn t, as specified in Eq. (2). The resulting trajectory is $\tau = (h_0, y_0^{\sigma_0}, \ldots, h_{H_\tau-1}, y_{H_\tau-1}^{\sigma_{H_\tau-1}})$. Both models condition on the same interaction history, and responses from either executor contribute to the next-turn context.

$$y_t^{\sigma_t} = (r_t^{\sigma_t}, a_t^{\sigma_t}) \sim \begin{cases} \pi_\theta(\cdot \mid h_t), & \sigma_t = S, \\ \pi_T(\cdot \mid h_t), & \sigma_t = T. \end{cases} \tag{2}$$

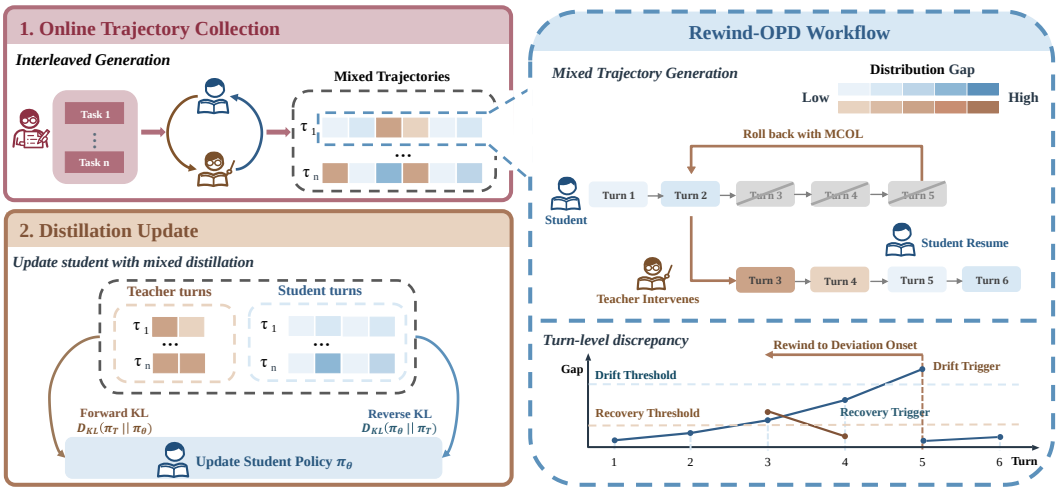

Figure 2: Overview of REWIND-OPD. **Top left:** Online collection of mixed student–teacher trajectories. **Bottom left:** Distillation updates using student- and teacher-generated turns. **Right:** An enlarged generation sequence illustrates how a drift trigger leads to rollback to an earlier estimated onset, followed by teacher generation and a return to the student after recovery.

Discrepancy Measure between Student and Teacher. Recent work uses the probability gap to measure distribution discrepancy (Zhong et al., 2026; Xia et al., 2026). Denote π_{σ_t} and $\pi_{\overline{\sigma_t}}$ as the t-th step generator and scorer. For y_t, the binary mask M_t selects action tokens, and $|M_t|$ denotes the number of valid tokens. Then the teacher–student discrepancy is expressed as the average log-probability gap on action tokens, as shown in Eq. (3).

$$d_t^{\sigma} = \frac{1}{|M_t|} \sum_i M_{t,i} \big[\log \pi_{\sigma_t}(y_{t,i} \mid h_t, y_{t,<i}) - \log \pi_{\overline{\sigma_t}}(y_{t,i} \mid h_t, y_{t,<i}) \big], \quad y_t \sim \pi_{\sigma_t}(\cdot \mid h_t) \tag{3}$$

Training Objective of Interleaved Trajectories. While the discrepancy signal determines intervention, the interleaved trajectory can provide supervision from both executors, i.e., reverse KL for student-generated turns and forward KL for teacher-generated turns. The joint objective is shown in Eq. (4), where $\mathbb{1}_t^S = \mathbb{1}\{\sigma_t = S\}$ and $\mathbb{1}_t^T = \mathbb{1}\{\sigma_t = T\}$ are the indicator functions for student and teacher turns.

$$\begin{aligned} \mathcal{L}_{\text{Agentic-OPD}}(\theta) &= \mathbb{E}_\tau \left[\sum_{t=0}^{H_\tau - 1} D_{\text{KL}}(\pi_{\sigma_t}(\cdot \mid h_t) \| \pi_{\overline{\sigma_t}}(\cdot \mid h_t)) \right] \\ &= \mathbb{E}_\tau \Big[\sum_{t=0}^{H_\tau - 1} \big(\mathbb{1}_t^S D_{\text{KL}}(\pi_\theta(\cdot \mid h_t) \| \pi_T(\cdot \mid h_t)) + \mathbb{1}_t^T D_{\text{KL}}(\pi_T(\cdot \mid h_t) \| \pi_\theta(\cdot \mid h_t)) \big) \Big] \end{aligned} \tag{4}$$

4 REWIND-OPD: DISCREPANCY-GUIDED TRAJECTORY ROLLBACK

4.1 OVERVIEW

Figure 2 illustrates the two stages of REWIND-OPD. During online trajectory collection, a discrepancy-based controller accumulates evidence of drift and recovery to determine when to switch between student and teacher generation (Section 4.2). When student drift triggers intervention, MCOL estimates a candidate intervention point, and the interaction is rolled back to it for teacher-guided reconstruction (Section 4.3). During the distillation update, student- and teacher-generated turns contribute different terms to the training objective (Section 4.4).

4.2 Discrepancy-based Switching Controller

We use the distribution gap d_t^σ defined in Eq. (3) to control teacher intervention and recovery, where t indexes interaction turns and $\sigma_t \in \{S, T\}$ identifies the active executor (student or teacher). Following Xia et al. (2026), we maintain separate running statistics for each executor using Welford's algorithm (Welford, 1962). For $\sigma \in \{S, T\}$, let N_{t-1}^σ, μ_{t-1}^σ, and v_{t-1}^σ denote the count, mean, and variance of its previously observed gaps. The normalized gap is

$$\widetilde{d}_t^\sigma = \begin{cases} 0, & N_{t-1}^\sigma = 0, \\ \dfrac{d_t^\sigma - \mu_{t-1}^\sigma}{\max\{\sqrt{v_{t-1}^\sigma}, \varepsilon\}}, & N_{t-1}^\sigma > 0. \end{cases} \tag{5}$$

Here $\varepsilon > 0$ is a scale floor; the statistics are updated only after scoring the current gap. $N_{t-1}^\sigma = 0$ identifies the executor's first turn, for which the normalized gap is set to zero. Starting from zero, the evidence for the active executor $\sigma = \sigma_t$ is updated as follows:

$$E_t^\sigma = \begin{cases} \max\{0, E_{t-1}^S + \widetilde{d}_t^S\}, & \sigma = S, \\ \max\{0, E_{t-1}^T - \widetilde{d}_t^T\}, & \sigma = T. \end{cases} \tag{6}$$

Student gaps above their running mean accumulate drift evidence, whereas teacher gaps below their running mean accumulate recovery evidence. We group each executor's running statistics and evidence into a controller state $W_\sigma = (N^\sigma, \mu^\sigma, v^\sigma, E^\sigma)$. Only W_{σ_t} is updated on turn t; the other executor's state remains unchanged. The Boolean flag switch_t^σ determines whether to switch away from the active executor:

$$\mathrm{switch}_t^\sigma = \begin{cases} E_t^S > \eta_S, & \sigma = S, \\ E_t^T > \eta_T, & \sigma = T, \end{cases} \qquad \eta_\sigma = q_\sigma H. \tag{7}$$

The evidence threshold η_σ is set by the dimensionless coefficient q_σ and maximum interaction horizon H. A switch resets only the departing executor's evidence; both executors' Welford statistics persist across switches and rollbacks within a rollout. A student-side trigger initiates rollback as described next; a teacher-side trigger returns control to the student. Detailed settings are provided in Appendix C.2.

4.3 Online Rollback with Median-centered CUSUM Onset Localization

A switching trigger identifies pronounced disagreement, but not where the drift began. Median-centered CUSUM Onset Localization (MCOL) uses the student-gap history to select an earlier turn from which teacher guidance should begin. Suppose the current student segment spans turns i through t, with intervention triggered after turn t. Its gap history is $\mathcal{G}_t = (d_i^S, \ldots, d_t^S)$, where d_j^S is the gap at turn j. To reduce the reference's sensitivity to isolated extreme gaps, we use the median and median absolute deviation (MAD) (Rousseeuw & Croux, 1993):

$$\mu_t^{\mathrm{loc}} = \mathrm{median}(\mathcal{G}_t), \qquad \mathrm{MAD}_t = \mathrm{median}_{i \le j \le t} |d_j^S - \mu_t^{\mathrm{loc}}|. \tag{8}$$

We then standardize each gap in the segment, $i \le j \le t$, using a common scale:

$$\sigma_t^{\mathrm{loc}} = \max\{c_{\mathrm{MAD}} \mathrm{MAD}_t, \varepsilon_{\mathrm{loc}}\}, \qquad u_j^{(t)} = \frac{d_j^S - \mu_t^{\mathrm{loc}}}{\sigma_t^{\mathrm{loc}}}. \tag{9}$$

Here c_{MAD} is the normal-consistency calibration for MAD, while $\varepsilon_{\mathrm{loc}} > 0$ ensures numerical stability. MCOL then compares cumulative excess gaps with the cumulative-sum principle (CUSUM; Page, 1954) across candidate suffixes ending at the triggering turn t:

$$k^* = \operatorname*{arg\,max}_{k \in \mathcal{K}_t} \sum_{j=t-k+1}^{t} \left(u_j^{(t)} - \delta\right), \qquad \mathcal{K}_t = \{1, \ldots, t-i+1\}, \tag{10}$$

where the allowance $\delta \geq 0$ imposes a fixed penalty per rewound turn, and $\mathcal{K}_t$ includes all feasible rewind depths. k^* is the selected rewind depth yielding the estimated intervention turn $j^* = t - k^* + 1$, where teacher-guided regeneration begins.

4.4 Training Objective and Algorithm

Initialization with Teacher. To encourage more teacher participation at the beginning of training, we assign the first turn to the teacher with an annealing probability of $p_T(s) = 1 - s/S_{\max}$, where s and $S_{\max}$ are the current and total training steps.

Training Target. We collect student- and teacher-generated turns in buffers $\mathcal{D}_S$ and $\mathcal{D}_T$ until task termination or horizon H, with $\tau_{\text{rewind}} = \mathcal{D}_S \cup \mathcal{D}_T$. For eligible response-token positions $i \in \mathcal{I}_t$, write $p_{\theta,t,i} = \pi_\theta(y_{t,i} \mid h_t, y_{t,<i})$, and similarly for the teacher and frozen rollout student. The token-level objective combines a clipped PPO-style student surrogate with teacher log-likelihood (Schulman et al., 2017):

$$\begin{aligned} \mathcal{L}_{\text{Rewind}}(\theta) &= -\sum_{t=0}^{H_\tau - 1} \sum_{i \in \mathcal{I}_t} \Big[\mathbb{1}_t^S \ell_{t,i}^S + \mathbb{1}_t^T \log p_{\theta,t,i} \Big], \\ \ell_{t,i}^S &= \min\{\rho_{t,i} A_{t,i}^S, \text{clip}(\rho_{t,i}, 1-\epsilon, 1+\epsilon) A_{t,i}^S\}. \end{aligned} \tag{11}$$

Here $\rho_{t,i} = p_{\theta,t,i}/p_{\bar{\theta},t,i}$ and $A_{t,i}^S = \log p_{T,t,i} - \log p_{\bar{\theta},t,i}$ are token-level ratios and advantages; teacher advantages are one. ϵ is the clipping parameter, and H_τ counts training turns. Appendix A provides the algorithm and subroutines.

