STOCHASTIC TEACHER INTERVENTION FOR AGENTIC ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) efficiently transfers capabilities from a stronger teacher to a student language model through dense token-level supervision on student-generated rollouts and has shown promise on complex tasks such as mathematical reasoning. However, in multi-turn agentic tasks, student decisions shape subsequent observations, causing early errors to accumulate across turns. The resulting trajectories can drift away from the teacher's rollout distribution, making the teacher's token-level supervision less reliable or even counterproductive for OPD training. To address this issue, we introduce STI-OPD, a stochastic teacher intervention framework for multi-turn agentic OPD. During multi-turn interaction, STI-OPD uses teacher intervention guided by teacher–student policy discrepancy to replace the student's proposed action with a teacher-generated one to maximize the acquisition of reliable supervision. We further develop a stochastic intervention strategy, addressing the limitations of previous threshold-based or fixed-schedule approaches, that estimates policy discrepancy using KL divergence and maps it to an intervention probability. By sampling whether to intervene from this probability, STI-OPD adaptively balances teacher control with student exploration. To learn from the resulting mixed-policy trajectories, we introduce an Importance-Weighted Reverse KL objective that corrects the token sampling mismatch between teacher-generated responses and the student policy to preserve the original OPD objective. Across tool-integrated reasoning and long-horizon interaction, STI-OPD outperforms the strongest prior OPD baseline on every evaluated benchmark and student size. Ablations further show that both discrepancy-guided intervention and importance weighting contribute to these gains.

1 INTRODUCTION

On-policy distillation (OPD) (Agarwal et al., 2024; Gu et al., 2024; Yang et al., 2025; DeepSeek-AI, 2025; Lu & Lab, 2025; Zhao et al., 2026) transfers capabilities from a stronger teacher to a student large language model (LLM) by minimizing token-level KL divergence on student-generated rollouts. By supervising contexts induced by the student's own policy, OPD reduces the train–inference mismatch and provides dense token-level guidance, in contrast to the typically sparse outcome-level rewards of reinforcement learning with verifiable rewards (RLVR), leading to promising results on complex tasks such as mathematical reasoning (Lu & Lab, 2025; Jin et al., 2026). Extending OPD to multi-turn agentic tasks (Yang et al., 2024; Shridhar et al., 2021), however, introduces a trajectory drift challenge: because each action shapes subsequent observations, an incorrect tool call, or even a valid action unlikely under the teacher policy, can lead the student into *off-teacher-distribution states* rarely encountered in teacher-generated rollouts (Zhong et al., 2026; Wang et al., 2026b), as illustrated in Figure 1. As such deviations accumulate across turns, teacher–student policy discrepancy can increase, making the teacher's token-level guidance on the resulting histories less reliable or even counterproductive (Li et al., 2026b). Consequently, supervising student-generated tokens alone may be insufficient to control the distribution shift induced by multi-turn interaction.

Existing methods address this instability through two strategies. One masks or reweights the distillation loss using signals derived from KL divergence (Zhong et al., 2026; Zhou et al., 2026a) or model confidence (Zheng et al., 2026). The other controls training-trajectory construction by progressively adjusting rollout depth (Wang et al., 2026b) or mixing teacher- and student-generated turns through a curriculum (Li et al., 2026a). However, loss filtering will lose supervision signals and cannot

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.

directly correct the action that determines the next environment transition. Predefined curricula and rollout-depth control can prevent the student's exposure to unreliable supervision, but do not adaptively assess whether its current response requires correction. Recent work has further explored *teacher intervention* that allows the teacher to take over from the students' generation at specific points (Xu et al., 2026). By replacing the student's proposed action with a teacher-generated one, these methods could limit the accumulation of trajectory drift and provide more reliable supervision on subsequent turns.

However, making teacher intervention effective in multi-turn OPD still requires addressing two key challenges. First, ***when should the teacher intervene?*** Predefined intervention schedules fail to account for immediate changes in teacher–student policy discrepancy during multi-turn interactions (Li et al., 2026a). On the other hand, recent adaptive approaches use discrepancy estimates to guide intervention through decision boundaries, hysteretic thresholds, or lookahead validation (Wang et al., 2026a; Xia et al., 2026; Chen et al., 2026). However, these estimates are subject to statistical noise, and threshold-based decisions can amplify estimation errors into abrupt changes in teacher control. A poor intervention strategy may therefore miss opportunities to correct trajectory drift or intervene unnecessarily, increasing teacher dependence and computational cost.

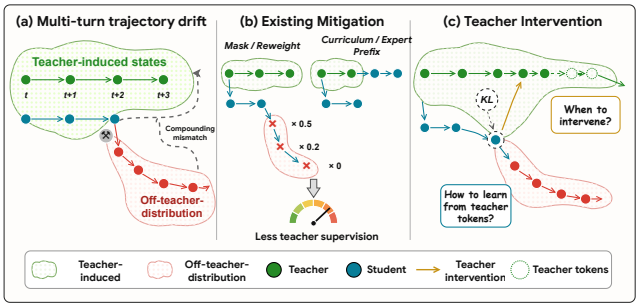

Figure 1: Illustration of trajectory-level distribution shift and turn-level teacher intervention in on-policy distillation.

Second, ***how should the student learn from teacher-generated tokens?*** Teacher intervention introduces tokens sampled from the teacher policy into training, whereas standard OPD defines its reverse-KL objective under student-generated tokens. Existing intervention-based OPD methods handle these teacher-generated tokens using different surrogate objectives (Li et al., 2026a; Xu et al., 2026; Chen et al., 2026). Some adopt forward-KL or SFT-style objectives, which depart from the reverse-KL formulation of standard OPD and may affect the retention of previously acquired capabilities. Others directly apply a sampled reverse-KL objective to teacher-generated tokens. However, because these tokens are sampled from the teacher rather than the student policy, directly evaluating the standard reverse-KL integrand does not recover its intended expectation under the student policy. Thus, learning from teacher intervention requires correcting this token-sampling mismatch while preserving the conditional reverse-KL objective of OPD.

To address these challenges, we propose STI-OPD, a discrepancy-guided Stochastic Teacher Intervention framework for multi-turn agentic OPD. We first develop a stochastic intervention strategy that samples whether to intervene based on teacher–student policy discrepancy. Specifically, STI-OPD uses KL divergence estimated by k_3 estimator (Schulman, 2020) to measure the divergence between the teacher and student policies at each turn, and maps this estimate into an intervention probability. This probability increases smoothly with the discrepancy, favoring teacher intervention when the policies diverge while preserving student exploration, without a hand-tuned threshold or fixed intervention budget. To learn from the resulting mixed-policy trajectories, we further introduce an importance-weighted reverse KL objective that uses importance sampling (Schulman et al., 2017) to correct the token sampling mismatch between teacher-generated responses and the student policy while preserving the original OPD objective formulation. Intervention thus controls which histories are visited, while importance weighting corrects the token-sampling mismatch introduced by teacher generation.

Extensive experiments are conducted in two multi-turn agentic settings, spanning tool-integrated reasoning (Zhang & Math-AI, 2024; 2025; 2026; Dekoninck et al., 2026; Rein et al., 2023; Jain et al., 2025) and long-horizon agentic interaction (Shridhar et al., 2021; Wang et al., 2022). STI-OPD outperforms all distillation baselines on each benchmark and at both student scales, with its largest margins on long-horizon tasks, where compounding error is most severe. Ablations show that different components of STI-OPD contribute to these gains and further analysis reveals the mechanism, scalability, and efficiency of STI-OPD.

2 Related Work

On-policy Distillation for Multi-Turn Agents. Small errors compound across turns and move student rollouts into histories on which teacher supervision is unreliable (Zhong et al., 2026; Wang et al., 2026b; Li et al., 2026a). Existing methods manage this instability by controlling either the supervision signal or the rollout distribution. The first group reweights or masks the distillation loss: SCOPE (Zheng et al., 2026) calibrates supervision using trajectory correctness together with teacher or student confidence, and SAGE-OPD (Zhou et al., 2026a) applies guidance selectively within multi-turn rollouts. The second group reshapes which trajectories the student is trained on: TCOD (Wang et al., 2026b) progressively expands the trajectory depth exposed to the student, Guided-OPD (Li et al., 2026a) mixes teacher- and student-generated turns by a curriculum, TurnOPD (Zhou et al., 2026b) allocates rollout depth and loss weight across turns, and ReOPD (Liao et al., 2026) replays teacher trajectories as prefixes, trading student occupancy against teacher reliability. Closest to STI-OPD, Relay-OPD (Xu et al., 2026) temporarily hands generation to the teacher after detecting a failed prefix, and FutureBridge-OPD (Chen et al., 2026) scores a short teacher bridge by its effect on the subsequent student continuation. Methods in the second group that hand control to the teacher face an estimation problem that neither group addresses: the tokens produced during intervention are sampled from the teacher, yet they are supervised with an integrand that assumes student sampling.

Expert Intervention and Distribution Shift. Teacher intervention is an instance of interactive imitation learning, where the learner's own actions change its future state distribution. DAgger repeatedly queries an expert on states induced by the learner (Ross et al., 2011), scheduled sampling gradually replaces ground-truth prefixes with model-generated prefixes in autoregressive prediction (Bengio et al., 2015), and more recent methods treat expert interventions themselves as learning signals (Luo et al., 2024). The same idea has been carried into LLM distillation: MOTAB monitors student trajectories, backtracks when they cross an adaptive boundary, and invokes the teacher from the last safe state (Wang et al., 2026a), while SMRC-SD applies privileged reference guidance to multi-turn agents only when the student's execution state matches a supported reference state (Liu et al., 2026). The intervention rules closest to ours are all specified by hand, making them inflexible and sensitive to noisy turn-level estimates: Guided-OPD follows a predefined curriculum (Li et al., 2026a), Relay-OPD uses a continuation-asymmetry trigger (Xu et al., 2026), FutureBridge-OPD validates guidance through future rollouts (Chen et al., 2026), and DASH-OPD switches executors on hysteretic thresholds over accumulated log-probability-ratio evidence (Xia et al., 2026). STI-OPD differs in both respects: each decision is stochastic, with a probability read off the instantaneous teacher–student KL divergence rather than a threshold, and the estimation mismatch that intervention introduces is corrected by importance sampling so that teacher-generated tokens still serve the target OPD objective.

3 Methodology

As illustrated in Figure 2, STI-OPD first estimates teacher–student policy discrepancy from a student-proposed response using the k_3 estimator. The resulting score determines the probability of retaining the proposal or replacing it with a teacher-regenerated response. The selected response is executed to update the interaction history and added to the training buffer. During optimization, an importance ratio corrects the sampling distribution of teacher-generated tokens to account for the generating policy, recovering the conditional reverse-KL expectation for teacher-generated tokens.