5 Experiments

5.1 Experimental Setup

Models, Benchmarks and Metrics. We conduct experiments on three long-horizon agentic tasks: ALFWorld (Shridhar et al., 2021), WebShop (Yao et al., 2022), and ScienceWorld (Wang et al., 2022). We report environment scores (Score), success rate (SR), and interaction rounds (Rnd). For ALFWorld and WebShop, we report the mean and standard deviation across three random seeds. We use models from the Qwen3 series (Yang et al., 2025), with Qwen3-30B-A3B and post-trained Qwen3-8B as teachers and Qwen3-1.7B and Qwen3-4B as students. A detailed description of the benchmarks and metrics is provided in Appendix B.

Training and Evaluation Settings. We implement different methods using the Trinity-RFT framework (Pan et al., 2025). Following previous setups (Wang et al., 2026), we train each student for 250 steps on ALFWorld and ScienceWorld and 150 steps on WebShop. We use the PPO-style (Schulman et al., 2017) training pipeline with rollout and update batch sizes of 16 and 64, respectively. We limit input and output lengths to 10,240 and 512 tokens, respectively, and use a sampling temperature of 1.0. For evaluation, we limit input and output lengths to 10,240 and 4,096 tokens, respectively, and use a sampling temperature of 0.4. We allow at most 30 interaction rounds on ALFWorld and ScienceWorld and 15 on WebShop. Full settings and the prompt protocol are provided in Appendices C and F.

Configurations of Different Methods. We compare REWIND-OPD with six baselines spanning three categories: (i) Zero-shot, the untrained model; (ii) standard on-policy distillation and its variants, including Vanilla OPD (Agarwal et al., 2024), TCOD-F2B (Wang et al., 2026), and TurnOPD (Zhou et al., 2026); (iii) Teacher-guidance methods, including Guided-OPD (Li et al., 2026a) and DASH-OPD (Xia et al., 2026). We reproduce each baseline using the settings in its original paper, while keeping shared training and evaluation settings consistent. For REWIND-OPD, we set the drift and recovery threshold coefficients to $q_{\text{on}} = 0.1$ and $q_{\text{off}} = 1/30$. Repetition of an action or observation for three turns also triggers teacher takeover without rollback. The minimum teacher span is $\ell_{\min} = 2$ turns. MCOL uses the default standardization coefficient $c_{\text{MAD}} = 1.4826$ and $\delta = 0$, $\varepsilon_{\text{loc}} = 10^{-3}$, with $K_{\max} = 5$ and 3 for the 1.7B and 4B students, respectively. Appendix C.3 documents the detailed configuration of each method.

Table 1: Distillation from a general Qwen3-30B-A3B teacher to Qwen3-1.7B and 4B students. ALFWorld reports success rate (SR); WebShop and ScienceWorld report score and SR; Rnd denotes interaction rounds. **Bold** and underlined values denote the best and second-best performance, respectively. Light blue highlights trained-student results matching or exceeding the teacher.

	ALFWorld						WebShop			ScienceWorld		
	Seen		Unseen		Hard							
Method	SR↑	Rnd↓	SR↑	Rnd↓	SR↑	Rnd↓	Score↑	SR↑	Rnd↓	Score↑	SR↑	Rnd↓
Teacher: Qwen3-30B-A3B												
Zero-shot	44.0	22.8	39.1	23.5	21.8	27.7	41.6	15.3	7.2	50.6	22.6	21.1
Student: Qwen3-1.7B												
Zero-shot	2.4	29.5	3.5	29.2	0.3	30.0	3.7	0.7	14.4	10.8	1.8	16.7
Vanilla OPD	26.0	25.4	30.3	24.8	10.7	28.7	37.6	12.0	7.3	32.3	7.6	25.5
TCOD-F2B	26.4	25.5	34.1	24.6	11.6	28.6	39.5	11.0	6.8	36.1	11.2	24.3
Guided-OPD	26.9	25.7	30.3	25.1	12.1	28.7	37.0	11.0	7.5	32.1	10.2	25.2
TurnOPD	26.0	25.2	33.1	24.3	10.2	**28.5**	40.3	10.3	7.0	32.5	9.3	24.9
DASH-OPD	28.6	25.4	33.6	24.4	11.0	28.7	38.9	12.0	6.9	36.6	12.0	23.8
REWIND-OPD	**31.9**	**24.3**	**36.6**	**24.2**	**14.3**	**28.5**	**42.9**	**14.0**	**6.5**	**41.5**	**12.7**	**23.7**
Student: Qwen3-4B												
Zero-shot	30.5	25.0	37.6	24.0	9.6	29.0	28.6	10.0	10.5	39.8	15.9	19.5
Vanilla OPD	35.0	24.2	35.3	24.0	16.3	28.7	37.4	12.7	7.7	46.0	18.7	22.4
TCOD-F2B	35.0	24.0	38.1	23.4	17.6	28.0	39.9	15.0	7.7	46.9	20.1	21.8
Guided-OPD	36.4	23.9	39.3	24.1	17.4	28.3	38.4	14.3	7.8	46.1	19.7	21.1
TurnOPD	36.2	23.9	34.6	24.6	17.9	28.2	36.0	13.0	8.3	45.3	17.2	22.5
DASH-OPD	38.1	23.9	37.8	24.2	19.3	28.1	40.9	13.7	7.3	51.0	22.2	21.4
REWIND-OPD	**39.8**	**23.3**	**42.8**	**23.1**	**22.3**	**27.8**	**42.1**	**16.7**	**7.1**	**52.0**	**22.6**	**20.5**

5.2 MAIN RESULTS WITH A GENERAL TEACHER

As shown in Table 1, REWIND-OPD achieves the best task performance across all three benchmarks and both student sizes. After weighting the ALFWorld splits by task count (140/134/121) and averaging across benchmarks, it improves mean SR over Vanilla OPD by 4.14 and 4.66 percentage points for the 1.7B and 4B students, respectively. The corresponding gains over DASH-OPD, the strongest baseline by mean SR, are 1.97 and 2.21 percentage points. REWIND-OPD also achieves the fewest mean interaction rounds among distillation methods, reducing them by 5.52% and 6.11% relative to Vanilla OPD, and by 1.79% and 3.29% relative to DASH-OPD, respectively.

On ALFWorld Unseen, the 4B student exceeds the teacher in SR (42.8% vs. 39.1%), while the 1.7B student reaches 36.6%. The 4B student also exceeds the teacher on the harder Hard split (22.3% vs. 21.8%). On WebShop, both students achieve higher scores with fewer interaction rounds than the teacher, consistent with more efficient trajectories. On ScienceWorld, the 4B student surpasses the teacher in score (52.0 vs. 50.6), matches its SR (22.6%), and uses fewer rounds (20.5 vs. 21.1).

5.3 MAIN RESULTS WITH SPECIALIZED TEACHERS

To evaluate REWIND-OPD with task-specialized teachers, we train separate Qwen3-8B models on ALFWorld and WebShop using GiGPO and distill them into Qwen3-4B students; Appendix C details the settings. We micro-average ALFWorld by split size, then macro-average the two benchmarks. In Table 2, REWIND-OPD achieves the highest overall SR and the fewest average rounds. Relative to Vanilla OPD and Guided-OPD, the strongest baseline in average SR, it improves SR by 12.88 and 1.72 percentage points and reduces rounds by 15.29% and 6.33%, respectively. It also exceeds the teachers by 1.49 percentage points in SR while reducing rounds by 1.51% (0.16 rounds). This aggregate advantage does not hold on every benchmark: ALFWorld weighted SR is 70.6%, versus 71.0% for the teacher, whereas WebShop SR improves from 31.3% to 34.7%, alongside a 4.2-point score gain. These results support the effectiveness of rollback-based teacher intervention with specialized teachers.

Table 2: Distillation from post-trained Qwen3-8B teachers to Qwen3-4B students. The teachers are trained with GiGPO on ALFWorld and WebShop separately. **Bold** and underlined values denote the best and second-best performance. Light blue highlights student results matching or exceeding the teacher.

	ALFWorld						WebShop			Overall	
	Seen		Unseen		Hard						
Method	SR↑	Rnd↓	SR↑	Rnd↓	SR↑	Rnd↓	Score↑	SR↑	Rnd↓	SR↑	Rnd↓
Teacher: Qwen3-8B-GIGPO											
Zero-shot	86.2	11.3	82.3	13.2	41.0	23.2	58.1	31.3	5.2	51.2	10.4
Student: Qwen3-4B											
Zero-shot	30.5	25.0	37.6	24.0	9.6	29.0	28.6	10.0	10.5	18.3	18.2
Vanilla OPD	66.2	14.9	65.4	15.3	8.5	28.4	58.9	31.3	5.0	39.8	12.1
TCOD-F2B	71.4	13.9	71.4	14.2	10.5	28.0	60.7	32.3	4.8	42.5	11.6
Guided-OPD	82.1	12.1	81.1	13.0	43.3	23.7	57.2	32.0	5.9	50.9	10.9
TurnOPD	69.0	14.7	64.2	15.9	14.0	27.5	61.4	33.0	**4.6**	41.8	11.8
DASH-OPD	80.5	12.1	78.4	13.2	41.6	23.6	57.5	31.7	5.1	49.8	10.5
REWIND-OPD	**82.6**	**12.0**	**82.3**	**12.9**	**43.8**	**23.0**	**62.3**	**34.7**	4.8	**52.7**	**10.2**

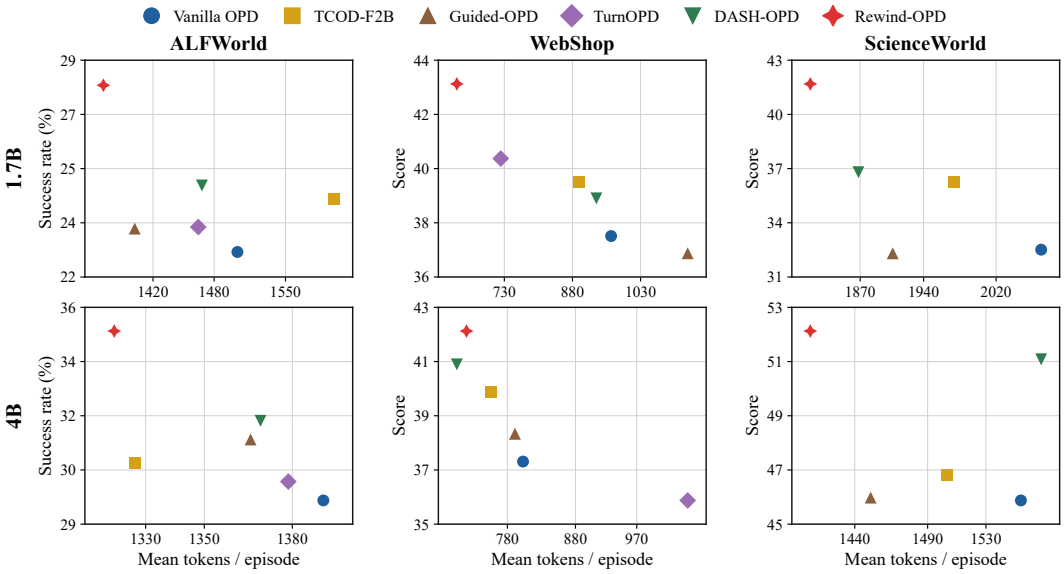

Figure 3: Task performance versus average output tokens per episode for each OPD method. ALFWorld metrics are pooled across Seen, Unseen, and Hard by task counts (140/134/121). Higher performance with fewer tokens (upper left) is preferable.