3.1 Problem Formulation

Multi-Turn Agentic Task. Given a query $q \sim \mathcal{D}$, an agent interacts with an environment over multiple turns to solve the task (Yao et al., 2023; Shridhar et al., 2021; Wang et al., 2022). At turn t, the agent observes o_t and samples a response $a_t \sim \pi_\theta(\cdot \mid h_t)$ from its policy π_θ, conditioned on the interaction history $h_t = (q, o_1, a_1, \ldots, o_{t-1}, a_{t-1}, o_t)$. Each response contains natural-language reasoning and possible subsequent tool calls, environment actions, and the final answer. For each tool call or environment action, the environment executes the operation specified in a_t and returns a new observation $o_{t+1} \sim P(\cdot \mid h_t, a_t)$. This loop continues until termination, yielding a trajectory $\tau = (q, o_1, a_1, \ldots, o_H, a_H)$ of H turns, with terminal feedback omitted. We denote its distribution by $p_{\pi_\theta}(\tau \mid q)$. Within each turn, the response $a_t = (x_{t,1}, \ldots, x_{t,L_t})$ is generated autoregressively,

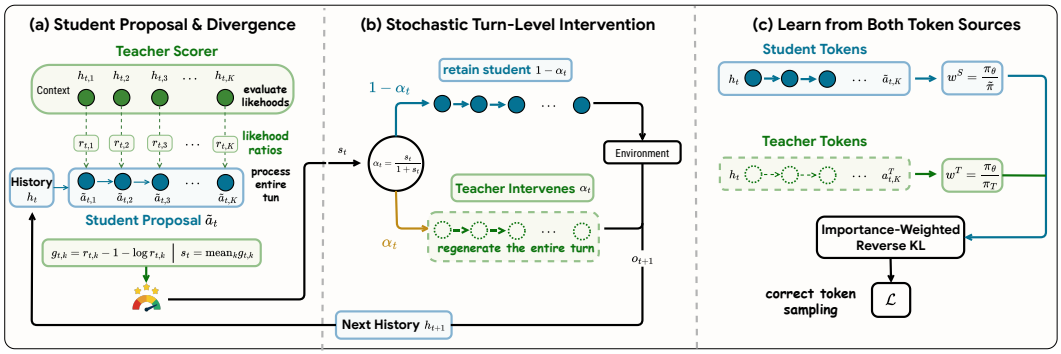

Figure 2: Overview of STI-OPD. A non-negative k_3 discrepancy estimate determines whether to retain the student proposal or replace it with a teacher-generated response before execution. The executed response is used for training, with importance ratios recovering the conditional reverse-KL expectation for teacher-generated tokens.

where L_t is its length. Given the token-level context $c_{t,k} = (h_t, x_{t,<k})$, its probability factorizes as

$$\pi_\theta(a_t \mid h_t) = \prod_{k=1}^{L_t} \pi_\theta(x_{t,k} \mid c_{t,k}). \tag{1}$$

On-Policy Distillation. Let π_T be a fixed teacher policy and π_θ the trainable student policy. In OPD, the teacher provides token-level supervision on trajectories generated by the student. Compared with offline distillation from fixed expert trajectories, OPD reduces the mismatch between training contexts and those encountered by the student at inference time (Agarwal et al., 2024; Gu et al., 2024; Lu & Lab, 2025). Using reverse KL divergence, the ideal multi-turn OPD objective is

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, \tau\sim p_{\pi_\theta}(\cdot|q)} \left[\sum_{t=1}^{H} \sum_{k=1}^{L_t} D_{\mathrm{KL}}(\pi_\theta(\cdot \mid c_{t,k}) \,\|\, \pi_T(\cdot \mid c_{t,k})) \right]. \tag{2}$$

With a fixed context c, to reduce the heavy overhead of computing the full KL divergence, each reverse-KL term can adopt a sampling form $\mathbb{E}_{x\sim\pi_\theta(\cdot|c)} [\log \pi_\theta(x \mid c) - \log \pi_T(x \mid c)]$. Sampled OPD therefore relies on both student-induced contexts and student-generated tokens.

Policy Discrepancy along Trajectories. As discussed in § 1, multi-turn interaction can drive the student toward histories rarely encountered under teacher rollouts, where teacher guidance may become less reliable. Large policy discrepancies at these histories can hinder effective distillation (Li et al., 2026b). We quantify the teacher–student policy discrepancy at history h_t by

$$\kappa(h_t) = D_{\mathrm{KL}}(\pi_\theta(\cdot \mid h_t) \,\|\, \pi_T(\cdot \mid h_t)), \tag{3}$$

where each policy defines a distribution over complete turn-level responses. Because student actions shape subsequent observations, a deviation at one turn can affect all later decisions. The resulting distribution shift may compound across turns and further increase policy discrepancy (Wang et al., 2026b; Li et al., 2026a). This motivates intervening before a proposed response is executed, allowing teacher guidance to influence the next environment transition rather than only the training loss.

3.2 Stochastic Teacher Intervention

An effective intervention strategy should respond to the policy discrepancy at the current turn while accounting for uncertainty in its estimate. Predefined schedules overlook immediate changes in discrepancy, while hard thresholds can turn estimation errors into abrupt changes in teacher control (Li et al., 2026a; Wang et al., 2026b). Poor intervention decisions may therefore allow trajectory drift to persist or introduce unnecessary teacher intervention. These considerations motivate combining an unbiased and low-variance discrepancy estimate with an intervention probability that varies smoothly with the estimated score.

Turn-Level Teacher Intervention. We first specify how intervention modifies a rollout. Let $\pi_{\bar{\theta}}$ denote the rollout student, which may be a lagged copy of the trainable policy π_θ. In the strictly on-policy setting, $\pi_{\bar{\theta}} = \pi_\theta$. At turn t, the rollout student first proposes a complete response $\widetilde{a}_t \sim \pi_{\bar{\theta}}(\cdot \mid h_t)$. An indicator $z_t \in \{0, 1\}$ determines whether to retain this proposal ($z_t = 0$) or replace it with a teacher

response $a_t^T \sim \pi_T(\cdot \mid h_t)$ generated from the same history ($z_t = 1$). The resulting trajectory is

$$\tau = (q, o_1, a_1, \ldots, o_H, a_H), \qquad a_t = \begin{cases} \widetilde{a}_t, & z_t = 0, \\ a_t^T, & z_t = 1. \end{cases} \tag{4}$$

Each intervention replaces the entire response, keeping its reasoning and selected operation consistent. For a tool call or environment action, the environment executes the operation specified in a_t and returns $o_{t+1} \sim P(\cdot \mid h_t, a_t)$. If the episode continues, the student resumes generation from $h_{t+1} = (h_t, a_t, o_{t+1})$. Thus, h_t consists of previously executed responses from either the student or the teacher, together with the corresponding environment observations. Both policies condition on this same history at the current turn. A discarded student proposal is neither appended to the history nor included in the training buffer.

An optimal intervention strategy must balance teacher guidance with autonomous student execution. Unlike random intervention with predefined probabilities (Li et al., 2026a) or switching at fixed thresholds, we propose a discrepancy-conditioned *stochastic intervention strategy*. The intervention probability varies smoothly with the estimated teacher–student discrepancy, allowing teacher guidance to adapt at each turn while preserving opportunities for student execution (Figure 2 (b)).

Discrepancy Estimation. Following the policy discrepancy defined in Equation (3), we assess the current student proposal $\widetilde{a}_t = (\widetilde{x}_{t,1}, \ldots, \widetilde{x}_{t,\widetilde{L}_t})$ using token-level reverse KL. At each token position k, we measure the discrepancy between the student's and teacher's next-token distributions conditioned on $\widetilde{c}_{t,k} = (h_t, \widetilde{x}_{t,<k})$ using $D_{\mathrm{KL}}(\pi_{\bar{\theta}}(\cdot \mid \widetilde{c}_{t,k}) \| \pi_T(\cdot \mid \widetilde{c}_{t,k}))$.

Exact evaluation requires summing over the full vocabulary at every position. For efficiency, the common practice is to use the KL estimator (Schulman, 2020), which requires only the teacher and student probabilities of each sampled token $\widetilde{x}_{t,k}$. We define the teacher-to-student likelihood ratio as

$$r_{t,k} = \frac{\pi_T(\widetilde{x}_{t,k} \mid \widetilde{c}_{t,k})}{\pi_{\bar{\theta}}(\widetilde{x}_{t,k} \mid \widetilde{c}_{t,k})}. \tag{5}$$

To provide an unbiased and low-variance estimate of the token-level reverse KL, we use the k_3 estimator (Schulman, 2020),

$$g_{t,k} = r_{t,k} - 1 - \log r_{t,k}. \tag{6}$$

For $r_{t,k} > 0$, $g_{t,k}$ is non-negative because $\log r_{t,k} \leq r_{t,k} - 1$. Under common support, $g_{t,k}$ is a pointwise non-negative and unbiased estimator of the token-level reverse KL under student sampling. Moreover, in the near-policy regime, k_3 removes the first-order fluctuation present in k_1, yielding lower-order variance, which makes it a more stable signal for intervention. We aggregate the token-level estimates into the length-normalized turn score to remove the direct scaling of the score with response length.

$$s_t = \frac{1}{\widetilde{L}_t} \sum_{k=1}^{\widetilde{L}_t} g_{t,k}. \tag{7}$$

The score s_t summarizes token-level discrepancy along the sampled proposal and serves as a length-normalized proxy for turn-level policy discrepancy. It is distinct from the full response-level KL $\kappa(h_t)$ in Equation (3).

Discrepancy-Conditioned Intervention. To make intervention responsive to discrepancy without imposing a hard decision boundary, we set the intervention odds equal to s_t. In this paper, we adopt the following simple monotone mapping:

$$\alpha_t = \frac{s_t}{1 + s_t}, \qquad z_t \sim \mathrm{Bernoulli}(\alpha_t), \tag{8}$$

where $z_t = 1$ denotes teacher intervention. This mapping assigns higher intervention probabilities to turns with larger estimated discrepancies, without a hand-tuned threshold. When the teacher and student policies are close, the k_3-based score is second order in the likelihood-ratio deviation, so α_t remains close to zero to allow student execution and exploration. When the policies diverge, α_t approaches one, increasing the likelihood of teacher intervention to correct the student's trajectory but still allowing nonzero student execution probability with the stochastic Bernoulli sampling.

3.3 Learning from Teacher Intervention

Teacher intervention helps reduce error propagation across turns, but it produces mixed-policy trajectories in which executed responses may come from either the student or the teacher. Standard

OPD assumes student-generated trajectories, with the sampled reverse-KL loss taking an expectation over tokens drawn from $\pi_\theta(\cdot \mid c)$. In intervened turns, however, the executed tokens are drawn from $\pi_T(\cdot \mid c)$. Directly applying this loss to teacher-generated tokens without correction does not estimate the intended conditional reverse KL. We therefore use importance-weighted reverse KL to correct this token-sampling mismatch and recover the expectation under the student policy (Figure 2 (c)).

Importance-Weighted Reverse KL. For a teacher-generated token, we define the importance ratio as

$$w_\theta(x, c) = \frac{\pi_\theta(x \mid c)}{\pi_T(x \mid c)}. \tag{9}$$

Assuming that $\pi_\theta(\cdot \mid c)$ is absolutely continuous with respect to $\pi_T(\cdot \mid c)$, importance sampling yields

$$\begin{aligned} \mathbb{E}_{x\sim\pi_T(\cdot|c)}\left[w_\theta(x,c)\log\frac{\pi_\theta(x \mid c)}{\pi_T(x \mid c)}\right] &= \sum_x \pi_T(x \mid c)\frac{\pi_\theta(x \mid c)}{\pi_T(x \mid c)}\log\frac{\pi_\theta(x \mid c)}{\pi_T(x \mid c)} \\ &= \mathbb{E}_{x\sim\pi_\theta(\cdot|c)}\left[\log\frac{\pi_\theta(x \mid c)}{\pi_T(x \mid c)}\right] = D_{\mathrm{KL}}(\pi_\theta(\cdot \mid c)\,\|\,\pi_T(\cdot \mid c))\,. \end{aligned} \tag{10}$$

Teacher-generated tokens can therefore estimate the same conditional reverse-KL term once their token-sampling distribution is corrected.