5.4 PERFORMANCE–EFFICIENCY TRADE-OFF

Figure 3 compares task performance against average output tokens per episode. Across the three benchmarks, REWIND-OPD reduces average token usage relative to Vanilla OPD by 16.0% and 8.0% for the 1.7B and 4B students, respectively, and by 10.7% and 5.3% relative to DASH-OPD, the strongest baseline in average SR. REWIND-OPD lies on the Pareto frontier in all six settings. Compared with DASH-OPD, REWIND-OPD improves both performance and token usage in five settings; on WebShop at 4B, its higher performance uses 2.0% more tokens.

5.5 ABLATION STUDIES

The ablation results are shown in Table 3, where the full method achieves the highest task performance and efficiency across all evaluated settings. Removing teacher intervention recovers Vanilla OPD and degrades performance across both benchmarks, supporting the value of teacher-

Table 3: Ablation results with the Qwen3-30B-A3B teacher and Qwen3-4B student. **Bold** and underlined values denote the best and second-best means among evaluated variants.

	ALFWorld						WebShop		
	Seen		Unseen		Hard				
Method	SR↑	Rnd↓	SR↑	Rnd↓	SR↑	Rnd↓	Score↑	SR↑	Rnd↓
w/o teacher intervention (Vanilla OPD)	35.0	24.2	35.3	24.0	16.3	28.7	37.4	12.7	7.7
w/o rewind (DASH-OPD)	38.1	23.9	37.8	24.2	19.3	28.1	40.9	13.7	7.3
w/o rewind, small q_{on}	36.2	24.0	35.3	24.2	16.5	28.4	38.8	13.0	7.6
w/ rewind, depth $k = 1$	32.4	24.6	41.0	23.2	17.1	28.2	40.0	13.3	7.2
w/ rewind, depth $k = 2$	35.5	23.8	37.8	23.7	14.9	28.5	40.9	15.0	**7.1**
REWIND-OPD	**39.8**	**23.3**	**42.8**	**23.1**	**22.3**	**27.8**	**42.1**	**16.7**	**7.1**

guided trajectory collection. Removing rewinding while retaining intervention with switch threshold $q_{\text{on}} = 1/10$ yields DASH-OPD. Lowering this threshold to $q_{\text{on}} = 1/30$ produces mixed results and remains worse than the full method. Thus, earlier intervention alone does not recover the benefits of rewinding. We then replace MCOL with fixed depths $k \in \{1, 2\}$, where $k = 1$ rejects only the triggering action and $k = 2$ additionally reverses one preceding executed action. Both fixed-depth variants underperform the adaptive method. The gap is pronounced on ALFWorld Hard, where adaptive rewinding improves SR by 5.2 and 7.4 percentage points, supporting the importance of adapting the rewind depth.

5.6 COMPARISON OF ONSET LOCALIZATION METHODS

We compare four localization rules in a diagnostic setting: MCOL, Causal-EMA (the suffix after the latest in-band turn), Positive-Run (the terminal run of positive normalized gaps), and Evidence-Onset (the current positive accumulated-drift episode). Diagnostic MCOL uses pre-trigger history; the main method uses the current student segment, including the triggering turn. Appendix C.4 gives implementation details.

Figure 4 shows that MCOL ranks first among the four onset localization methods on the pooled Seen/Unseen group, the Hard split, and the WebShop validation set. On the ALFWorld Hard split, MCOL improves SR by 33.3% relative to the no-rewind counterpart, supporting discrepancy-guided selection of an earlier intervention point. Every alternative also improves over no rewind, suggesting that adaptive rollback is beneficial.

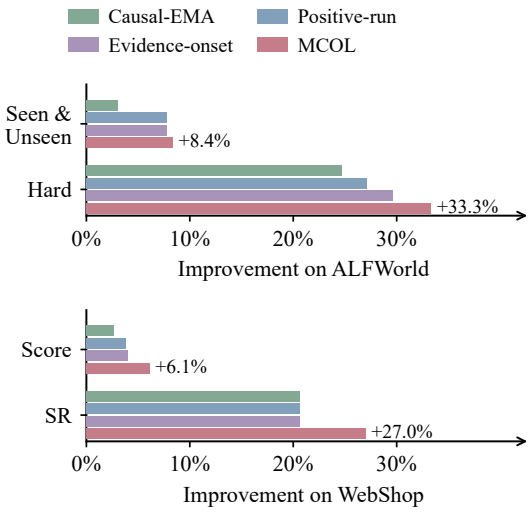

Figure 4: Comparison of different onset localization methods on ALFWorld and WebShop.

6 CONCLUSION

Our results support discrepancy-guided selection of an earlier intervention point in agentic OPD. REWIND-OPD estimates a candidate point from discrepancy history and rolls back the environment and interaction history for teacher-guided reconstruction. Across three benchmarks and three teacher–student pairs, it outperforms five competitive baselines in task performance. With the general teacher, the 1.7B and 4B students respectively gain 4.14 and 4.66 percentage points in average SR over Vanilla OPD, with 16.0% and 8.0% fewer tokens and 5.52% and 6.11% fewer rounds. With specialized teachers, average SR exceeds the strongest baseline by 1.72 percentage points. These results support the effectiveness and generalizability of rollback-based teacher intervention.

AI USE STATEMENT

We used an LLM to assist with literature review, manuscript organization, drafting and language editing. The original research idea and core method were developed by the authors, who remain responsible for all scientific decisions and for verifying the manuscript and references. The tool was not used to generate experimental data or reported numerical results. The authors take full responsibility for the final work.

ETHICS STATEMENT

All experiments use open-source datasets and involve no personally identifiable information or sensitive content. This research aims to improve agent training and efficiency through on-policy distillation. Distillation may transfer the teacher's errors and biases alongside useful behavior, so teacher intervention does not guarantee correctness. We report teacher usage and compute costs for transparency and use all models, datasets, and environments in accordance with their applicable licenses and terms.

REPRODUCIBILITY STATEMENT

Section 5.1 and Appendix C specify the benchmarks, model pairs, baselines, training protocol, controller configuration, and evaluation metrics used in our study. The final supplementary material will provide complete pseudocode, prompts, task splits, all method-specific hyperparameters, random seeds, hardware and software versions, and per-run results, together with the ablations needed to reconstruct each reported table and figure. We will also submit an anonymized implementation and evaluation scripts as supplementary code, with any external service dependencies documented explicitly.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

Chishui Chen, Yaoyou Fan, Te Sun, Yi Yang, Chenghao Sun, Delin Mao, Hongbo Qiao, Zuowei Zhang, Junxi Wang, Chenxing Sun, Yangen Hu, Lu Pan, Xuyang Liu, and Linfeng Zhang. Look Ahead Before You Distill: Future Trajectory Validation of Teacher Guidance for Agentic On-Policy Distillation, 2026. URL https://arxiv.org/abs/2608.01953.

DeepSeek-AI et al. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

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

GLM-5 Team. GLM-5: from Vibe Coding to Agentic Engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, 2015. URL https://arxiv.org/abs/1503.02531.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement Learning via Self-Distillation, 2026. URL https://arxiv.org/abs/2601.20802.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. doi: 10.1145/3600006.3613165.

Gengsheng Li, Mao Zheng, Mingyang Song, Ruiqi Liu, Tianyu Yang, Jie Sun, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Dan Zhang, and Jinqiao Wang. On-Policy Distillation with Curriculum Turn-level Guidance for Multi-turn Agents, 2026a. URL https://arxiv.org/abs/2606.15912.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking On-Policy Distillation of Large Language Models: Phenomenology, Mechanism, and Recipe, 2026b. URL https://arxiv.org/abs/2604.13016.

Kevin Lu and Thinking Machines Lab. On-Policy Distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. URL https://thinkingmachines.ai/blog/on-policy-distillation/.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Self-Distilled Agentic Reinforcement Learning, 2026. URL https://arxiv.org/abs/2605.15155.

E. S. Page. Continuous Inspection Schemes. *Biometrika*, 41(1–2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100. URL https://doi.org/10.1093/biomet/41.1-2.100.

Xuchen Pan, Yanxi Chen, Yushuo Chen, Yuchang Sun, Daoyuan Chen, Wenhao Zhang, Yuexiang Xie, Yilun Huang, Yilei Zhang, Dawei Gao, Weijie Shi, Yaliang Li, Bolin Ding, and Jingren Zhou. Trinity-RFT: A General-Purpose and Unified Framework for Reinforcement Fine-Tuning of Large Language Models, 2025. URL https://arxiv.org/abs/2505.17826.

Peter J. Rousseeuw and Christophe Croux. Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993. doi: 10.1080/01621459.1993.10476408. URL https://doi.org/10.1080/01621459.1993.10476408.

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

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.03768.

Jiaqi Wang, Wenhao Zhang, Weijie Shi, Yaliang Li, and James Cheng. TCOD: Exploring Temporal Curriculum in On-Policy Distillation for Multi-turn Autonomous Agents, 2026. URL https://arxiv.org/abs/2604.24005.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your Agent Smarter than a 5th Grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, 2022. doi: 10.18653/v1/2022.emnlp-main.775. URL https://aclanthology.org/2022.emnlp-main.775/.

B. P. Welford. Note on a Method for Calculating Corrected Sums of Squares and Products. *Technometrics*, 4(3):419–420, 1962. doi: 10.1080/00401706.1962.10490022.

Jinyang Wu, Shuo Yang, Zhengxi Lu, Fan Zhang, Yuhao Shen, Lang Feng, Haoran Luo, Zheng Lian, Shuai Zhang, Zhengqi Wen, and Jianhua Tao. SEED: Self-Evolving On-Policy Distillation for Agentic Reinforcement Learning, 2026. URL https://arxiv.org/abs/2607.14777.

Yuchen Xia, Qianguo Sun, Chao Song, Junlong Wu, Yiyan Qi, and Yunjian Xu. DASH-OPD: Discrepancy-Aware Switching with Hysteresis for On-Policy Distillation, 2026. URL https://arxiv.org/abs/2607.29078.

Haolei Xu, Xiaowen Xu, Haiwen Hong, Zixuan Ni, Hongxing Li, Yiwen Qiu, Weiming Lu, and Yongliang Shen. Pass the Baton: Trajectory-Relayed On-Policy Distillation, 2026. URL https://arxiv.org/abs/2607.26057.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

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*, volume 35, pp. 20744–20757, 2022. URL https://arxiv.org/abs/2207.01206.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models, 2026. URL https://arxiv.org/abs/2601.18734.

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. SOD: Step-wise On-policy Distillation for Small Language Model Agents, 2026. URL https://arxiv.org/abs/2605.07725.

Yuhang Zhou, Kai Zheng, Haoling Li, Dengyun Peng, Can Xu, and Jingjing Chen. TurnOPD: Making On-Policy Distillation Turn-Aware for Efficient Long-Horizon Agent Training, 2026. URL https://arxiv.org/abs/2607.05804.

A Detailed Algorithm

Algorithm 1 summarizes trajectory collection and the student update. The active executor generates a response and interacts with the environment; accumulated discrepancy determines whether to invoke MCOL for rollback or return control to the student. Algorithm 2 specifies these decisions using the equations in the main text.

Algorithm 1 REWIND-OPD trajectory collection and student update

Require: Student π_θ, teacher π_T, and training tasks
Freeze the rollout student and initialize student and teacher training buffers
for each task **do**
    Initialize the environment, interaction history, and both switching controllers
    Select the initial executor, choosing the teacher with probability $p_T(s) = 1 - s/S_{\max}$
    while the task is not terminal and the retained history has fewer than H turns **do**
        Generate a response with the active executor σ_t and score it with both models
        Store the response and its original context in the corresponding training buffer
        Parse the response into an action; execute it and extend the history with the returned observation
        Compute the discrepancy using Eq. (3)
        Update the active controller and evaluate the switching rule UPDATECONTROLLER
        if the task is terminal or the horizon is reached **then break**
        if student drift triggers a switch **then**
            Select the intervention turn j^* using LOCALIZE
            Roll back to j^* and let the teacher resume generation
        else if teacher recovery triggers a switch **then**
            Return control to the student and begin a new student segment
        end if
        On a switch, reset the departing executor's evidence but retain its Welford statistics
    end while
end for
Update the student on eligible student and teacher records using Eq. (11)

Algorithm 2 Switching controller and MCOL localization

function UPDATECONTROLLER
    Normalize the current gap using prior Welford statistics: Eq. (5)
    Incorporate the gap into the active executor's running statistics
    Accumulate drift or recovery evidence: Eq. (6)
    return the switching decision in Eq. (7)
end function

function LOCALIZE
    Collect gaps from the current uninterrupted student segment, including the triggering turn
    Compute the median reference and standardize the gaps: Eqs. (8) and (9)
    Select the highest-scoring suffix using Eq. (10); prefer a shallower depth on ties
    return its starting turn j^*
end function

Rollback shortens the active interaction history but does not discard collected training records or reset Welford statistics. Removed interactions do not count toward the horizon H. Eligible responses from both the retained path and replaced branches are trained under their original contexts, with the generating executor determining the loss term in Eq. (11).

B Benchmarks and Evaluation Metrics

We evaluate agents on three text-interactive benchmarks covering household tasks, online shopping, and scientific experimentation. Their dataset characteristics are summarized in Table 4. We describe these benchmarks below.

ALFWorld. ALFWorld is a text-based embodied environment for household navigation and object manipulation (Shridhar et al., 2021). Its six task categories cover object placement, placement after

Table 4: Introduction of Evaluation Benchmark Environments.

Characteristic	ALFWorld	WebShop	ScienceWorld
Type	Embodied	E-Commerce	Scientific Reasoning
Difficulty	Easy / Middle / Hard	Medium	Hard
Number of Training Tasks	3,553	4,096	2,294
Number of Evaluation Tasks	140 / 134 / 121	100	1,308
Evaluation Split	Seen / Unseen / Hard	Test	Test
Maximum Interaction Steps	30	15	30

cleaning, heating or cooling, examination under a light source, and placing two objects of the same type. We follow the standard seen and unseen splits, reported as IID and OOD, respectively. We also use TCOD's Hard set, comprising training-split tasks on which the teacher fails under pass@10 sampling, to evaluate generalization beyond the teacher's capability boundary (Wang et al., 2026). The environment reward is a binary indicator of task completion, where 1 indicates success and 0 indicates failure. We report success rate (SR) and mean interaction rounds.

WebShop. WebShop is a simulated e-commerce environment in which an agent purchases a product satisfying a natural-language request (Yao et al., 2022). At each turn, the agent observes a rendered HTML page, and the available actions are updated to reflect the current page. These primarily include search[query], available when a search bar is present, and click[button_name], whose argument must match a currently clickable element. The agent searches, navigates product pages, selects options, and completes a purchase while preserving the requested constraints. The environment then returns a continuous score in $[0, 100]$ that measures how well the purchased product matches the requested attributes, with a score of 100 indicating success. We report average score (Score), success rate (SR), and mean interaction rounds.

ScienceWorld. ScienceWorld is a text-based interactive environment for scientific reasoning (Wang et al., 2022). It provides 30 task types aligned with an elementary-school science curriculum, covering thermodynamics, electrical circuits, chemistry, biology, and matter classification. At each turn the agent observes a structured textual state listing its current location, visible objects, and inventory, and chooses from a large action space (e.g., `focus on`, `activate`, `mix` and `move ... to ...`). Each episode is limited to 30 interaction steps. The environment returns a continuous score in $[0, 100]$ measuring the fraction of completed subgoals; we report the highest score attained in an episode (`best_score`), treat a perfect score as success (SR), and additionally report mean interaction rounds. Completing these tasks usually requires multi-step tool use and long-horizon cause–effect reasoning, which makes ScienceWorld substantially harder than the other two benchmarks in our suite.

Evaluation Metrics. We follow the evaluation metrics used in previous work (Wang et al., 2026; Li et al., 2026a) and report success rate (SR), average score (Score), and mean interaction rounds (Rnd). Each ALFWorld split and the WebShop evaluation set contains fewer than 150 tasks, so results can vary across runs. We therefore report the mean and standard deviation across seeds 41, 42, and 43. For efficiency comparisons, we report output tokens per trajectory in Appendix D and training time in Appendix E.1. For an evaluation set of N episodes, let $r_i \in [0, 1]$ denote the normalized environment score of episode i, and let $\mathbb{1}(\cdot)$ denote the indicator function. We compute success rate (SR) and average score (Score) as follows:

$$\mathrm{SR} = \frac{100}{N} \sum_{i=1}^{N} \mathbb{1}\left(r_i = 1\right), \quad \mathrm{Score} = \frac{100}{N} \sum_{i=1}^{N} r_i, \tag{12}$$

C Experimental Settings

C.1 Teacher Model Training

We initialize separate ALFWorld and WebShop teachers from Qwen3-8B and train them with GiGPO using the verl-agent framework (Feng et al., 2025). Both use AdamW with learning rate

10^{-6} and weight decay 0.01, discount factor $\gamma = 0.95$, and 8 trajectories per task. A `low_var_kl` regularizer with coefficient 0.01 is included in the policy loss, not in the environment reward. The invalid-action penalty coefficient is 0.1. Rollouts are generated with vLLM (Kwon et al., 2023) using temperature 1.0 during training and 0.4 during validation. Dedicated thinking mode is disabled. Each turn allows at most 4,096 input tokens and 512 generated tokens. Both teachers use a single node with eight NVIDIA H20 GPUs, FSDP, gradient checkpointing, and rollout tensor parallelism of two. Teacher checkpoints are selected by the highest score on the corresponding validation set, yielding global step 70 for ALFWorld and global step 110 for WebShop. The detailed training setup is shown in Table 5.

Table 5: GiGPO Teacher-Training Configurations For ALFWorld And WebShop. Shared Settings Are Listed Once; Benchmark-Specific Settings Are Reported Separately.

Configuration	ALFWorld	WebShop
Shared configs.		
Base Model / Framework	Qwen3-8B / GiGPO (verl-agent)	
Optimizer / Learning Rate / Weight Decay	AdamW / 10^{-6} / 0.01	
Discount γ / Trajectories Per Task	0.95 / 8	
Advantage Norm. / Step-Level Weight	`mean_norm` / 1.0	
Policy-Ratio Clip $(\epsilon_{\text{low}}, \epsilon_{\text{high}})$	(0.2, 0.2)	
KL Regularizer / Coefficient	`low_var_kl` / 0.01 (policy loss)	
Invalid-Action Penalty Coefficient	0.1	
Rollout Engine / Thinking Mode	vLLM / Disabled	
Train / Validation Temperature	1.0 / 0.4	
Input / Output Tokens Per Turn	4,096 / 512	
Hardware / Parallelism	8×H20; FSDP; grad. ckpt; TP = 2	
PPO Mini-Batch / Micro-Batch / Log-Prob Micro-Batch	64 / 4 / 8 (per GPU)	
Benchmark-specific configs.		
Training Task Batch Size	32	16
Validation Batch Size	64	32
Maximum Interaction Steps	30	15
Selected Checkpoint (Global Step)	70	110

C.2 Training and Evaluation Settings

For training, we implement each OPD variant with the Trinity-RFT framework (Pan et al., 2025). We use the prompting protocol described in Appendix F for both training and evaluation. We sample one trajectory for each task at a temperature of 1.0. The input and generation limits are set to 10,240 and 512 tokens. For context management, we retain the cumulative history of all past user–assistant pairs and additionally include a summary of the two most recent observation–action pairs from turn four onward. For context truncation, we stop sampling when the input token limit is reached.

For evaluation, each model executes independently, without teacher assistance or the training-time controller. We sample one trajectory for each task at a temperature of 0.4. The input and generation limits are set to 10,240 and 4,096 tokens. For context management, we retain the cumulative history of all past user–assistant pairs and additionally include a summary of the two most recent observation–action pairs from turn four onward. We remove the oldest user–assistant pairs to keep the context within the input limit. Each ALFWorld split and the WebShop evaluation set contains fewer than 150 tasks, so the results of a single evaluation can vary substantially. Therefore, we report the mean and standard deviation of three evaluation seeds (41, 42, 43) for these benchmarks. For ScienceWorld, which contains 1,308 evaluation tasks, we report results using seed 42. Table 6 reports the shared training and evaluation configurations.

Table 6: Shared Training And Evaluation Configurations.

Configuration	Value
Training Configs.	
Training Framework / Hardware	Trinity-RFT / 8 NVIDIA H20 GPUs
Optimizer	AdamW
Learning Rate η / Schedule	10^{-6} / Constant, No Warm-Up
Adam Moments (β_1, β_2) / Weight Decay λ_{wd}	$(0.9, 0.999)$ / 0.01
Gradient-Norm Limit $g_{\max}$ / Distillation Weight λ_{KD}	1.0 / 1.0
Rollout Batch B_{roll} / Update Batch B_{upd}	16 / 64
PPO Clip ϵ_{PPO} / Dual Clip c_{dual}	0.2 / 3.0
Critic / Reference-Policy KL Penalty / Entropy Bonus	None
Rollouts Per Task	1
Top-p / Top-k / Min-p / Sampling Temperature T_{samp}	1 / -1 / 0 / 1.0
Prompt Limit L_{in} / Generation Limit L_{out}	10,240 / 512 Tokens Per Turn
Context Management	Cumulative Context of All Past User-Assistant Pairs
Context Truncation	Stop Sampling when Reaching Maximum Input Tokens
Thinking Mode / Numerical Precision	Disabled / bfloat16
Dynamic Batch Token Budget Per GPU	16,384
Ulysses Sequence Parallelism / Training Seed	2 / 42
Maximum Sample Staleness	2 Student-Policy Versions
Evaluation Configs.	
Top-p / Top-k / Min-p / Sampling Temperature T_{eval}	1 / -1 / 0 / 0.4
Maximum Input / Output Tokens Per Turn	10,240 / 4,096
Thinking Mode	Disabled
Rollouts Per Task Per Evaluation Seed	1
Context Management	Cumulative Context of All Past User-Assistant Pairs
Context Truncation	Remove Oldest User–Assistant Pairs
Reported Statistics	Mean Across Evaluation Seeds
Evaluation Seeds	41, 42, 43 (ScienceWorld: 42)

C.3 Introduction and Implementation of Different Methods

We describe each method and its key configuration below; shared settings follow Section C.2. We use π_T for the frozen teacher and π_θ for the student in the definitions below.

Vanilla OPD. Following the agent-level formulation of TCOD (Wang et al., 2026), Vanilla OPD samples trajectories from the student and minimizes the student-to-teacher reverse KL divergence at every visited history:

$$\mathcal{L}_{\text{OPD}}(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{H_\tau-1} D_{\text{KL}}(\pi_\theta(\cdot \mid h_t) \,\|\, \pi_T(\cdot \mid h_t))\right], \tag{13}$$

where y denotes a complete response, as in the main text, and

$$D_{\text{KL}}(\pi_\theta \| \pi_T \mid h_t) = \sum_y \pi_\theta(y \mid h_t) \log \frac{\pi_\theta(y \mid h_t)}{\pi_T(y \mid h_t)}. \tag{14}$$

Training uses token-level losses conditioned on each sampled response prefix, as in Section 4.4.

TCOD. TCOD controls the student rollout horizon through a temporal curriculum (Wang et al., 2026). In TCOD-F2B, the student rolls out for at most k turns and optimizes

$$\mathcal{L}_{\text{TCOD-F2B}}(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{k-1} D_{\text{KL}}(\pi_\theta(\cdot \mid h_t) \,\|\, \pi_T(\cdot \mid h_t))\right]. \tag{15}$$

The horizon follows the linear pacing rule

$$k = k_{\text{start}} + \left\lfloor \frac{n}{\eta} \right\rfloor, \qquad n \in \{1, \dots, N\}, \tag{16}$$

and is capped by the environment horizon in implementation. We set $k_{\text{start}} = 1$, with $\eta = 4$ rollout batches on WebShop and $\eta = 6$ on ALFWorld and ScienceWorld.