Combined Training Objective. Once the task is complete, the full executed trajectory is added to the training buffer. Consequently, each turn contributes either the retained student proposal when $z_t = 0$ or the regenerated teacher response when $z_t = 1$. For a retained student token, no sampling correction is required because $\widetilde{x}_{t,k}$ is already sampled from π_θ. For a teacher-generated token, we define the importance ratio as $\omega^T_{t,k}(\theta) = \pi_\theta(x^T_{t,k} \mid c^T_{t,k})/\pi_T(x^T_{t,k} \mid c^T_{t,k})$. Following the standard OPD formulation based on sampled tokens, the optimization objective is the empirical mean of the sampled reverse-KL terms over all generated tokens:

$$\widehat{\mathcal{L}}_{\text{STI-OPD}}(\theta) = \frac{1}{N_{\text{tok}}}\sum_{t=1}^{H}\left[(1-z_t)\sum_{k=1}^{\widetilde{L}_t}\ell_\theta(\widetilde{x}_{t,k},\widetilde{c}_{t,k}) + z_t\sum_{k=1}^{L^T_t}\omega^T_{t,k}(\theta)\ell_\theta(x^T_{t,k},c^T_{t,k})\right], \tag{11}$$

where $\ell_\theta(x, c) = \log \pi_\theta(x \mid c) - \log \pi_T(x \mid c)$ is the token-level student-to-teacher log-ratio, whose expectation under $x \sim \pi_\theta(\cdot \mid c)$ equals the conditional reverse KL, and $N_{\text{tok}} = \sum_{t=1}^{H}((1-z_t)\widetilde{L}_t + z_t L^T_t)$ is the number of optimization tokens in the episode. For retained student turns, the first term is exactly the standard sampled OPD objective because the tokens are drawn from π_θ. For intervened turns, the second term uses importance weighting to correct the change in token-sampling distribution from π_T to π_θ. By Equation (10), both branches therefore estimate the same conditional reverse-KL objective at their respective sampled contexts. Importantly, the identity in Equation (10) holds at a fixed context. The importance correction accounts for the token-sampling distribution of the teacher branch, but does not undo the adaptive intervention decision or correct the distribution over contexts induced by intervention, including teacher-generated prefixes and subsequent interaction histories.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate STI-OPD in two complementary settings: tool-integrated reasoning and long-horizon agentic tasks. For tool-integrated reasoning, the training set contains 34,501 prompts drawn from 14,116 examples from the English subset of DAPO-Math, 10,385 examples from the code subset of Skywork-OR1-RL-Data, and 10,000 examples from MegaScience (Yu et al., 2025; He et al., 2025; Fan et al., 2025). We evaluate mathematical reasoning on AIME 2024–2026 and HMMT February 2026 (Zhang & Math-AI, 2024; 2025; 2026; Dekoninck et al., 2026), scientific reasoning on GPQA (Rein et al., 2023), and code generation on LiveCodeBench v6 (Jain et al., 2025). For long-horizon interaction, we train and evaluate on the official splits of ALFWorld (Shridhar et al., 2021) and ScienceWorld (Wang et al., 2022). Both environments require the agent to condition successive actions on environment feedback, which makes them a direct test of robustness to compounding inter-turn error.

Baselines. We compare STI-OPD with the original student checkpoint (Vanilla), standard OPD (Agarwal et al., 2024; Gu et al., 2024), and five OPD methods that improve training stability

Table 1: Results on tool-integrated reasoning benchmarks. Bold and underlined values denote the best and second-best results, respectively.

Method	AIME24	AIME25	AIME26	HMMT26	GPQA	LiveCodeBench v6
	Avg@32				Avg@8	Avg@8
Qwen3-4B						
TEACHER	$71.7_{\pm 0.3}$	$65.0_{\pm 0.3}$	$60.0_{\pm 0.3}$	$39.8_{\pm 0.3}$	$50.9_{\pm 0.3}$	$70.5_{\pm 0.3}$
Qwen3-0.6B						
VANILLA	$8.3_{\pm 0.4}$	$15.8_{\pm 0.6}$	$9.6_{\pm 0.5}$	$10.2_{\pm 0.4}$	$26.0_{\pm 0.7}$	$20.7_{\pm 0.5}$
OPD	$11.2_{\pm 0.5}$	$18.5_{\pm 0.5}$	$12.4_{\pm 0.4}$	$11.8_{\pm 0.6}$	$27.1_{\pm 0.6}$	$23.5_{\pm 0.7}$
SAGE-OPD	$14.5_{\pm 0.6}$	$21.0_{\pm 0.7}$	$14.8_{\pm 0.6}$	$13.9_{\pm 0.5}$	$28.6_{\pm 0.5}$	$26.2_{\pm 0.6}$
GUIDED-OPD	$18.2_{\pm 0.7}$	$25.1_{\pm 0.6}$	$17.5_{\pm 0.7}$	$16.2_{\pm 0.5}$	$29.8_{\pm 0.6}$	$31.4_{\pm 0.8}$
RELAY-OPD	$16.8_{\pm 0.5}$	$23.4_{\pm 0.8}$	$16.1_{\pm 0.5}$	$14.7_{\pm 0.6}$	$28.2_{\pm 0.7}$	$29.1_{\pm 0.7}$
TCOD	$20.4_{\pm 0.8}$	$27.9_{\pm 0.7}$	$19.3_{\pm 0.6}$	$15.8_{\pm 0.7}$	$\underline{31.2}_{\pm 0.6}$	$33.6_{\pm 0.8}$
FUTUREBRIDGE-OPD	$\underline{22.1}_{\pm 0.7}$	$\underline{30.2}_{\pm 0.8}$	$\underline{20.8}_{\pm 0.7}$	$\underline{17.5}_{\pm 0.5}$	$30.9_{\pm 0.5}$	$\underline{35.8}_{\pm 0.9}$
STI-OPD	$\mathbf{24.2}_{\pm \mathbf{0.4}}$	$\mathbf{33.5}_{\pm \mathbf{0.4}}$	$\mathbf{22.7}_{\pm \mathbf{0.3}}$	$\mathbf{18.6}_{\pm \mathbf{0.3}}$	$\mathbf{32.5}_{\pm \mathbf{0.3}}$	$\mathbf{38.8}_{\pm \mathbf{0.4}}$
Qwen3-1.7B						
VANILLA	$42.9_{\pm 0.6}$	$35.8_{\pm 0.7}$	$32.8_{\pm 0.5}$	$25.0_{\pm 0.6}$	$34.2_{\pm 0.8}$	$44.1_{\pm 0.7}$
OPD	$44.5_{\pm 0.7}$	$37.1_{\pm 0.6}$	$33.9_{\pm 0.6}$	$26.1_{\pm 0.5}$	$35.4_{\pm 0.7}$	$45.3_{\pm 0.6}$
SAGE-OPD	$46.2_{\pm 0.8}$	$38.6_{\pm 0.8}$	$34.8_{\pm 0.7}$	$27.0_{\pm 0.6}$	$37.1_{\pm 0.6}$	$46.7_{\pm 0.8}$
GUIDED-OPD	$48.6_{\pm 0.7}$	$41.0_{\pm 0.7}$	$37.0_{\pm 0.8}$	$28.9_{\pm 0.7}$	$38.8_{\pm 0.7}$	$48.4_{\pm 0.9}$
RELAY-OPD	$47.3_{\pm 0.6}$	$39.8_{\pm 0.9}$	$36.2_{\pm 0.6}$	$27.8_{\pm 0.5}$	$36.9_{\pm 0.8}$	$47.5_{\pm 0.7}$
TCOD	$49.8_{\pm 0.8}$	$\underline{42.1}_{\pm 0.8}$	$\underline{38.3}_{\pm 0.7}$	$28.5_{\pm 0.6}$	$\underline{40.2}_{\pm 0.7}$	$49.3_{\pm 0.8}$
FUTUREBRIDGE-OPD	$\underline{50.7}_{\pm 0.7}$	$41.4_{\pm 0.7}$	$37.9_{\pm 0.6}$	$\underline{29.1}_{\pm 0.8}$	$39.9_{\pm 0.6}$	$\underline{49.9}_{\pm 0.7}$
STI-OPD	$\mathbf{52.2}_{\pm \mathbf{0.4}}$	$\mathbf{44.7}_{\pm \mathbf{0.5}}$	$\mathbf{39.7}_{\pm \mathbf{0.3}}$	$\mathbf{31.4}_{\pm \mathbf{0.6}}$	$\mathbf{42.2}_{\pm \mathbf{0.4}}$	$\mathbf{52.8}_{\pm \mathbf{0.4}}$

through selective guidance or curricula: SAGE-OPD (Zhou et al., 2026a), GUIDED-OPD (Li et al., 2026a), RELAY-OPD (Xu et al., 2026), TCOD (Wang et al., 2026b), and FUTUREBRIDGE-OPD (Chen et al., 2026). All distillation methods share the same teacher and student initializations, prompts, rollout budget, and evaluation protocol, so the main comparison controls for differences in data and rollout configuration. We additionally provide a comparison under a matched teacher-intervention budget in § C.4.

Implementation Details. We use Qwen3-0.6B and Qwen3-1.7B as students and Qwen3-4B as the teacher (Yang et al., 2025) and additional experiments with Qwen3-32B examine scaling to a stronger teacher. Evaluation uses the student policy alone, with no teacher intervention. All reported numbers are means over multiple independent runs (refer to the detailed experimental results), and the reported uncertainties are standard deviations across those runs. Complete hyperparameters and prompts are given in §§ B.3 and B.4.

4.2 MAIN RESULTS

STI-OPD consistently outperforms all distillation baselines. As shown in Tables 1 and 2, STI-OPD attains the best mean score among the distillation methods in every reported cell, at both student sizes, spanning mathematical and scientific reasoning, code generation, and interactive decision making. The advanced OPD baselines generally improve over standard OPD, yet STI-OPD remains ahead of the most competitive of them, TCOD and FUTUREBRIDGE-OPD.

STI-OPD improves tool-integrated reasoning across domains. On the tool-integrated benchmarks, STI-OPD improves over standard OPD and over the best-performing baseline on every dataset. The gains extend beyond mathematics to GPQA and LiveCodeBench v6, which suggests that the method helps both structured reasoning and tool-supported generation. The margin over OPD is larger for the 0.6B student than for the 1.7B student, consistent with adaptive teacher intervention being most useful when the student is less capable and therefore more likely to leave teacher-familiar contexts.

STI-OPD provides its largest gains on long-horizon agentic tasks. Long-horizon tasks like ALFWorld and ScienceWorld are representative tasks because actions change subsequent observations and errors propagate across turns. The largest gains therefore fall on the longest-horizon settings, ScienceWorld and the ALFWorld types that require the most environment interaction, which is the direction turn-adaptive intervention predicts under trajectory-level distribution shift. On some tasks,

Table 2: Success rates on long-horizon agentic benchmarks. We report results for six ALFWorld task categories and the average across ScienceWorld tasks. Bold and underlined values denote the best and second-best results, respectively.

Method	ALFWorld Pick	Pick Two	Pick & Clean	Pick & Heat	Pick & Cool	Look	ScienceWorld
Qwen3-4B							
TEACHER	$87.5_{\pm 0.4}$	$66.2_{\pm 0.4}$	$35.5_{\pm 0.4}$	$57.6_{\pm 0.4}$	$46.4_{\pm 0.4}$	$83.3_{\pm 0.4}$	$26.1_{\pm 0.4}$
Qwen3-0.6B							
VANILLA	$4.2_{\pm 0.3}$	0.0	$1.1_{\pm 0.2}$	0.0	0.0	$3.5_{\pm 0.3}$	$0.8_{\pm 0.2}$
OPD	$65.2_{\pm 0.4}$	$18.5_{\pm 0.5}$	$14.8_{\pm 0.4}$	$4.5_{\pm 0.3}$	$27.4_{\pm 0.5}$	$6.8_{\pm 0.4}$	$7.3_{\pm 0.5}$
SAGE-OPD	$66.5_{\pm 0.5}$	$23.2_{\pm 0.4}$	$15.9_{\pm 0.3}$	$9.8_{\pm 0.4}$	$28.7_{\pm 0.4}$	$13.5_{\pm 0.5}$	$9.5_{\pm 0.4}$
RELAY-OPD	$67.8_{\pm 0.4}$	$28.0_{\pm 0.5}$	$17.2_{\pm 0.4}$	$17.5_{\pm 0.5}$	$30.5_{\pm 0.6}$	$21.2_{\pm 0.4}$	$10.2_{\pm 0.6}$
GUIDED-OPD	$69.1_{\pm 0.5}$	$33.6_{\pm 0.6}$	$18.3_{\pm 0.5}$	$24.6_{\pm 0.4}$	$31.8_{\pm 0.5}$	$30.8_{\pm 0.5}$	$12.6_{\pm 0.5}$
TCOD	$70.4_{\pm 0.4}$	$39.8_{\pm 0.5}$	$19.8_{\pm 0.4}$	$32.7_{\pm 0.5}$	$33.9_{\pm 0.4}$	$39.5_{\pm 0.6}$	$13.0_{\pm 0.4}$
FUTUREBRIDGE-OPD	$71.6_{\pm 0.5}$	$44.5_{\pm 0.5}$	$20.9_{\pm 0.4}$	$37.8_{\pm 0.6}$	$33.4_{\pm 0.5}$	$46.8_{\pm 0.5}$	$15.4_{\pm 0.5}$
STI-OPD	$\mathbf{73.5}_{\pm \mathbf{0.4}}$	$\mathbf{48.8}_{\pm \mathbf{0.6}}$	$\mathbf{22.4}_{\pm \mathbf{0.5}}$	$\mathbf{41.8}_{\pm \mathbf{0.5}}$	$\mathbf{36.2}_{\pm \mathbf{0.6}}$	$\mathbf{53.4}_{\pm \mathbf{0.6}}$	$\mathbf{18.7}_{\pm \mathbf{0.6}}$
Qwen3-1.7B							
VANILLA	$62.5_{\pm 0.3}$	$1.5_{\pm 0.2}$	$12.9_{\pm 0.3}$	$7.7_{\pm 0.2}$	$3.6_{\pm 0.2}$	$36.1_{\pm 0.4}$	$6.3_{\pm 0.3}$
OPD	$85.5_{\pm 0.5}$	$38.7_{\pm 0.4}$	$34.7_{\pm 0.5}$	$21.7_{\pm 0.4}$	$47.6_{\pm 0.5}$	$22.2_{\pm 0.4}$	$17.1_{\pm 0.5}$
SAGE-OPD	$86.8_{\pm 0.4}$	$43.5_{\pm 0.5}$	$35.8_{\pm 0.4}$	$27.4_{\pm 0.5}$	$48.9_{\pm 0.6}$	$29.8_{\pm 0.5}$	$21.4_{\pm 0.4}$
RELAY-OPD	$88.1_{\pm 0.5}$	$48.2_{\pm 0.6}$	$37.1_{\pm 0.5}$	$35.6_{\pm 0.4}$	$50.8_{\pm 0.5}$	$38.4_{\pm 0.6}$	$17.8_{\pm 0.5}$
GUIDED-OPD	$89.4_{\pm 0.4}$	$53.9_{\pm 0.5}$	$38.0_{\pm 0.4}$	$42.8_{\pm 0.5}$	$51.5_{\pm 0.4}$	$48.6_{\pm 0.5}$	$16.2_{\pm 0.6}$
TCOD	$90.6_{\pm 0.5}$	$60.1_{\pm 0.5}$	$39.5_{\pm 0.5}$	$51.3_{\pm 0.6}$	$53.7_{\pm 0.5}$	$58.2_{\pm 0.5}$	$18.7_{\pm 0.5}$
FUTUREBRIDGE-OPD	$91.8_{\pm 0.4}$	$64.7_{\pm 0.6}$	$40.2_{\pm 0.3}$	$57.5_{\pm 0.5}$	$53.2_{\pm 0.4}$	$66.5_{\pm 0.6}$	$22.2_{\pm 0.4}$
STI-OPD	$\mathbf{93.7}_{\pm \mathbf{0.5}}$	$\mathbf{69.1}_{\pm \mathbf{0.4}}$	$\mathbf{41.2}_{\pm \mathbf{0.3}}$	$\mathbf{62.0}_{\pm \mathbf{0.5}}$	$\mathbf{56.0}_{\pm \mathbf{0.4}}$	$\mathbf{73.7}_{\pm \mathbf{0.5}}$	$\mathbf{27.5}_{\pm \mathbf{0.5}}$

Table 3: Ablation results on the long-horizon agentic benchmarks with the Qwen3-1.7B student. Each variant modifies one component of STI-OPD: replacing k_3 with k_1, removing importance weighting, or replacing reverse KL with forward KL.

Method	ALFWorld Pick	Pick Two	Pick & Clean	Pick & Heat	Pick & Cool	Look	ScienceWorld
STI-OPD	$\mathbf{93.7}_{\pm \mathbf{0.5}}$	$\mathbf{69.1}_{\pm \mathbf{0.4}}$	$\mathbf{41.2}_{\pm \mathbf{0.3}}$	$\mathbf{62.0}_{\pm \mathbf{0.5}}$	$\mathbf{56.0}_{\pm \mathbf{0.4}}$	$\mathbf{73.7}_{\pm \mathbf{0.5}}$	$\mathbf{27.5}_{\pm \mathbf{0.5}}$
w/ k_1 estimator	$89.5_{\pm 0.5}$	$65.2_{\pm 0.4}$	$39.4_{\pm 0.5}$	$59.7_{\pm 0.5}$	$51.3_{\pm 0.4}$	$67.4_{\pm 0.5}$	$23.8_{\pm 0.5}$
w/o importance weighting	$86.2_{\pm 0.4}$	$66.7_{\pm 0.4}$	$36.8_{\pm 0.3}$	$56.2_{\pm 0.4}$	$46.4_{\pm 0.5}$	$61.9_{\pm 0.4}$	$21.5_{\pm 0.3}$
w/ forward KL	$91.3_{\pm 0.5}$	$67.2_{\pm 0.4}$	$38.2_{\pm 0.4}$	$58.8_{\pm 0.5}$	$54.9_{\pm 0.4}$	$69.6_{\pm 0.5}$	$23.1_{\pm 0.4}$

STI-OPD even outperforms its teacher, demonstrating the effectiveness of turn-adaptive intervention in mitigating student errors and enhancing long-horizon performance.

4.3 ABLATION STUDY

Table 3 isolates the contributions of the components in STI-OPD. Replacing k_3 with k_1 reduces performance in both environments, supporting the use of a non-negative discrepancy estimator with lower variance near policy agreement for intervention. We further compare the proposed stochastic intervention strategy against deterministic thresholding in § C.3. The stochastic strategy consistently performs better across ALFWorld and ScienceWorld without requiring a manually selected decision boundary. Removing importance weighting causes the largest degradation on each reported task, highlighting the importance of correcting the sampling mismatch introduced by teacher-generated tokens. Replacing reverse KL with forward KL also consistently reduces performance (we also study its impact on mitigating forgetting in § C.2), favoring preservation of the conditional reverse-KL objective through importance weighting over changing the distillation objective.

4.4 MECHANISM, SCALING, AND COST ANALYSIS OF STI-OPD

Training Dynamics. As shown in Figure 3, STI-OPD achieves higher success rates early in training and maintains its advantage over standard OPD. It also preserves higher policy entropy while requiring fewer interaction turns, suggesting greater policy diversity alongside more efficient task execution. Although the training losses become similar at later stages, the success-rate gap persists, indicating that comparable distillation losses do not necessarily translate into comparable agentic performance.

Mechanism Analysis. Figure 4 shows that STI-OPD shifts progressively from teacher-generated to student-generated tokens as training proceeds, and that the shift is more pronounced than for FUTUREBRIDGE-OPD. Alongside this shift, the teacher–student policy discrepancy decreases

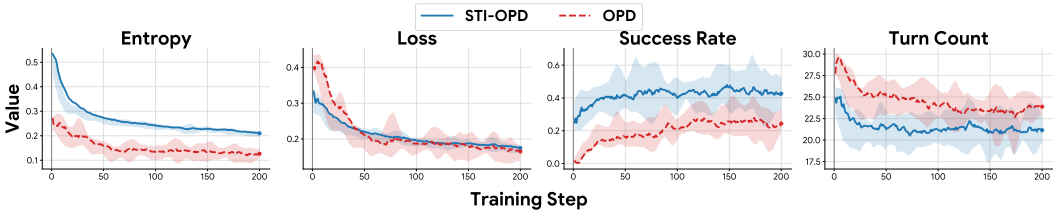

Figure 3: Training dynamics of STI-OPD and standard OPD on ALFWorld with Qwen3-1.7B as the student and Qwen3-4B as the teacher.

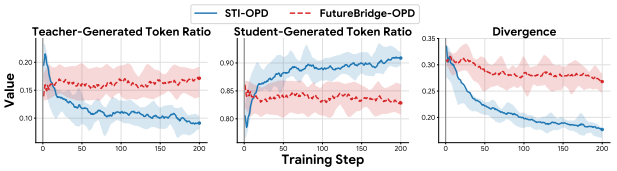

Figure 4: Intervention and policy discrepancy dynamics of STI-OPD and FUTUREBRIDGE-OPD on ALFWorld, using Qwen3-1.7B as the student and Qwen3-4B as the teacher.

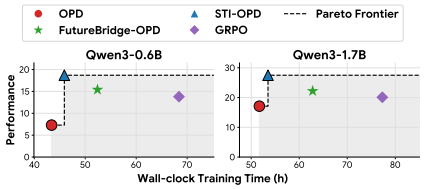

Figure 5: Performance versus training wall-clock time. The dashed line indicates the empirical Pareto frontier.

toward zero, as measured by the turn-level score in Equation (7). These coupled trends are what the intervention rule prescribes: the teacher takes over more often when policy discrepancy is large, then cedes control as the student aligns. Training therefore approaches the student-controlled regime used at inference time on its own, without a predefined intervention schedule. § C.5 shows cases and discussions about the mechanism of STI-OPD.

Scalability. Replacing the Qwen3-4B teacher with large-scale Qwen3-32B lifts each method accordingly. STI-OPD keeps the lead on each ALFWorld subtask and on ScienceWorld: with the 1.7B student it reaches 97.8 on LOOK and 56.3 on ScienceWorld, against 90.4 and 46.2 for FUTUREBRIDGE-OPD (see details in § C.1). The mechanism is therefore not specific to the Qwen3-4B teacher, and its margin does not close as the teacher grows.