Guided-OPD. Guided-OPD samples the executor independently at each turn (Li et al., 2026a). At update s, a teacher turn is sampled with probability $p_T(s) = \frac{1}{2}[1 + \cos(\pi \min\{s/S_{\text{ann}}, 1\})]$. Let $z_t \sim \text{Bernoulli}(p_T(s))$ indicate a teacher turn and let $\mathcal{L}_t^{\text{rKL}}$ and $\mathcal{L}_t^{\text{fKL}}$ denote the token-averaged reverse and forward KL losses. Guided-OPD optimizes

$$\mathcal{L}_{\text{Guided}}(\theta; s) = \mathbb{E}_{\tau,z}\left[\sum_t (1 - z_t)\mathcal{L}_t^{\text{rKL}}(\theta) + z_t \mathcal{L}_t^{\text{fKL}}(\theta)\right]. \tag{17}$$

After annealing, all turns are student generated. We anneal over 120 rollout batches on WebShop and 200 on ALFWorld and ScienceWorld, followed by 30 and 50 student-only batches, respectively.

TurnOPD. TurnOPD jointly adapts rollout depth and turn-level loss allocation using periodic full-depth probes (Zhou et al., 2026). Let κ_t be the top-K_{turn} reverse-KL signal and n_t the number of probe trajectories reaching turn t. It forms $m_t = [\kappa_t]_+ n_t / n_1$ and $q_t = m_t / (\sum_j m_j + \varepsilon_{\text{mass}})$, giving $H_{\text{eff}} = \text{round}(\sum_t t q_t)$; the complementary coverage horizon is $H_{\text{cov}} = Q^{q_{\text{cov}}}(L_{\text{succ}})$. With $H_{\text{ctrl}} = \max\{H_{\text{eff}}, H_{\text{cov}}\}$, the next rollout horizon is

$$\bar{H}_s = (1 - \beta_H)\bar{H}_{s-1} + \beta_H H_{\text{ctrl}}, \qquad H_{s+1} = \text{clip}\big(\text{round}(\bar{H}_s) + 1, H_{\min}, \min\{H_{\max}^{\text{req}}, H\}\big). \tag{18}$$

For loss allocation, TurnOPD uses $q_t^{\text{blend}} = (1 - \lambda_s) n_t / (\sum_j n_j) + \lambda_s / H$, where λ_s increases linearly from 0 to 1 over training. Full-depth probes use $4 \times 4 = 16$ trajectories and run during the first three updates and every eight updates thereafter. We use $K_{\text{turn}} = 50$, $q_{\text{cov}} = 0.8$, and $\beta_H = 0.30$.

DASH-OPD. DASH-OPD measures directional teacher–student discrepancy on action tokens and accumulates separate drift and recovery evidence (Xia et al., 2026). On student- and teacher-controlled turns, respectively, the original directional signals are

$$\begin{aligned} d_t^S &= \frac{1}{N_t^A} \sum_i A_{t,i} \left[\log \pi_{\bar{\theta}}(y_{t,i} \mid h_t, y_{t,<i}) - \log \pi_T(y_{t,i} \mid h_t, y_{t,<i})\right], \quad y_t \sim \pi_{\bar{\theta}}, \\ d_t^T &= \frac{1}{N_t^A} \sum_i A_{t,i} \left[\log \pi_T(y_{t,i} \mid h_t, y_{t,<i}) - \log \pi_{\bar{\theta}}(y_{t,i} \mid h_t, y_{t,<i})\right], \quad y_t \sim \pi_T, \end{aligned} \tag{19}$$

where $A_{t,i}$ is the action-token mask and $N_t^A = \sum_i A_{t,i}$. Using the statistics of the corresponding executor, DASH-OPD standardizes the active signal and updates the two evidence variables as

$$\bar{d}_t^{\sigma_t} = \frac{d_t^{\sigma_t} - \mu_{t-1}^{\sigma_t}}{\sqrt{v_{t-1}^{\sigma_t}}}, \qquad \sigma_t \in \{S, T\}, \tag{20}$$

$$D_t = \max(D_{t-1} + \bar{d}_t^S, 0), \quad \sigma_t = S; \qquad R_t = \max(R_{t-1} - \bar{d}_t^T, 0), \quad \sigma_t = T. \tag{21}$$

The switching thresholds and decisions follow the original definitions

$$\tau_{\text{on}} = q_{\text{on}} H, \qquad \tau_{\text{off}} = q_{\text{off}} H, \tag{22}$$

$$\text{intervention}_t = (D_t > \tau_{\text{on}}) \vee g_t, \qquad \text{return}_t = (R_t > \tau_{\text{off}}) \wedge (\ell_t \geq \ell_{\min}), \tag{23}$$

where g_t is the stagnation flag and ℓ_t is the current teacher-segment length. We use the official configuration $q_{\text{on}} = 0.1$, $q_{\text{off}} = q_{\text{on}}/3$, $\ell_{\min} = 2$, and stagnation threshold $K = 3$, with a normalization floor of 10^{-3}. The initial teacher probability decays linearly as $p_T(s) = 1 - s/S_{\max}$.

REWIND-OPD. REWIND-OPD adds MCOL rollback to the discrepancy-based switching controller. We use $q_{\text{on}} = 0.1$, $q_{\text{off}} = 1/30$, a minimum teacher span of two turns, and a stagnation threshold of three turns; stagnation triggers takeover without rollback. The maximum rewind depth is five for the 1.7B student and three for the 4B student, with MCOL allowance $\delta = 0$. Teacher intervention is limited to 15, 8, and 15 turns on ALFWorld, WebShop, and ScienceWorld, respectively. Replaced branches contribute to training on ALFWorld and WebShop but not on ScienceWorld.

C.4 REWIND-POSITION SELECTORS

This section specifies the four selectors compared in Section 5.6. They determine *where* to rewind after a shared intervention trigger, rather than deciding whether to invoke the teacher. Let $d_1, \ldots, d_n$ be the available finite student–teacher action-logprob gaps, with n denoting the triggering observation. A depth k removes the last k student actions before teacher intervention. Requested depths are clipped to $[1, K]$ and further constrained to preserve earlier teacher segments and previous rewind boundaries. In the online comparison, $K = 5$; in the fixed-two-action experiment, K is the largest feasible positive depth among the saved branches (at most 4 on ALFWorld and 3 on WebShop). Predictions are mapped to the nearest available positive depth, with ties favoring the shallower branch. These diagnostic settings are distinct from the main distillation configurations in Section C.3.

Median-centered CUSUM Onset Localization (MCOL). MCOL selects the suffix with the largest accumulated discrepancy relative to a robust historical reference. For these experiments, the reference $\mathcal{H}_n^- = \{d_1, \ldots, d_{n-1}\}$ excludes the triggering observation. Define

$$\begin{aligned} m_n &= \operatorname{median}(\mathcal{H}_n^-), \\ s_n &= \max\Big\{c_{\text{MAD}} \operatorname{median}_{d \in \mathcal{H}_n^-} |d - m_n|, c_{\text{SD}} \operatorname{std}(\mathcal{H}_n^-), \varepsilon_{\text{R}}\Big\}, \\ \widehat{k}_{\text{MCOL}} &= \min \operatorname*{arg\,max}_{1 \le k \le \min(K,n)} \sum_{j=0}^{k-1} \left(\frac{d_{n-j} - m_n}{s_n} - \delta \right). \end{aligned} \tag{24}$$

We use all available pre-trigger history, $c_{\text{MAD}} = 1.4826$, $c_{\text{SD}} = 0.25$, $\varepsilon_{\text{R}} = 10^{-3}$, and $\delta = 0$; a single available observation yields depth one. The factor 1.4826 calibrates MAD for consistency as an estimator of the standard deviation under normality (Rousseeuw & Croux, 1993); it is not a tuned hyperparameter or an unbiasedness correction for arbitrary distributions. The minimum over maximizers implements shallower tie-breaking. Unlike a consecutive-run rule, the cumulative score can span isolated negative deviations. With zero allowance, the common positive scale does not change the maximizing depth; selection depends on the historical median and cumulative centered gaps.

Causal-EMA Onset Localization. This selector searches backward for the latest observation within an upper band estimated only from its preceding history. Initialize the exponential moving average as $e_1 = d_1$ and update $e_i = (1 - \alpha)e_{i-1} + \alpha d_i$. For $i \ge 2$, let σ_i^- be the population standard deviation of the preceding $\min(w, i - 1)$ observations. The upper band and selected depth are

$$B_i = e_{i-1} + \beta \max(\sigma_i^-, \varepsilon_{\text{E}}), \qquad j^* = \max\{i < n : d_i \le B_i\}, \qquad \widehat{k}_{\text{EMA}} = \text{clip}(n - j^*, 1, K). \tag{25}$$

The first band is $B_1 = d_1 + \beta\varepsilon_{\text{E}}$, and a history shorter than two observations has zero estimated standard deviation. A single available observation yields depth one; if no earlier in-band observation exists, the entire available suffix is selected before clipping. We use $\alpha = 0.2$, $\beta = 1.5$, $w = 5$, and $\varepsilon_{\text{E}} = 10^{-6}$. The window limits the scale estimate, whereas the EMA center remains recursive.

Positive-Run Onset Localization. Using Welford statistics computed before incorporating d_i, define $z_i = (d_i - \mu_{i-1})/\max(\sigma_{i-1}, \varepsilon_{\text{W}})$, with $z_1 = 0$ and $\varepsilon_{\text{W}} = 10^{-3}$. The selector counts the terminal run of positive standardized gaps:

$$P_0 = 0, \qquad P_i = \mathbb{1}\{z_i > 0\}(P_{i-1} + 1), \qquad \widehat{k}_{\text{positive}} = \text{clip}(P_n, 1, K). \tag{26}$$

Any nonpositive standardized gap ends the run, even if earlier observations provide substantial discrepancy evidence.

Evidence-Onset Localization. This selector uses the duration of the current positive accumulated-drift episode:

$$D_0 = E_0 = 0, \qquad D_i = \max(0, D_{i-1} + z_i),$$
$$E_i = \mathbb{1}\{D_i > 0\}(E_{i-1} + 1), \qquad \widehat{k}_{\text{evidence}} = \text{clip}(E_n, 1, K). \tag{27}$$

It can span negative individual gaps provided the accumulated evidence remains positive. Unlike MCOL, it uses the last evidence reset rather than comparing cumulative scores across candidate suffixes. Positive-Run and Evidence-Onset introduce no additional selector hyperparameters beyond the shared normalization and depth limits.

D Experimental Results

This appendix details the results in Tables 1 and 2. Bold and underlined entries indicate the best and second-best trained results; light blue highlights results matching or exceeding the teacher. Rnd and Token denote mean interaction rounds and output tokens per task. ALFWorld and WebShop report mean ± standard deviation across evaluation seeds; ScienceWorld reports a single full-set evaluation. We compare REWIND-OPD with Vanilla OPD and the strongest baseline by SR on each ALFWorld split or Score on WebShop and ScienceWorld.