Cost Analysis. Figure 5 shows that STI-OPD lies on the empirical Pareto frontier at both student sizes. STI-OPD improves substantially over standard OPD with modest additional training time. Compared with FUTUREBRIDGE-OPD, STI-OPD achieves higher performance at lower computational cost. To further rule out the possibility that these gains arise simply from greater teacher usage, we additionally compare intervention-based methods under the same teacher-intervention budget in § C.4. STI-OPD remains consistently better under this matched budget, indicating that its advantage comes from more effective allocation and utilization of teacher guidance rather than simply more frequent teacher intervention. Its efficiency relative to FUTUREBRIDGE-OPD is consistent with estimating discrepancy from the existing student proposal without additional lookahead rollouts. Moreover, the intervention frequency of STI-OPD decreases as the student policy aligns with the teacher, further reducing teacher dependence over training. Compared with GRPO, STI-OPD's advantage lies in its lower rollout computational cost, because it does not require group rollouts for reward normalization. Together, these results support a favorable performance–cost trade-off.

5 CONCLUSION

In this work, we studied on-policy distillation for multi-turn LLM agents, where an early student error changes the observations that follow and pushes the trajectory toward histories on which teacher supervision is less reliable. We introduced STI-OPD, which answers the two questions this setting raises: a per-sample non-negative k_3 estimate of teacher–student policy discrepancy sets a stochastic turn-level intervention probability, and an importance-sampling correction keeps supervision on teacher-generated responses faithful to the conditional reverse-KL objective. Across tool-integrated reasoning and long-horizon agentic benchmarks, STI-OPD outperforms standard OPD and every intervention-based baseline we evaluate, at both student scales and with its largest margins on the longest-horizon tasks, while adding only modest wall-clock overhead over standard OPD.

AI USE STATEMENT

We use generative AI tools to refine the writing and visualization of this paper for readability and use AI coding assistance for vibing code development. No other disclosure categories apply. We have reviewed all AI-assisted text and take full responsibility for the final content of this work, including all claims and artifacts.

ETHICS STATEMENT

We focus on general scientific research questions and pose no risks to safety, personal security, or privacy. No new datasets are released as part of this study. Additionally, the research does not encompass potentially harmful insights, methods, or applications, nor does it raise issues related to privacy, security, legal compliance, or research integrity. We foresee no ethical risks or conflicts of interest. We are committed to adhering to ethical guidelines throughout the research process.

REPRODUCIBILITY STATEMENT

The intervention rule and the corrected objective are stated in full in §§ 3.2 and 3.3, and Algorithm 1 in § A gives the complete training loop, including how the student and teacher buffers are populated and how student- and teacher-generated tokens enter the loss and how the teacher-branch importance correction is applied.

All training and evaluation data are based on publicly available datasets. § B.1 names each source, the subset used, and the per-source example counts, and states that the original responses are never used as fixed targets. § B.4 reproduces every prompt template verbatim, including the environment-specific system prompts and the observation-handling rules for ALFWorld and ScienceWorld, so the interaction protocol can be reconstructed exactly. Table 4 lists the student and teacher checkpoints, prompt counts, batch sizes, training length, sampling parameters, context and response limits, turn caps, learning rate, gradient clipping, and the hardware and software versions used. § B.2 describes each baseline and states which quantities are held fixed across methods. Evaluation runs the student policy alone, with no teacher intervention, in every reported setting.

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 *ICLR*. OpenReview.net, 2024. 1, 3.1, 4.1, B.2

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, pp. 1171–1179, 2015. 2

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. *CoRR*, abs/2608.01953, 2026. 1, 1, 2, 4.1, B.2

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *NeurIPS*, 2022. 4

DeepSeek-AI. Deepseek-v3.2: Pushing the frontier of open large language models. *CoRR*, abs/2512.02556, 2025. 1

Jasper Dekoninck, Nikola Jovanovic, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin T. Vechev. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms. *CoRR*, abs/2605.00674, 2026. 1, 4.1, B.1

Run-Ze Fan, Zengzhi Wang, and Pengfei Liu. Megascience: Pushing the frontiers of post-training datasets for science reasoning. *CoRR*, abs/2507.16812, 2025. 4.1, B.1

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *ICLR*. OpenReview.net, 2024. 1, 3.1, 4.1, B.2

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. Skywork open reasoner 1 technical report. *CoRR*, abs/2505.22312, 2025. 4.1, B.1

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *ICLR*. OpenReview.net, 2025. 1, 4.1, B.1

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *CoRR*, abs/2603.07079, 2026. 1

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *SOSP*, pp. 611–626. ACM, 2023. 4

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. *CoRR*, abs/2606.15912, 2026a. 1, 1, 2, 3.1, 3.2, 3.2, 4.1, B.2

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From live data to high-quality benchmarks: The arena-hard pipeline. https://lmsys.org/blog/2024-04-19-arena-hard/, 2024. C.2

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *CoRR*, abs/2604.13016, 2026b. 1, 3.1

Baohao Liao, Hanze Dong, Christof Monz, Xinxing Xu, Li Dong, and Furu Wei. Multi-turn on-policy distillation with prefix replay. *CoRR*, abs/2607.04763, 2026. 2

Junzhuo Liu, Weiwei Li, Jun Ling, and Peng Wang. When privileged guidance misaligns: State-matched routing and contextualized self-distillation for multi-turn agents. *CoRR*, abs/2608.05219, 2026. 2

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. https://thinkingmachines.ai/blog/on-policy-distillation. 1, 3.1

Jianlan Luo, Perry Dong, Yuexiang Zhai, Yi Ma, and Sergey Levine. RLIF: interactive imitation learning as reinforcement learning. In *ICLR*. OpenReview.net, 2024. 2

Xuchen Pan, Yanxi Chen, Yushuo Chen, Yuchang Sun, Daoyuan Chen, Wenhao Zhang, Yuexiang Xie, Yilun Huang, Yilei Zhang, Dawei Gao, Yaliang Li, Bolin Ding, and Jingren Zhou. Trinity-rft: A general-purpose and unified framework for reinforcement fine-tuning of large language models. *CoRR*, abs/2505.17826, 2025. 4

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8024–8035, 2019. 4

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022, 2023. 1, 4.1, B.1

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, volume 15 of *JMLR Proceedings*, pp. 627–635. JMLR.org, 2011. 2

John Schulman. Approximating KL divergence, 2020. URL http://joschu.net/blog/kl-approx.html. Blog post. 1, 3.2, 3.2

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. 1

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *ICLR*. OpenReview.net, 2021. 1, 1, 3.1, 4.1, B.1

Bing Wang, Shaotian Yan, Chen Shen, Kaiyuan Liu, Sinan Fan, Ximing Li, Rui Miao, Xiaosong Yuan, Zhanming Shen, and Jieping Ye. Backtracking when it strays: Mitigating dual exposure biases in LLM reasoning distillation. *CoRR*, abs/2605.19433, 2026a. 1, 2

Jiaqi Wang, Wenhao Zhang, Weijie Shi, Yaliang Li, and James Cheng. TCOD: exploring temporal curriculum in on-policy distillation for multi-turn autonomous agents. *CoRR*, abs/2604.24005, 2026b. 1, 2, 3.1, 3.2, 4.1, B.2

Ruoyao Wang, Peter A. Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? In *EMNLP*, pp. 11279–11298. Association for Computational Linguistics, 2022. 1, 3.1, 4.1, B.1

Yuchen Xia, Qianguo Sun, Chao Song, Junlong Wu, Yiyan Qi, and Yunjian Xu. DASH-OPD: discrepancy-aware switching with hysteresis for on-policy distillation. *CoRR*, abs/2607.29078, 2026. 1, 2

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. *CoRR*, abs/2607.26057, 2026. 1, 1, 2, 4.1, B.2

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *CoRR*, abs/2409.12122, 2024. 1

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, Jian Yang, Jiaxi Yang, Jingren 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. *CoRR*, abs/2505.09388, 2025. 1, 4.1, B.3

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR*. OpenReview.net, 2023. 3.1

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476, 2025. 4.1, B.1

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2024, 2024. 1, 4.1, B.1

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2025, 2025. 1, 4.1, B.1

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2026, 2026. 1, 4.1, B.1

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. *CoRR*, abs/2601.18734, 2026. 1

Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. SCOPE: signal-calibrated on-policy distillation enhancement with dual-path adaptive weighting. *CoRR*, abs/2604.10688, 2026. 1, 2

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. *CoRR*, abs/2605.07725, 2026. 1, 2

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911, 2023. C.2

Yuhang Zhou, Lizhu Zhang, Yifan Wu, Mingyi Wang, Bo Peng, Jiayi Liu, Xiangjun Fan, and Zhuokai Zhao. SAGE-OPD: selective agent-guided intervention for multi-turn on-policy distillation. *CoRR*, abs/2606.19659, 2026a. 1, 2, 4.1, B.2

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. *CoRR*, abs/2607.05804, 2026b. 2

Appendix

Contents of Appendix

A Algorithmic Details of STI-OPD **15**

B Additional Experimental Details **15**

B.1 Datasets 15

B.2 Baselines 17

B.3 Training Configuration and Parameters 18

B.4 Prompt Templates 19

C Additional Experimental Results **21**

C.1 STI-OPD Results with Larger Teacher 21

C.2 Preserving the Reverse-KL Objective Helps Mitigate Forgetting 21

C.3 Stochastic Intervention v.s. Deterministic Threshold 22

C.4 Teacher Computation Matched Comparison 23

C.5 Case Studies 23

D Limitations and Future Work **25**

A Algorithmic Details of STI-OPD

We present the complete algorithmic details of STI-OPD in Algorithm 1.

Algorithm 1 STI-OPD: Stochastic Teacher Intervention for Agentic On-Policy Distillation

Require: Task distribution $\mathcal{D}$; fixed teacher π_T; student π_θ; maximum horizon H
Ensure: Optimized student parameters θ

while not converged **do**
  Sample a task minibatch $\mathcal{Q} \sim \mathcal{D}$
  Initialize student and teacher buffers $\mathcal{B}_S, \mathcal{B}_T \leftarrow \emptyset$
  for all $q \in \mathcal{Q}$ **do**
    Sample $o_1 \sim \rho_0(\cdot \mid q)$ and set $h_1 \leftarrow (q, o_1)$
    for $t = 1, \ldots, H$ **do**
      Sample a student proposal $\widetilde{a}_t = (\widetilde{x}_{t,1}, \ldots, \widetilde{x}_{t,\widetilde{L}_t}) \sim \pi_\theta(\cdot \mid h_t)$
      for $k = 1, \ldots, \widetilde{L}_t$ **do**
        $\widetilde{c}_{t,k} \leftarrow (h_t, \widetilde{x}_{t,<k})$
        $r_{t,k} \leftarrow \dfrac{\pi_T(\widetilde{x}_{t,k} \mid \widetilde{c}_{t,k})}{\pi_\theta(\widetilde{x}_{t,k} \mid \widetilde{c}_{t,k})}$
        $g_{t,k} \leftarrow r_{t,k} - 1 - \log r_{t,k}$
      end for
      $s_t \leftarrow \widetilde{L}_t^{-1} \sum_{k=1}^{\widetilde{L}_t} g_{t,k}$
      $\alpha_t \leftarrow s_t/(1 + s_t)$
      Sample $z_t \sim \mathrm{Bernoulli}(\alpha_t)$
      if $z_t = 1$ **then**
        Sample a complete teacher turn $a_t^T = (x_{t,1}^T, \ldots, x_{t,L_t^T}^T) \sim \pi_T(\cdot \mid h_t)$
        for $k = 1, \ldots, L_t^T$ **do**
          $c_{t,k}^T \leftarrow (h_t, x_{t,<k}^T)$
          Add $(x_{t,k}^T, c_{t,k}^T)$ to $\mathcal{B}_T$
        end for
        Execute $a_t \leftarrow a_t^T$
      else
        for $k = 1, \ldots, \widetilde{L}_t$ **do**
          Add $(\widetilde{x}_{t,k}, \widetilde{c}_{t,k})$ to $\mathcal{B}_S$
        end for
        Execute $a_t \leftarrow \widetilde{a}_t$
      end if
      Observe $o_{t+1} \sim P(\cdot \mid h_t, a_t)$ and update h_{t+1}
      if the environment terminates **then**
        break
      end if
    end for
  end for
  Define $\ell_\theta(x, c) \leftarrow \log \pi_\theta(x \mid c) - \log \pi_T(x \mid c)$
  $\mathcal{L}_S \leftarrow \sum_{(x,c) \in \mathcal{B}_S} \ell_\theta(x, c)$
  $\mathcal{L}_T \leftarrow \sum_{(x,c) \in \mathcal{B}_T} \dfrac{\pi_\theta(x \mid c)}{\pi_T(x \mid c)} \ell_\theta(x, c)$
  $\widehat{\mathcal{L}}_{\text{STI-OPD}} \leftarrow \dfrac{\mathcal{L}_S + \mathcal{L}_T}{|\mathcal{B}_S| + |\mathcal{B}_T|}$
  Update θ using $\nabla_\theta \widehat{\mathcal{L}}_{\text{STI-OPD}}$
end while