D.1 Main Results with Qwen3-30B-A3B Teacher

ALFWorld. In Tables 7 and 8, REWIND-OPD achieves 31.9%, 36.6%, and 14.3% SR on Seen, Unseen, and Hard with the 1.7B student. These improve over Vanilla OPD by 5.9, 6.3, and 3.6 percentage points and over the strongest baselines on the respective splits by 3.3, 2.5, and 2.2 points. At 4B, the corresponding SR values are 39.8%, 42.8%, and 22.3%, with gains of 4.8, 7.5, and 6.0 points over Vanilla OPD and 1.7, 3.5, and 3.0 points over the strongest baselines. REWIND-OPD uses fewer rounds than Vanilla OPD on every split, with the lowest or tied-lowest means among trained methods. Its token reductions relative to Vanilla OPD range from 7.3% to 11.4% at 1.7B and 2.7% to 6.6% at 4B. On Unseen at 1.7B, it also reduces token usage by 17.1% relative to the strongest SR baseline, TCOD-F2B.

Table 7: ALFWorld Seen/Unseen results for Qwen3-1.7B/4B students distilled from Qwen3-30B-A3B.

Method	Seen			Unseen			Training Time↓
	SR↑	Rnd↓	Token↓	SR↑	Rnd↓	Token↓	
Teacher: Qwen3-30B-A3B							
Zero-Shot	44.0 ± 2.4	22.8 ± 0.4	1205.9 ± 13.7	39.1 ± 0.9	23.5 ± 0.2	1215.1 ± 20.4	–
Student: Qwen3-1.7B							
Zero-Shot	2.4 ± 1.9	29.5 ± 0.4	587.3 ± 44.6	3.5 ± 1.9	29.2 ± 0.4	674.6 ± 24.2	–
Vanilla OPD	26.0 ± 0.7	25.4 ± 0.2	1435.5 ± 21.6	30.3 ± 1.5	24.8 ± 0.3	1443.4 ± 38.6	1.00×
TCOD-F2B	26.4 ± 2.5	25.5 ± 0.4	1536.0 ± 25.5	34.1 ± 4.0	24.6 ± 0.5	1541.9 ± 33.9	**0.54**×
Guided-OPD	26.9 ± 1.8	25.7 ± 0.2	1379.3 ± 16.3	30.3 ± 0.9	25.1 ± 0.1	1320.2 ± 16.1	0.90×
TurnOPD	26.0 ± 0.9	25.2 ± 0.2	1418.1 ± 24.0	33.1 ± 3.4	24.3 ± 0.4	1376.6 ± 25.8	1.30×
DASH-OPD	28.6 ± 1.0	25.4 ± 0.2	1416.1 ± 17.6	33.6 ± 2.6	24.4 ± 0.5	1375.2 ± 12.9	0.68×
REWIND-OPD	**31.9 ± 2.2**	**24.3 ± 0.4**	**1330.8 ± 4.90**	**36.6 ± 0.8**	**24.2 ± 0.3**	**1278.3 ± 60.0**	0.63×
Student: Qwen3-4B							
Zero-Shot	30.5 ± 3.7	25.0 ± 0.3	1341.1 ± 30.5	37.6 ± 6.1	24.0 ± 0.7	1295.9 ± 47.4	–
Vanilla OPD	35.0 ± 1.2	24.2 ± 0.2	1326.8 ± 18.4	35.3 ± 2.3	24.0 ± 0.4	1284.3 ± 26.9	1.00×
TCOD-F2B	35.0 ± 1.0	24.0 ± 0.2	**1231.5 ± 11.2**	38.1 ± 1.2	23.4 ± 0.5	1277.4 ± 30.5	**0.65**×
Guided-OPD	36.4 ± 3.2	23.9 ± 0.4	1284.8 ± 44.7	39.3 ± 2.3	24.1 ± 0.6	**1245.3 ± 23.2**	0.96×
TurnOPD	36.2 ± 0.7	23.9 ± 0.0	1288.2 ± 28.1	34.6 ± 2.5	24.6 ± 0.1	1335.8 ± 8.60	0.75×
DASH-OPD	38.1 ± 0.3	23.9 ± 0.1	1314.2 ± 17.4	37.8 ± 0.7	24.2 ± 0.3	1271.6 ± 21.9	0.76×
REWIND-OPD	**39.8 ± 1.9**	**23.3 ± 0.4**	1249.3 ± 13.4	**42.8 ± 2.0**	**23.1 ± 0.4**	1250.1 ± 10.9	0.76×

Table 8: ALFWorld Hard results for Qwen3-1.7B/4B students distilled from Qwen3-30B-A3B.

	Student: Qwen3-1.7B				Student: Qwen3-4B			
Method	SR↑	Rnd↓	Token↓	Time↓	SR↑	Rnd↓	Token↓	Time↓
Teacher: Qwen3-30B-A3B								
Zero-Shot	21.8 ± 2.4	27.7 ± 0.2	1487.1 ± 35.5	–	21.8 ± 2.4	27.7 ± 0.2	1487.1 ± 35.5	–
Zero-Shot	0.3 ± 0.4	30.0 ± 0.1	603.7 ± 40.1	–	9.6 ± 2.4	29.0 ± 0.2	1508.4 ± 48.0	–
Vanilla OPD	10.7 ± 0.7	28.7 ± 0.1	1646.6 ± 2.90	1.00×	16.3 ± 3.0	28.7 ± 0.2	1582.1 ± 16.2	1.00×
TCOD-F2B	11.6 ± 0.7	28.6 ± 0.3	1731.3 ± 11.1	**0.54**×	17.6 ± 0.4	28.0 ± 0.2	1489.9 ± 3.40	**0.65**×
Guided-OPD	12.1 ± 1.0	28.7 ± 0.1	**1519.1 ± 11.1**	0.90×	17.4 ± 2.3	28.3 ± 0.1	1592.9 ± 24.8	0.96×
TurnOPD	10.2 ± 0.8	**28.5 ± 0.2**	1615.3 ± 37.6	1.30×	17.9 ± 1.0	28.2 ± 0.2	1530.8 ± 21.1	0.75×
DASH-OPD	11.0 ± 3.5	28.7 ± 0.5	1630.8 ± 33.2	0.68×	19.3 ± 1.4	28.1 ± 0.2	1540.8 ± 26.7	0.76×
REWIND-OPD	**14.3 ± 2.1**	**28.5 ± 0.3**	1521.3 ± 29.5	0.63×	**22.3 ± 3.1**	**27.8 ± 0.2**	**1476.9 ± 16.2**	0.76×

WebShop. In Table 9, the 1.7B REWIND-OPD student achieves 42.9 Score and 14.0% SR, improving over Vanilla OPD by 5.3 score points and 2.0 SR points, and over the strongest Score baseline, TurnOPD, by 2.6 and 3.7 points. It uses 6.5 rounds and 625.4 tokens per task, reducing token usage by 35.2% and 13.4% relative to these two baselines. At 4B, REWIND-OPD reaches 42.1 Score and 16.7% SR, exceeding Vanilla OPD by 4.7 score points and 4.0 SR points, and the strongest Score baseline, DASH-OPD, by 1.2 and 3.0 points. Its 7.1 rounds save 0.6 and 0.2 rounds, respectively. Token usage decreases by 10.3% relative to Vanilla OPD, while the gain over DASH-OPD uses 14.2 additional tokens per task.

Table 9: WebShop results for Qwen3-1.7B/4B students distilled from Qwen3-30B-A3B.

Method	Score↑	SR↑	Rnd↓	Token↓	Time↓
Teacher: Qwen3-30B-A3B					
Zero-Shot	41.6 ± 2.1	15.3 ± 0.5	7.2 ± 0.5	681.9 ± 59.1	–
Student: Qwen3-1.7B					
Zero-Shot	3.7 ± 0.6	0.7 ± 0.5	14.4 ± 0.3	423.2 ± 36.0	–
Vanilla OPD	37.6 ± 2.7	12.0 ± 0.8	7.3 ± 0.5	965.1 ± 279	1.00×
TCOD-F2B	39.5 ± 0.3	11.0 ± 0.8	6.8 ± 0.2	895.4 ± 249	0.64×
Guided-OPD	37.0 ± 0.4	11.0 ± 0.8	7.5 ± 0.1	1133.7 ± 201	1.02×
TurnOPD	40.3 ± 0.0	10.3 ± 0.9	7.0 ± 0.2	722.1 ± 41.0	**0.16**×
DASH-OPD	38.9 ± 1.4	12.0 ± 1.7	6.9 ± 0.3	932.7 ± 288.1	0.89×
REWIND-OPD	**42.9 ± 0.8**	**14.0 ± 0.8**	**6.5 ± 0.2**	**625.4 ± 9.40**	0.94×
Student: Qwen3-4B					
Zero-Shot	28.6 ± 1.5	10.0 ± 0.8	10.5 ± 0.3	1097.7 ± 72.4	–
Vanilla OPD	37.4 ± 0.7	12.7 ± 0.5	7.7 ± 0.2	803.0 ± 9.60	1.00×
TCOD-F2B	39.9 ± 2.1	15.0 ± 0.0	7.7 ± 0.4	755.3 ± 11.0	**0.69**×
Guided-OPD	38.4 ± 2.1	14.3 ± 1.2	7.8 ± 0.3	791.1 ± 36.5	0.99×
TurnOPD	36.0 ± 1.7	13.0 ± 0.0	8.3 ± 0.2	1044.8 ± 232	0.86×
DASH-OPD	40.9 ± 1.4	13.7 ± 0.5	7.3 ± 0.0	**705.8 ± 30.5**	0.83×
REWIND-OPD	**42.1 ± 3.3**	**16.7 ± 0.9**	**7.1 ± 0.5**	720.0 ± 42.4	0.91×

ScienceWorld. In Table 10, REWIND-OPD reaches 41.5 Score and 12.7% SR at 1.7B. It improves over Vanilla OPD by 9.2 score points and 5.1 SR points, and over the strongest baseline, DASH-OPD, by 4.9 and 0.7 points. Its 23.7 rounds and 1815.0 tokens reduce rounds by 1.8 and 0.1 and token usage by 12.3% and 2.8%, respectively. At 4B, REWIND-OPD achieves 52.0 Score and 22.6% SR, with gains of 6.0 score points and 3.9 SR points over Vanilla OPD and 1.0 and 0.4 points over DASH-OPD. It uses 20.5 rounds and 1409.5 tokens, saving 1.9 and 0.9 rounds and reducing token usage by 9.3% and 10.1%, respectively.

Table 10: ScienceWorld results for Qwen3-1.7B/4B students distilled from Qwen3-30B-A3B.

Method	Score↑	SR↑	Rnd↓	Token↓	Time↓
Teacher: Qwen3-30B-A3B					
Zero-Shot	50.6	22.6	21.1	1357.0	–
Student: Qwen3-1.7B					
Zero-Shot	10.8	1.8	16.7	163.9	–
Vanilla OPD	32.3	7.6	25.5	2069.7	1.00×
TCOD-F2B	36.1	11.2	24.3	1973.2	0.49×
Guided-OPD	32.1	10.2	25.2	1905.9	0.73×
DASH-OPD	36.6	12.0	23.8	1868.1	0.94×
REWIND-OPD	**41.5**	**12.7**	**23.7**	**1815.0**	**0.45**×
Student: Qwen3-4B					
Zero-Shot	39.8	15.9	19.5	1448.9	–
Vanilla OPD	46.0	18.7	22.4	1554.0	1.00×
TCOD-F2B	46.9	20.1	21.8	1503.3	0.60×
Guided-OPD	46.1	19.7	21.1	1450.9	0.83×
DASH-OPD	51.0	22.2	21.4	1567.9	**0.56**×
REWIND-OPD	**52.0**	**22.6**	**20.5**	**1409.5**	0.65×

D.2 MAIN RESULTS WITH POST-TRAINED QWEN3-8B TEACHER

ALFWorld. In Tables 11 and 12, REWIND-OPD achieves 82.6%, 82.3%, and 43.8% SR on Seen, Unseen, and Hard. The gains over Vanilla OPD are 16.4, 16.9, and 35.3 percentage points; those over Guided-OPD, the strongest SR baseline on all three splits, are 0.5, 1.2, and 0.5 points. REWIND-OPD uses 12.0, 12.9, and 23.0 rounds, saving 2.9, 2.4, and 5.4 rounds relative to Vanilla OPD and 0.1, 0.1, and 0.7 rounds relative to Guided-OPD. Compared with Guided-OPD, its token usage decreases by 58.3%, 0.6%, and 36.1%, respectively, alongside higher SR on each split.