B Additional Experimental Details

B.1 Datasets

Training Corpus. Our tool-integrated reasoning corpus combines mathematical, coding, and scientific problems from three public sources. We use the English subset of DAPO-Math, the code

subset of Skywork-OR1-RL-Data, and 10,000 examples from MegaScience. The resulting mixture contains 34,501 training prompts: 14,116 from DAPO-Math, 10,385 from Skywork-OR1-RL-Data, and 10,000 from MegaScience. The original responses associated with these datasets are not used as fixed distillation targets, instead, each prompt initiates a fresh rollout under the policy specified by the corresponding distillation method.

DAPO-Math. DAPO-Math-17K is a collection of 17,000 mathematical reasoning prompts curated from Web sources and official competition materials (Yu et al., 2025). The source dataset transforms each problem so that its target is an integer-valued answer, which enables reliable rule-based verification while retaining a broad range of competition-style reasoning patterns. We retain only its English-language examples for training.

Skywork-OR1-RL-Data. Skywork-OR1-RL-Data was developed for reinforcement-learning post-training of long-chain-of-thought reasoning models (He et al., 2025). Its released code portion contains 13,700 executable programming problems after quality filtering and deduplication, including 2,700 problems derived from LeetCode and 11,000 from TACO. We use this code subset to expose the student to algorithm design, program synthesis, and execution-based correctness signals.

MegaScience. MegaScience is a 1.25-million-example mixture of open scientific reasoning datasets spanning seven disciplines (Fan et al., 2025). It is designed to broaden post-training beyond mathematics and coding by providing diverse, verifiable scientific questions. We sample 10,000 examples from MegaScience and include them as the scientific component of our training mixture.

AIME 2024–2026. The American Invitational Mathematics Examination (AIME) consists of 15 short-answer problems whose answers are integers between 000 and 999. We use the 2024, 2025, and 2026 contests (Zhang & Math-AI, 2024; 2025; 2026). Their recency and emphasis on multi-step algebraic, geometric, combinatorial, and number-theoretic reasoning make them useful for evaluating mathematical generalization across successive competition years.

HMMT February 2026. The February 2026 Harvard–MIT Mathematics Tournament benchmark contains recent olympiad-style problems covering several areas of high-school mathematics. We include it as an additional out-of-corpus evaluation of challenging mathematical reasoning (Dekoninck et al., 2026).

GPQA. GPQA is a graduate-level multiple-choice benchmark comprising 448 questions written and validated by domain experts in biology, physics, and chemistry (Rein et al., 2023). The questions are designed to remain difficult even for skilled non-experts with unrestricted Web access. GPQA therefore evaluates whether the distilled student transfers scientific reasoning ability beyond the scientific prompts seen during post-training.

LiveCodeBench v6. LiveCodeBench is a continuously updated code benchmark constructed from recent problems collected from LeetCode, AtCoder, and Codeforces (Jain et al., 2025). It evaluates generated programs with executable test cases and reduces contamination risk by introducing newly released contest problems over time. We use version 6 for all code-generation evaluations.

ALFWorld. ALFWorld aligns a text-based simulator with embodied household tasks from the ALFRED benchmark (Shridhar et al., 2021). An agent receives a natural-language goal and must interact with the environment through a sequence of textual actions, such as navigation, object manipulation, cleaning, heating, cooling, and placement. Because every action changes the observations available at later turns, early mistakes can alter the remainder of the trajectory.

ScienceWorld. ScienceWorld is an interactive text environment grounded in an elementary science curriculum (Wang et al., 2022). Its tasks require agents to explore the environment, manipulate objects, conduct simple experiments, and use the resulting observations to achieve a scientific goal. The environment tests both procedural knowledge and long-horizon adaptation to feedback, making it a complementary setting to ALFWorld for studying inter-turn error accumulation.

Unless stated otherwise, all methods use identical training instances and evaluation tasks within each setting. This controls for variation in task composition and isolates the effect of the distillation procedure.

Table 4: Training and rollout parameters. Unless stated otherwise, these settings are shared by STI-OPD and all distillation baselines.

Category	Parameter	Value
Model	Student	Qwen3-0.6B; Qwen3-1.7B
	Teacher	Qwen3-4B; Qwen3-32B
Data	*Tool-Integrated Reasoning*	
	# Training prompts	34,501
	Training epochs	1
	Batch size	128
	ALFWorld	
	# Training prompts	3,553
	Training steps	200
	Batch size	32
	ScienceWorld	
	# Training prompts	2,294
	Training steps	200
	Batch size	32
Rollout	Rollouts per training prompt	1
	Sampling temperature	1.0
	Top-p	1.0
	Maximum response length per turn	32,768 tokens
	Maximum model context length	40,960 tokens
	Maximum Turns	
	Tool-integrated reasoning	5
	ALFWorld	30
	ScienceWorld	30
Optimization	Learning rate	1×10^{-6}
	Gradient clipping	1.0
Evaluation	Policy used at inference	Student only; no teacher intervention
	Reported aggregation	Mean across different runs
Hardware	GPU	4 NVIDIA GB200 GPUs
	CPU	Neoverse-v2
Software	CUDA	13.0
	Pytorch (Paszke et al., 2019)	2.11.0
	vLLM (Kwon et al., 2023)	0.23.0
	Flash-Attention (Dao et al., 2022)	0.8.3
	Trinity-RFT (Pan et al., 2025)	0.6.0

B.2 Baselines

Vanilla. Vanilla is the original student checkpoint without additional post-training. It measures the capability supplied by the pretrained model and provides a reference for quantifying the net effect of distillation.

OPD. OPD denotes standard on-policy distillation (Agarwal et al., 2024; Gu et al., 2024). It collects rollouts from the student and minimizes the token-level reverse KL divergence to a fixed teacher

on the student-visited contexts. It therefore reduces the sequence-level mismatch between offline training and student inference, but does not explicitly control trajectory drift in multi-turn interaction.

SAGE-OPD. SAGE-OPD selectively allocates teacher supervision at the turn level (Zhou et al., 2026a). After observing environment feedback, it uses teacher judgment to decide whether a student turn should be skipped or supervised. It further weights token-level distillation by teacher confidence and normalizes the loss to preserve the scale of the standard OPD objective.

GUIDED-OPD. GUIDED-OPD mixes teacher- and student-generated turns within each trajectory (Li et al., 2026a). A curriculum begins with frequent teacher intervention to keep early rollouts close to the teacher-induced history distribution and gradually decays the intervention probability to zero, recovering student-only rollouts toward the end of training.

RELAY-OPD. RELAY-OPD targets prefix failures by detecting locations where teacher and student continuations exhibit different tendencies (Xu et al., 2026). At a detected trigger, the teacher temporarily takes over to generate a short teacher relay segment, after which control returns to the student. A limited relay budget concentrates guidance at selected positions while restricting the deviation from student-generated trajectories.

TCOD. TCOD introduces a temporal curriculum over interaction depth (Wang et al., 2026b). It initially exposes the student to short trajectory prefixes and progressively increases the number of student-controlled turns. This curriculum is designed to prevent early inter-turn errors from immediately pushing training trajectories far from the teacher-induced distribution.

FUTUREBRIDGE-OPD. FUTUREBRIDGE-OPD evaluates teacher guidance through its downstream effect on the student trajectory (Chen et al., 2026). At a high-discrepancy state, it inserts a short teacher-generated bridge and then examines the subsequent student continuation to determine whether the bridge increases the density of useful distillation signals. It thereby selects interventions using future trajectory validation rather than only the current turn's discrepancy.

All methods use the same student initialization, teacher checkpoint, training prompts, rollout budget, and evaluation protocol for a given comparison. We also apply the same context and response limits across methods. This setup ensures that performance differences are attributable to the distillation and intervention strategies rather than to differences in models or data budgets.

B.3 TRAINING CONFIGURATION AND PARAMETERS

We use Qwen3-0.6B and Qwen3-1.7B as students and Qwen3-4B as the default teacher (Yang et al., 2025). We additionally use Qwen3-32B as a stronger teacher to study whether STI-OPD scales with teacher capability. Shared training and rollout parameters are summarized in Table 4.

For every training prompt, a distillation method collects one rollout using temperature 1.0 and top-p 1.0. At each turn, the model generates at most 32,768 response tokens and maximum model context length is 40,960 tokens. We cap tool-integrated reasoning trajectories at five turns and trajectories in ALFWorld and ScienceWorld at 30 turns. All methods are trained for one epoch for TIR and 200 steps for ALFWorld and ScienceWorld with a learning rate of 1×10^{-6}.

During STI-OPD training, the teacher scores each student-proposed turn, the intervention decision is sampled according to Equation (8), and the student is optimized on teacher-generated tokens using the importance-weighted objective in Equation (11). The teacher is used only during training and evaluation runs the student policy alone. We report the mean over multiple independent runs, with standard deviations taken across those runs, and reproduce all task prompts verbatim in § B.4.

In terms of engineering implementation, we optimize STI-OPD based on a policy gradient-style workflow. Specifically, we treat $-\ell_\theta$ in Equation (11) as the reward and then compute the PPO-style loss. This allows us to leverage well-established techniques such as clipping to achieve more stable optimization. For instance, $\omega_{t,k}^{T}$ in Equation (11) will be clipped to enable stable and controllable step-wise optimization.

B.4 Prompt Templates

For reproducibility, we provide the complete task prompts used in our experiments. Mathematical and scientific reasoning examples share the same `MATH_PROMPT`. Only the task-specific `{question}` field is substituted. Code-generation examples use a separate template that enforces a Python-only output format. ALFWorld and ScienceWorld use environment-specific system prompts that define the interaction protocol and valid action space.

For interactive tasks, each environment observation is supplied in a user message prefixed by `Observation:`. In ALFWorld, an observation containing `Nothing happens.` is additionally followed by a reminder to check action validity and output formatting. In ScienceWorld, the initial observation is also prefixed by the task description. The prompt text below is otherwise reproduced verbatim from the training workflow.