Table 11: ALFWorld Seen/Unseen results for Qwen3-4B distilled from a GiGPO-trained Qwen3-8B teacher.

Method	Seen			Unseen			Training Time↓
	SR↑	Rnd↓	Token↓	SR↑	Rnd↓	Token↓	
Teacher: Qwen3-8B-GIGPO							
Zero-Shot	86.2 ± 1.2	11.3 ± 0.1	509.4 ± 23.3	82.3 ± 2.5	13.2 ± 0.3	871.9 ± 155	–
Student: Qwen3-4B							
Zero-Shot	30.5 ± 3.7	25.0 ± 0.3	1341.1 ± 30.5	37.6 ± 6.1	24.0 ± 0.7	1295.9 ± 47.4	–
Vanilla OPD	66.2 ± 1.8	14.9 ± 0.4	181.9 ± 5.00	65.4 ± 3.1	15.3 ± 0.6	184.4 ± 6.70	1.00×
TCOD-F2B	71.4 ± 1.2	13.9 ± 0.2	**170.3 ± 2.70**	71.4 ± 2.0	14.2 ± 0.3	**172.7 ± 3.80**	**0.42**×
Guided-OPD	82.1 ± 1.5	12.1 ± 0.5	925.2 ± 322	81.1 ± 2.0	13.0 ± 0.3	799.4 ± 23.7	0.56×
TurnOPD	69.0 ± 1.2	14.7 ± 0.3	178.6 ± 3.00	64.2 ± 2.2	15.9 ± 0.4	191.8 ± 5.00	0.46×
DASH-OPD	80.5 ± 1.8	12.1 ± 0.4	487.4 ± 53.9	78.4 ± 2.2	13.2 ± 0.3	1266.5 ± 423	0.54×
REWIND-OPD	**82.6 ± 0.9**	**12.0 ± 0.1**	386.0 ± 39.2	**82.3 ± 2.5**	**12.9 ± 0.5**	795.0 ± 288	0.54×

WebShop. In Table 13, REWIND-OPD achieves 62.3 Score and 34.7% SR, exceeding Vanilla OPD by 3.4 score points and 3.4 SR points, and the strongest Score baseline, TurnOPD, by 0.9 and 1.7 points. It uses 4.8 rounds and 79.0 tokens, saving 0.2 rounds and 5.3% of tokens relative to Vanilla OPD. Relative to TurnOPD, the performance gain requires 0.2 more rounds while matching its token usage.

Table 12: ALFWorld Hard results for Qwen3-4B distilled from a GiGPO-trained Qwen3-8B teacher.

Method	SR↑	Rnd↓	Token↓	Time↓
Teacher: Qwen3-8B-GIGPO				
Zero-Shot	41.0 ± 0.8	23.2 ± 0.1	2715.7 ± 127	–
Student: Qwen3-4B				
Zero-Shot	9.6 ± 2.4	29.0 ± 0.2	1508.4 ± 48.0	–
Vanilla OPD	8.5 ± 1.4	28.4 ± 0.2	336.6 ± 4.20	1.00×
TCOD-F2B	10.5 ± 1.0	28.0 ± 0.2	334.2 ± 3.30	**0.42**×
Guided-OPD	43.3 ± 0.4	23.7 ± 0.1	3414.2 ± 777	0.56×
TurnOPD	14.0 ± 2.4	27.5 ± 0.3	**327.1 ± 4.90**	0.46×
DASH-OPD	41.6 ± 1.7	23.6 ± 0.2	2566.2 ± 662	0.54×
REWIND-OPD	**43.8 ± 1.3**	**23.0 ± 0.2**	2182.0 ± 505	0.54×

Table 13: WebShop results for Qwen3-4B distilled from a GiGPO-trained Qwen3-8B teacher.

Method	Score↑	SR↑	Rnd↓	Token↓	Time↓
Teacher: Qwen3-8B-GIGPO					
Zero-Shot	58.1 ± 0.8	31.3 ± 0.5	5.2 ± 0.1	87.1 ± 2.50	–
Student: Qwen3-4B					
Zero-Shot	28.6 ± 1.5	10.0 ± 0.8	10.5 ± 0.3	1097.7 ± 72.0	–
Vanilla OPD	58.9 ± 1.1	31.3 ± 2.1	5.0 ± 0.1	83.4 ± 3.10	1.00×
TCOD-F2B	60.7 ± 1.6	32.3 ± 2.9	4.8 ± 0.1	79.4 ± 1.50	**0.65**×
Guided-OPD	57.2 ± 1.6	32.0 ± 2.2	5.9 ± 0.0	98.0 ± 0.900	0.73×
TurnOPD	61.4 ± 0.8	33.0 ± 0.8	**4.6 ± 0.1**	**79.0 ± 1.30**	0.69×
DASH-OPD	57.5 ± 1.9	31.7 ± 3.3	5.1 ± 0.1	85.1 ± 1.80	1.22×
REWIND-OPD	**62.3 ± 1.8**	**34.7 ± 3.1**	4.8 ± 0.1	**79.0 ± 1.50**	0.66×

E TRAINING EFFICIENCY

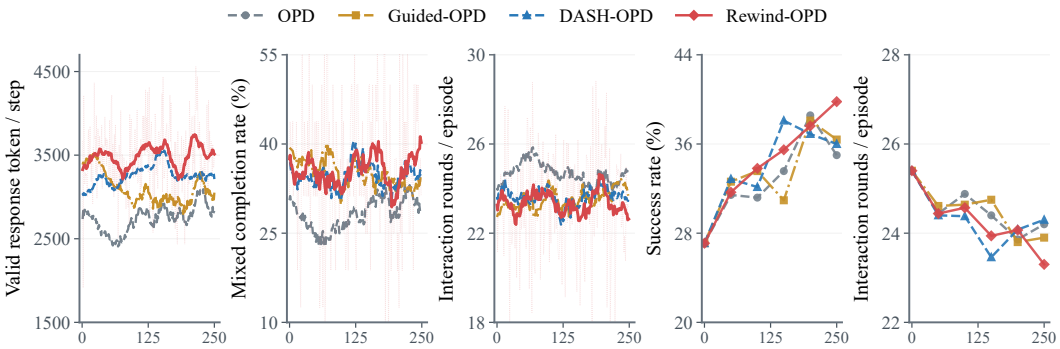

Figure 5: Training dynamics on ALFWorld with Qwen3-30B-A3B teacher and Qwen3-4B student.

Figure 5 examines training dynamics through the quality of collected rollouts and evaluations of student checkpoints. As shown in the left panel, REWIND-OPD produces the most valid response tokens per training step, providing more supervised tokens for each update. The middle two panels demonstrate that its mixed teacher–student rollouts achieve higher completion rates with fewer interaction rounds, particularly later in training, indicating that rewinding helps produce more effective training trajectories. Across checkpoint evaluations, REWIND-OPD's success rate rises steadily, while the baselines fluctuate or reach their peak at an early stage. Its average interaction rounds also remain low throughout training and show an overall downward trend. These results are consistent with the student learning more effective and concise action sequences from the reconstructed trajectories.

REWIND-OPD has the lowest average prompt-limit row fraction (2.2%, versus 6.5% for DASH-OPD) and the largest usable-response-token proxy (3,486 per update, versus 3,259 for DASH-OPD). The former measures length-limit pressure in optimizer batches, rather than the fraction of complete trajectories truncated. Across mixed-executor training rollouts, REWIND-OPD also has the highest mean completion rate (35.5%) and the fewest mean rounds (23.35). Its teacher-turn share decreases from 51.6% in the first 50 updates to 39.7% in the last 50, while rollout completion remains near 35%.

Student-only Seen success rises at every measured checkpoint through step 200, from 27.1% to 37.6%. The separately evaluated final checkpoint reaches 39.8% success and 23.3 rounds, giving REWIND-OPD the highest final success and shortest final trajectories among the compared methods. In contrast, Vanilla OPD, Guided-OPD, and DASH-OPD reach their highest recorded intermediate success at step 200, 200, and 150, respectively, and finish below REWIND-OPD. Its evaluation rounds decrease overall from 25.4 to 23.3, with small fluctuations between checkpoints.

E.1 TRAINING-TIME COMPARISON

Figure 6 compares total wall-clock distillation time, normalized to Vanilla OPD within each dataset and student scale. REWIND-OPD reduces training time by 37.5%/24% on ALFWorld, 6%/9% on WebShop, and 55%/35% on ScienceWorld for the 1.7B/4B students, respectively, while improving task performance. On ALFWorld at 1.7B, it takes 7 h 35 min, compared with 12 h 08 min for Vanilla OPD and 8 h 18 min for DASH-OPD, saving 8.6% relative to DASH-OPD. Its cost matches DASH-OPD on ALFWorld at 4B, and their normalized times differ by only 0.05 on WebShop at 1.7B. On ScienceWorld at 1.7B, REWIND-OPD takes 52.1% less time than DASH-OPD ($0.45\times$ versus $0.94\times$). Thus, regenerating a suffix does not necessarily increase total training time; the additional sampling can be offset by more efficient subsequent rollouts.

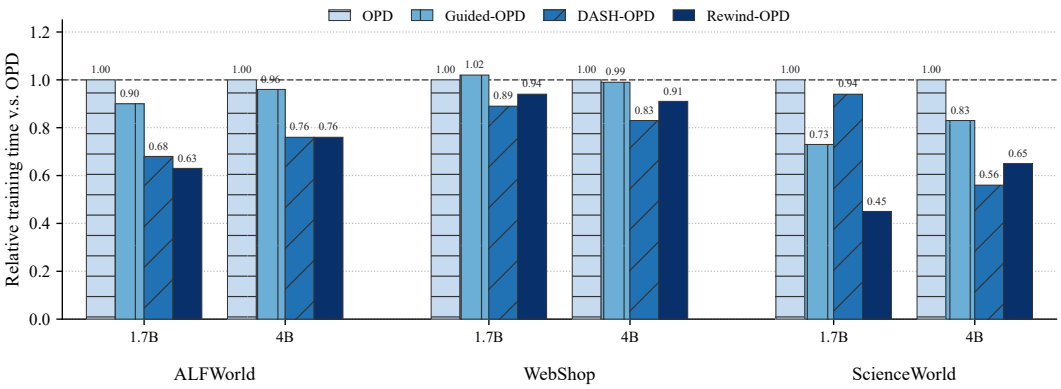

Figure 6: Relative wall-clock training time for 1.7B and 4B students distilled from Qwen3-30B-A3B. Each dataset–student setting normalizes Vanilla OPD to $1.00\times$; lower values indicate less training time.

E.2 Training Time with Post-Trained Qwen3-8B Teachers

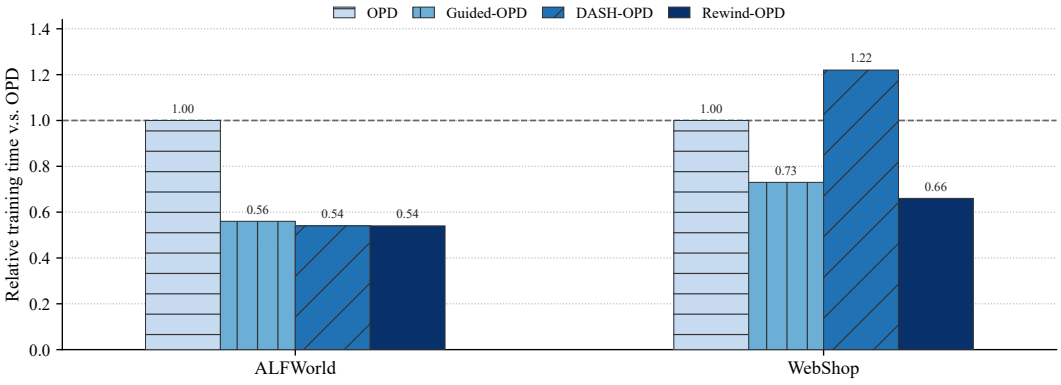

Figure 7: Relative wall-clock training time for Qwen3-4B distilled from GiGPO-trained Qwen3-8B teachers. Vanilla OPD is normalized to $1.00\times$ separately for each dataset.

Figure 7 shows the same pattern with task-specialized teachers. REWIND-OPD uses $0.54\times$ the Vanilla OPD time on ALFWorld and $0.66\times$ on WebShop, reducing training time by 46% and 34%, respectively. It matches DASH-OPD's time on ALFWorld and requires 45.9% less time on WebShop, while achieving higher task performance on both benchmarks (Tables 11–13). These results indicate that rollback can improve distillation without a substantial net training-time penalty relative to teacher intervention alone.

Why intervention can improve training efficiency. Both student and teacher are already deployed for OPD, but the teacher does not generate while the student executes a standard OPD rollout. Teacher intervention can use this otherwise idle generation capacity, and shared-prefix KV-cache reuse can reduce repeated prefill when supported by the serving implementation. Rollback adds suffix-generation work, yet it can also improve the trajectories used for learning. Without rollback, as in DASH-OPD, the teacher may remain stalled on an incorrect path and exhaust its intervention budget without resolving the underlying error. The ALFWorld dynamics in Figure 5 show higher rollout completion and fewer interaction rounds alongside stronger student performance, consistent with more effective supervision leading to shorter subsequent rollouts. This provides a plausible explanation for the large time reductions on ScienceWorld at 1.7B and WebShop with the post-trained teacher: correcting an unproductive continuation can avoid lengthy downstream generation. The

reported times measure the combined effect; they do not isolate cache reuse, learning progress, or rollback overhead.

F PROMPT PROTOCOL

We use the environment-specific user prompts from Xia et al. (2026). The same prompting protocol is used for teacher and student execution in both training and evaluation, with a shared action parser. The first three turns use the warm-up template. From turn four onward, the user message additionally contains the two most recent observation–action pairs in {action_history}, with each entry formatted as [Observation i: o_i, Action i: a_i]. This summary supplements, rather than replaces, the full user–assistant message history. Actions must appear inside <action> and </action>; no separate <think> or <thought> block is required. Prompts 1, 3, and 5 show the warm-up prompts for the three benchmarks. Prompts 2, 4, and 6 show the prompts that include the two most recent observation–action pairs.

Prompt 1: ALFWorld user prompt for turns 1–3.

```
You are an expert agent operating in the ALFRED Embodied Environment.
Your current observation is: {current_observation}
Your admissible actions of the current situation are:
[{admissible_actions}].

Now it's your turn to take an action.
You should first reason about the current situation.
Once you've finished your reasoning, you should choose the best
admissible action for the current step and present it within
<action> </action> tags.
Do not output any other text besides your reasoning and the action.
```

Prompt 2: ALFWorld user prompt from turn 4 onward.

```
You are an expert agent operating in the ALFRED Embodied Environment.
Your task is to: {task_description}
Prior to this step, you have already taken {step_count} step(s).
Below are the most recent {history_length} observations and the
corresponding actions you took: {action_history}
You are now at step {current_step} and your current observation is:
{current_observation}
Your admissible actions of the current situation are:
[{admissible_actions}].

Now it's your turn to take an action.
You should first reason about the current situation.
Once you've finished your reasoning, you should choose the best
admissible action for the current step and present it within
<action> </action> tags.
Do not output any other text besides your reasoning and the action.
```

Prompt 3: WebShop user prompt for turns 1–3.

```
You are an expert autonomous agent operating in the WebShop
e-commerce environment.
Your task is to: {task_description}.
Your current observation is: {current_observation}.
Your admissible actions of the current situation are:
[
{available_actions}
].

Now it's your turn to take one action for the current step.
You should first reason about the current situation, then consider which
admissible action best advances the shopping goal.
```

```
Once you've finished your reasoning, you should choose the best
admissible action for the current step and present it within
<action> </action> tags.
Do not output any other text besides your reasoning and the action.
```

Prompt 4: WebShop user prompt from turn 4 onward.

```
You are an expert autonomous agent operating in the WebShop
e-commerce environment.
Your task is to: {task_description}.
Prior to this step, you have already taken {step_count} step(s).
Below are the most recent {history_length} observations and the
corresponding actions you took: {action_history}
You are now at step {current_step} and your current observation is:
{current_observation}.
Your admissible actions of the current situation are:
[
{available_actions}
].

Now it's your turn to take one action for the current step.
You should first reason about the current situation, then think which
admissible action best advances the shopping goal.
Once you've finished your reasoning, you should choose the best
admissible action for the current step and present it within
<action> </action> tags.
Do not output any other text besides your reasoning and the action.
```

Prompt 5: ScienceWorld user prompt for turns 1–3.

```
    You are an expert agent operating in the ScienceWorld text
    environment.
    {task_description}
    Your current observation is: {current_observation}
    Available action commands: [{action_templates}]
    Available objects you can interact with: [{objects}]

    Now it's your turn to take an action. Combine an action command with
    appropriate object(s) to form a valid action.
    You should first reason about the current situation.
    Once you've finished your reasoning, you should choose the best
    valid
    action for the current step and present it within <action>
    </action> tags.
    Do not output any other text besides your reasoning and the action.
```

Prompt 6: ScienceWorld user prompt from turn 4 onward.

```
    You are an expert agent operating in the ScienceWorld text
    environment.
    {task_description}
    Prior to this step, you have already taken {step_count} step(s).
    Below are the most recent {history_length} observations and the
    corresponding actions you took: {action_history}
    You are now at step {current_step} and your current observation is:
    {current_observation}
    Available action commands: [{action_templates}]
    Available objects you can interact with: [{objects}]

    Now it's your turn to take an action. Combine an action command with
    appropriate object(s) to form a valid action.
    You should first reason about the current situation.
```

```
Once you've finished your reasoning, you should choose the best
valid
action for the current step and present it within <action>
</action> tags.
Do not output any other text besides your reasoning and the action.
```

G CASE STUDIES

We compare three inference branches of the same student model: student-only execution, Welford-triggered teacher intervention without rollback, and intervention after rollback. Within each case, the branches share the sampled student prefix; both intervention branches use the same trigger, teacher, and Welford-based adaptive return rule. Model parameters remain fixed. These are not comparisons between separately trained Vanilla OPD, DASH-OPD, and REWIND-OPD models. The selected rollback depths are fixed, and the teacher happens to generate two turns in each successful branch; this is an outcome of adaptive return, not a prescribed two-turn budget. All cases use the Qwen3-30B-A3B teacher and seed 42.

G.1 ALFWORLD: UNDOING AN INCORRECT OBJECT CHOICE

The task is to clean a plate and put it in the fridge. The Qwen3-1.7B student opens the fridge, takes a bowl instead of a plate, and cools it instead of cleaning it. Intervention is triggered after the cooling action, one turn after the incorrect object choice. Student-only execution returns the bowl and repeatedly looks around, ending in failure. Without rollback, the teacher also fails, issuing `look` for all 26 remaining turns.

Rolling back two turns removes the bowl pickup and cooling while preserving the useful prefix: the fridge is already open. The teacher visits a countertop and the sink, then returns control. In seven further turns, the student finds a plate, cleans it, and places it in the fridge (Table 14). The contrast illustrates why changing the executor at the detected disagreement is not always enough: rollback lets guidance begin before the incorrect object manipulation.

Table 14: ALFWorld task 107: three continuations of Qwen3-1.7B.

Shared student prefix: Go to fridge; open fridge; take bowl; cool bowl (trigger).		
Student Only	**Intervention without Rollback**	**Intervention after Rollback**
Return the bowl; check inventory; repeatedly `look`.	Teacher issues `look` for 26 turns; no return to the student.	Undo the last two turns, retaining the opened fridge. Teacher visits countertop 1 and sinkbasin 1, then returns control. Student finds plate 2, cleans it at the sink, and puts it in the fridge (7 turns).
Score 0; failure	Score 0; failure	**Score 100; success**

G.2 WEBSHOP: REVISING THE PRODUCT RATHER THAN ITS OPTIONS

The user requests a slim-fit, button-closure dress shirt in size 4X-large with a blue plaid pattern, priced below $50. The Qwen3-1.7B student opens a product whose available sizes stop at XL and selects green-black plaid, triggering intervention. Student-only execution then selects Large and buys it, receiving 75 points without success. The teacher intervening without rollback changes the pattern to black-blue plaid but selects XL and purchases the same product, also scoring 75.

A three-turn rollback removes the search, product click, and pattern selection. The teacher issues a revised search and opens a different product; the student then selects `4x-large tall` and `colony blue plaid` and completes a successful purchase (Table 15). Here, rollback changes the upstream product choice rather than only repairing downstream options. Returning to search

was available in the no-rollback branch; the observed failure was not an environmental prohibition, but a continuation that stayed with the unsuitable product.

Table 15: WebShop session 4174: three continuations of Qwen3-1.7B.

Shared student prefix: Search; open product B07FSTSW14 (no 4XL option); select green-black plaid (trigger).

Student Only	**Intervention without Rollback**	**Intervention after Rollback**
Student selects Large and purchases the product.	Teacher selects black-blue plaid, then XL, and purchases the same product (3 turns).	Undo all three turns. Teacher searches again and opens B079KBM5DZ (2 turns), then returns control. Student selects `4x-large tall` and `colony blue plaid`, then purchases (3 turns).
Score 75; failure	Score 75; failure	**Score 100; success**

G.3 ScienceWorld: Reorganizing Exploration

The task is to find an animal, focus on it, and move it to the yellow box in the workshop. The Qwen3-4B student first explores the art studio, and intervention is triggered when it opens a cupboard. Student-only execution later finds a bee but repeatedly fails to transfer it to the destination, ending with 67 points. Without rollback, the teacher continues inspecting the cupboard and drawer, and the branch eventually focuses on a peach tree rather than completing the animal-transfer task, scoring 17.

The requested five-turn rollback is limited by the three-turn student prefix, so the actual rollback is three turns. From the initial state, the teacher visits the workshop and observes the yellow box before returning control. The student subsequently finds and picks up a bee, returns to the workshop, and places it in the box, obtaining 100 points (Table 16). This successful branch reorganizes exploration around the destination, but still contains invalid navigation attempts. It does not imply that visiting the art studio is inherently wrong or that the repaired continuation is error-free.

Table 16: ScienceWorld task 19: three continuations of Qwen3-4B.

Shared student prefix: Enter art studio; look around; open cupboard (trigger).

Student Only	**Intervention without Rollback**	**Intervention after Rollback**
Student finds and focuses on a bee, but repeated transfer attempts are rejected.	Teacher continues cupboard/drawer exploration (11 turns); the branch eventually focuses on a peach tree.	Undo all three turns. Teacher visits the workshop and observes the yellow box (2 turns), then returns control. Student finds a bee, picks it up, returns via the hallway, and moves it to the yellow box (25 turns).
Score 67; failure	Score 17; failure	**Score 100; success**

Interpretation. In these selected cases, the trigger occurs after decisions have already shaped the environment and subsequent context. Direct intervention can continue along that branch, whereas rollback permits guidance from an earlier state. The examples illustrate this mechanism, not its average effectiveness, sampling cost, or MCOL localization accuracy. In particular, the successful fixed-depth branches do not establish that their rollback points are optimal.