Prompt B.1: Mathematical and Scientific Reasoning

```
Analyze and solve the following mathematics problem step by step.

Question:
{question}

You may use available tools to perform precise calculations,
validate intermediate results, and verify your final answer when
helpful.

Instructions:
1.  First, analyze the problem and formulate a high-level solution
strategy.
2.  Carry out the solution step by step, using tools only when they
improve accuracy or efficiency.
3.  Before finalizing, verify that the result satisfies all
conditions in the problem.
4.  Once you have determined the final answer, do not call any
additional tools.

Your final response must contain the answer in exactly this format:
\boxed{{The final answer goes here.}}
```

Prompt B.2: Code Generation

```
You will be given a programming problem specification.  Write a
correct Python program that solves the problem and passes all test
cases.

Question:
{question}

Instructions:
1.  Analyze the problem and formulate a high-level solution
strategy.
2.  Derive the algorithm, considering correctness, edge cases, and
complexity.
3.  You may use available tools to validate reasoning or test edge
cases when helpful.
4.  Do not execute the program directly on the provided sample
inputs.
5.  Once you are confident that the solution is correct, do not call
additional tools.

The program must read from stdin, write only the required output to
stdout, be self-contained, and be compatible with Python.

Return only the final code in this format:
```python
```
```

```
# YOUR CODE HERE
```

Prompt B.3: ALFWorld System

```
You are an agent interacting with a virtual text-based environment.

## Notes:
At each step, you should first think then perform action to fulfill
the instruction.  You should ALWAYS wrap your thinking with the
<think> </think> tag and wrap your action with the <action> </action>
tag.
You should ALWAYS take one action each step.
DO NOT try to interact with the user at anytime.  Finish the task by
yourself.

## Action Format:
Below are the available commands you can use:
 look:  look around your current location
 inventory:  check your current inventory(you can only have 1 item
in your inventory)
 go to (receptacle):  move to a receptacle
 open (receptacle):  open a receptacle
 close (receptacle):  close a receptacle
 take (object) from (receptacle):  take an object from a receptacle
 move (object) to (receptacle):  place an object in or on a
receptacle
 examine (something):  examine a receptacle or an object
 use (object):  use an object
 heat (object) with (receptacle):  heat an object using a receptacle
 clean (object) with (receptacle):  clean an object using a
receptacle
 cool (object) with (receptacle):  cool an object using a receptacle
 slice (object) with (object):  slice an object using a sharp object

For example your output should be like this:
<think> To solve the task, I need first to ...  </think><action>go to
cabinet 1</action>
```

Prompt B.4: ScienceWorld System

```
You are an agent, your job is to do some scientific experiment in a
virtual text-based environment.

## Notes:
At each step, you should first think then perform action to fulfill
the instruction.  You should ALWAYS wrap your thinking with the
<think> </think> tag and wrap your action with the <action> </action>
tag.
You should ALWAYS take one action each step.
DO NOT try to interact with the user at anytime.  Finish the task by
yourself.

## Action Format:
Below are the available commands you can use:
 open OBJ: open a container
 close OBJ: close a container
 activate OBJ: activate a device
 deactivate OBJ: deactivate a device
 connect OBJ to OBJ: connect electrical components
 disconnect OBJ: disconnect electrical components
 use OBJ [on OBJ]: use a device/item
```

```
look around:  describe the current room
examine OBJ: describe an object in detail
look at OBJ: describe a container's contents
read OBJ: read a note or book
move OBJ to OBJ: move an object to a container
pick up OBJ: move an object to the inventory
pour OBJ into OBJ: pour a liquid into a container
mix OBJ: chemically mix a container
teleport to LOC: teleport to a specific room
focus on OBJ: signal intent on a task object
wait:  task no action for 10 steps
wait1:  task no action for a step

For example your output should be like this:
<think> Now I will check the bedroom ...  </think><action>teleport to
bedroom</action>
```

C Additional Experimental Results

C.1 STI-OPD Results with Larger Teacher

To demonstrate the effectiveness of STI-OPD with a larger teacher model, we conduct experiments on ALFWorld and ScienceWorld using Qwen3-32B as the teacher. We focus on these benchmarks due to computational constraints and our primary interest in long-horizon agentic tasks.

STI-OPD Remains Effective With A Larger Teacher. As shown in Table 5, STI-OPD achieves the highest success rate among the distillation methods in every reported ALFWorld category and on ScienceWorld, at both student sizes, when Qwen3-32B is used as the teacher. Its advantage holds not only over standard OPD but also over the stronger intervention-based baselines, indicating that the benefit of discrepancy-adaptive intervention is preserved as teacher capacity increases. The framework is therefore not specific to the Qwen3-4B teacher used in the main experiments.

The Largest Gains Occur on Tasks with Longer Dependencies. Relative to the best competing method, the margins are most pronounced on ScienceWorld and Look ($+10.1$ and $+7.4$ points for the 1.7B student), while remaining positive on every task category. These settings require information and action outcomes to be propagated across multiple turns, making them especially sensitive to early trajectory deviations. This pattern is consistent with the intended role of STI-OPD: stronger teacher supervision is most useful when combined with adaptive intervention at histories where student errors would otherwise affect subsequent decisions.

C.2 Preserving the Reverse-KL Objective Helps Mitigate Forgetting

Some distillation methods learn from teacher-generated tokens using a forward-KL objective, which may compromise previously acquired capabilities. Beyond improving task performance (§ 4.3), we investigate whether preserving the reverse-KL objective helps mitigate forgetting during distillation with teacher intervention.

We evaluate capability retention on IFEval (Zhou et al., 2023) and Arena-Hard (Li et al., 2024), which are benchmarks designed to assess the subjective ability of LLMs. We compare STI-OPD with a variant that replaces reverse KL with forward KL, using the original student checkpoint (Vanilla) as the reference for measuring changes in performance.

Table 6: Impact of preserving the reverse-KL objective on mitigating forgetting.

Method	IFEval	Arena-Hard
Qwen3-0.6B		
Vanilla	60.1	8.6
STI-OPD $w/$ forward KL	55.8↓4.3	7.2↓1.4
STI-OPD	**64.3**↑4.2	**9.1**↑0.5
Qwen3-1.7B		
Vanilla	71.3	41.3
STI-OPD $w/$ forward KL	65.7↓5.6	36.4↓4.9
STI-OPD	**74.3**↑3.0	**43.6**↑2.3

Table 5: Success rates on long-horizon agentic benchmarks with Qwen3-32B as teacher. We report results for six ALFWorld task categories and the average across ScienceWorld tasks. Bold and underlined values denote the best and second-best results, respectively.

Method	ALFWorld						ScienceWorld
	Pick	Pick Two	Pick & Clean	Pick & Heat	Pick & Cool	Look	
Qwen3-32B							
TEACHER	$91.7_{\pm 0.4}$	$72.1_{\pm 0.3}$	$62.1_{\pm 0.3}$	$80.4_{\pm 0.5}$	$65.5_{\pm 0.4}$	$93.1_{\pm 0.3}$	$47.8_{\pm 0.4}$
Qwen3-0.6B							
VANILLA	$4.2_{\pm 0.3}$	$0.0_{\pm 0.2}$	$1.1_{\pm 0.2}$	$0.0_{\pm 0.2}$	$0.0_{\pm 0.2}$	$3.5_{\pm 0.3}$	$0.8_{\pm 0.2}$
OPD	$66.9_{\pm 0.4}$	$22.4_{\pm 0.5}$	$34.7_{\pm 0.4}$	$13.3_{\pm 0.3}$	$42.2_{\pm 0.5}$	$16.0_{\pm 0.4}$	$22.6_{\pm 0.5}$
SAGE-OPD	$68.3_{\pm 0.5}$	$27.5_{\pm 0.4}$	$37.9_{\pm 0.3}$	$21.5_{\pm 0.4}$	$44.6_{\pm 0.4}$	$25.7_{\pm 0.5}$	$25.0_{\pm 0.4}$
RELAY-OPD	$69.8_{\pm 0.4}$	$32.9_{\pm 0.5}$	$41.0_{\pm 0.4}$	$31.3_{\pm 0.5}$	$47.2_{\pm 0.6}$	$35.6_{\pm 0.4}$	$25.9_{\pm 0.6}$
GUIDED-OPD	$71.2_{\pm 0.5}$	$38.9_{\pm 0.6}$	$44.0_{\pm 0.5}$	$40.0_{\pm 0.4}$	$49.9_{\pm 0.5}$	$47.5_{\pm 0.5}$	$29.2_{\pm 0.5}$
TCOD	$72.6_{\pm 0.4}$	$45.7_{\pm 0.5}$	$47.1_{\pm 0.4}$	$49.8_{\pm 0.5}$	$\underline{53.1}_{\pm 0.4}$	$57.5_{\pm 0.6}$	$29.9_{\pm 0.4}$
FUTUREBRIDGE-OPD	$\underline{73.8}_{\pm 0.5}$	$\underline{51.0}_{\pm 0.5}$	$\underline{49.2}_{\pm 0.4}$	$\underline{56.5}_{\pm 0.6}$	$52.5_{\pm 0.5}$	$\underline{65.7}_{\pm 0.5}$	$\underline{32.4}_{\pm 0.5}$
STI-OPD	$\mathbf{75.6}_{\pm 0.4}$	$\mathbf{55.6}_{\pm 0.6}$	$\mathbf{52.2}_{\pm 0.5}$	$\mathbf{61.3}_{\pm 0.5}$	$\mathbf{56.4}_{\pm 0.6}$	$\mathbf{72.8}_{\pm 0.6}$	$\mathbf{40.5}_{\pm 0.6}$
Qwen3-1.7B							
VANILLA	$62.5_{\pm 0.3}$	$1.5_{\pm 0.2}$	$12.9_{\pm 0.3}$	$7.7_{\pm 0.2}$	$3.6_{\pm 0.2}$	$36.1_{\pm 0.4}$	$6.3_{\pm 0.3}$
OPD	$87.2_{\pm 0.4}$	$42.6_{\pm 0.5}$	$54.6_{\pm 0.4}$	$30.5_{\pm 0.5}$	$62.4_{\pm 0.5}$	$31.4_{\pm 0.4}$	$32.4_{\pm 0.5}$
SAGE-OPD	$88.6_{\pm 0.5}$	$47.8_{\pm 0.4}$	$57.8_{\pm 0.5}$	$39.1_{\pm 0.4}$	$64.8_{\pm 0.6}$	$42.0_{\pm 0.5}$	$36.9_{\pm 0.4}$
RELAY-OPD	$90.1_{\pm 0.4}$	$53.1_{\pm 0.6}$	$60.9_{\pm 0.3}$	$49.4_{\pm 0.5}$	$67.5_{\pm 0.4}$	$52.8_{\pm 0.6}$	$33.5_{\pm 0.5}$
GUIDED-OPD	$91.5_{\pm 0.5}$	$59.2_{\pm 0.5}$	$63.7_{\pm 0.4}$	$58.2_{\pm 0.4}$	$69.6_{\pm 0.5}$	$65.3_{\pm 0.4}$	$42.8_{\pm 0.6}$
TCOD	$92.8_{\pm 0.4}$	$66.0_{\pm 0.5}$	$66.8_{\pm 0.5}$	$68.4_{\pm 0.6}$	$\underline{72.9}_{\pm 0.5}$	$76.2_{\pm 0.5}$	$45.6_{\pm 0.5}$
FUTUREBRIDGE-OPD	$\underline{94.0}_{\pm 0.5}$	$\underline{71.2}_{\pm 0.6}$	$\underline{68.5}_{\pm 0.4}$	$\underline{76.2}_{\pm 0.5}$	$72.3_{\pm 0.4}$	$\underline{90.4}_{\pm 0.6}$	$\underline{46.2}_{\pm 0.4}$
STI-OPD	$\mathbf{95.8}_{\pm 0.5}$	$\mathbf{75.9}_{\pm 0.5}$	$\mathbf{71.0}_{\pm 0.3}$	$\mathbf{81.5}_{\pm 0.6}$	$\mathbf{76.2}_{\pm 0.6}$	$\mathbf{97.8}_{\pm 0.5}$	$\mathbf{56.3}_{\pm 0.6}$

As shown in Table 6, the forward-KL variant underperforms the original checkpoint on both benchmarks at both student sizes, indicating degradation in previously acquired capabilities. In contrast, STI-OPD improves over the original checkpoint across all four comparisons. These results suggest that preserving the reverse-KL objective helps retain general capabilities while acquiring task-specific behaviors, providing an additional benefit beyond the agentic performance gains.

C.3 STOCHASTIC INTERVENTION V.S. DETERMINISTIC THRESHOLD

Setpup. To isolate the effect of stochastic intervention, we replace the proposed stochastic decision rule with deterministic thresholding while keeping all other training and evaluation settings unchanged. Specifically, given the same turn-level discrepancy score s_t, the deterministic variant triggers teacher intervention whenever $s_t > \tau$, i.e., $z_t = \mathbf{1}[s_t > \tau]$. We consider three thresholds, $\tau \in \{0.5, 1.0, 2.0\}$, covering increasingly conservative intervention policies. All variants use the same student and teacher models, training data, optimization hyperparameters, rollout budget, and evaluation protocol as STI-OPD, so that the comparison isolates the effect of converting the discrepancy score into a stochastic intervention probability rather than a hard decision boundary. We report results on the two long-horizon interactive benchmarks, ALFWorld and ScienceWorld, where compounding inter-turn errors make intervention decisions particularly consequential.

Table 7: Comparison of stochastic intervention and deterministic threshold on ALFWorld and ScienceWorld benchmarks.

Method	ALFWorld						ScienceWorld
	Pick	Pick Two	Pick & Clean	Pick & Heat	Pick & Cool	Look	
Thre. 0.5	$88.5_{\pm 0.5}$	$62.7_{\pm 0.4}$	$36.5_{\pm 0.4}$	$55.1_{\pm 0.5}$	$49.7_{\pm 0.4}$	$65.3_{\pm 0.5}$	$20.8_{\pm 0.4}$
Thre. 1.0	$89.8_{\pm 0.4}$	$64.8_{\pm 0.4}$	$38.0_{\pm 0.3}$	$57.2_{\pm 0.5}$	$51.8_{\pm 0.4}$	$68.4_{\pm 0.4}$	$23.2_{\pm 0.5}$
Thre. 2.0	$89.0_{\pm 0.4}$	$63.3_{\pm 0.5}$	$35.8_{\pm 0.5}$	$54.5_{\pm 0.6}$	$49.2_{\pm 0.5}$	$64.7_{\pm 0.6}$	$20.3_{\pm 0.5}$
STI-OPD	$\mathbf{93.7}_{\pm 0.5}$	$\mathbf{69.1}_{\pm 0.5}$	$\mathbf{41.2}_{\pm 0.2}$	$\mathbf{62.0}_{\pm 0.6}$	$\mathbf{56.0}_{\pm 0.6}$	$\mathbf{73.7}_{\pm 0.5}$	$\mathbf{27.5}_{\pm 0.6}$

Results and Analysis. Table 7 compares our stochastic intervention strategy with deterministic intervention using several fixed discrepancy thresholds. Among the deterministic variants, a threshold of 1.0 performs best across all tasks, while both smaller and larger thresholds lead to consistent degradation. This non-monotonic behavior indicates that deterministic intervention is sensitive to the choice of decision boundary: a small threshold may trigger teacher intervention too frequently,

whereas a large threshold may fail to intervene when corrective guidance is beneficial. In contrast, STI-OPD consistently outperforms all deterministic variants across every ALFWorld category and ScienceWorld. These results indicate that the performance of deterministic intervention depends noticeably on the selected threshold, whereas the stochastic strategy avoids introducing a manually tuned decision boundary and performs consistently better across the evaluated tasks. By mapping discrepancy continuously to an intervention probability, STI-OPD allows teacher control to increase gradually with teacher–student disagreement while preserving opportunities for student execution. This provides empirical support for using discrepancy-conditioned stochastic intervention rather than a fixed threshold.

C.4 Teacher Computation Matched Comparison

Setup. To examine whether the gains of STI-OPD arise simply from increased teacher usage, we compare all intervention-based methods under a matched teacher-intervention budget. Specifically, we impose the same maximum number of teacher interventions for each method and terminate training once this budget is exhausted. For methods with configurable intervention schedules or frequencies, we adjust these controls so that all methods consume the same total number of teacher interventions over training. All other training and evaluation settings, including the student and teacher models, training data, optimization hyperparameters, rollout configuration, and evaluation protocol, are kept identical to the corresponding main experiments. This comparison therefore controls for the amount of teacher intervention and focuses on how effectively each method allocates and learns from a fixed teacher budget.

Table 8: Results of STI-OPD and baselines under the same teacher computation budget.

Method	ALFWorld						ScienceWorld
	Pick	Pick Two	Pick & Clean	Pick & Heat	Pick & Cool	Look	
RELAY-OPD	$84.9_{\pm 0.5}$	$44.8_{\pm 0.5}$	$33.5_{\pm 0.4}$	$32.4_{\pm 0.4}$	$47.3_{\pm 0.5}$	$35.1_{\pm 0.5}$	$14.6_{\pm 0.4}$
GUIDED-OPD	$86.2_{\pm 0.4}$	$50.3_{\pm 0.5}$	$34.8_{\pm 0.4}$	$39.5_{\pm 0.4}$	$48.2_{\pm 0.4}$	$45.4_{\pm 0.5}$	$13.1_{\pm 0.5}$
FUTUREBRIDGE-OPD	$\underline{88.3}_{\pm 0.4}$	$\underline{61.0}_{\pm 0.5}$	$\underline{36.7}_{\pm 0.3}$	$\underline{53.8}_{\pm 0.4}$	$49.7_{\pm 0.4}$	$\underline{62.9}_{\pm 0.5}$	$\underline{18.7}_{\pm 0.4}$
STI-OPD	$\mathbf{90.8}_{\pm 0.4}$	$\mathbf{65.8}_{\pm 0.4}$	$\mathbf{38.1}_{\pm 0.2}$	$\mathbf{58.6}_{\pm 0.5}$	$\mathbf{52.8}_{\pm 0.4}$	$\mathbf{69.9}_{\pm 0.4}$	$\mathbf{24.2}_{\pm 0.4}$

Results and Analysis. As shown in Table 8, STI-OPD consistently outperforms all competing intervention-based OPD methods under the same teacher-intervention budget. Compared with the strongest baseline, FUTUREBRIDGE-OPD, STI-OPD improves performance by 1.4–7.0 points across all seven reported tasks, with an average gain of approximately 4.2 points. The largest improvements are observed on LOOK (+7.0), ScienceWorld (+5.5), and PICK TWO/PICK & HEAT (+4.8 each), where successful completion requires sustained decision making over multiple interaction steps. These results indicate that the advantage of STI-OPD cannot be explained solely by using more teacher interventions. Instead, under a fixed teacher budget, its discrepancy-conditioned stochastic intervention allocates teacher guidance more effectively across interaction states, while the importance-weighted objective enables the student to learn from the resulting teacher-generated tokens without changing the conditional reverse-KL target. Overall, the comparison suggests that the gains of STI-OPD arise from more effective use of teacher supervision rather than simply greater teacher involvement.

C.5 Case Studies

We pair early- and late-training trajectories from the same ALFWorld dataset instances. The controller column identifies the policy that generated the executed action, and s_t is the turn-level discrepancy score of Equation (7). Scores on teacher-controlled turns are computed from the discarded student proposals. We abbreviate repeated search steps while preserving every action needed for the conclusions below.

Case C.1: Local Transfer — The Microwave Subtask Becomes Student-Controlled

Task: Cool an apple and place it in a microwave. **Dataset instance:** 15
Early trajectory: batch 1; 5 consecutive ineffective actions in the shown segment. **Later trajectory:** batch 112; 0 ineffective actions.
Outcome: Both trajectories succeed, but control of the previously problematic microwave subtask transfers from teacher to student.

Turn	Executed action and environment response (abbreviated)	Controller	s_t
Early training — the student reaches an invalid local state and requires correction			
15	`cool apple 1 with fridge 1` → apple cooled	Student	0.993
16–20	Attempts to move/open the microwave → `Nothing happens.` five times	Student	0.478–0.835
21	`go to microwave 1` → arrives at closed microwave	**Teacher**	0.845
22	`open microwave 1` → microwave opened	**Teacher**	0.846
23	`move apple 1 to microwave 1` → reward 1	Student	0.506
Later training — the same local subtask is completed without intervention			
23	`cool apple 1 with fridge 1` → apple cooled	**Student**	0.286
24	`go to microwave 1` → arrives at closed microwave	**Student**	0.231
25	`open microwave 1` → microwave opened	**Student**	0.287
26	`move apple 1 to microwave 1` → reward 1	**Student**	0.462

Analysis

Early in training, the student cools the apple but attempts to manipulate the microwave before navigating to it. The teacher restores a valid state by navigating to and opening the microwave. When the same instance reappears, the student independently executes this corrected sequence, and the navigation and opening scores decrease from $0.845/0.846$ to $0.231/0.287$. The case therefore provides evidence of *local* skill acquisition: intervention is converted into student-controlled behavior on the corrected subtask. Because two teacher actions still occur earlier in the later trajectory, we do not interpret this example as end-to-end teacher independence.

Case C.2: End-to-End Transfer — Search and Placement without Intervention

Task: Find two pens and place both in the desk. **Dataset instance:** 713
Early trajectory: batch 23; 24 turns, 9 teacher interventions, and 4 ineffective actions.
Later trajectory: batch 134; 21 turns, 0 teacher interventions, and 0 ineffective actions.
Outcome: Both trajectories receive reward 1 but the later trajectory is generated entirely by the student.

Phase	Executed action sequence (abbreviated)	Controller	s_t
Early training — intervention spans exploration, object acquisition, and return navigation			
Initialize search	`go drawer 1; open drawer 1; go drawer 2`	**Teacher**	0.292, 0.414, 0.698
Locate first pen	Search shelves; `go shelf 5; take pen 1 from shelf 5`	Mixed	0.430 at intervention
Locate second pen	`go shelf 6; take pen 2 from shelf 6`	**Teacher**	0.454, 0.614
Return and place	Teacher navigates to the desk; student eventually places both pens → reward 1	Mixed	0.554, 0.288 at interventions
Later training — all search and placement decisions are student-generated			
Initialize search	`go drawer 1; open drawer 1; go drawer 2; open drawer 2`	**Student**	0.322–0.516
Locate first pen	Search shelves; `go shelf 6; take pen 2 from shelf 6`	**Student**	0.406–0.472
Place first pen	`go desk 1; move pen 2 to desk 1`	**Student**	0.161–0.357
Locate and place second pen	`go shelf 5; take pen 1; go desk 1; move pen 1 to desk 1` → reward 1	**Student**	0.185–0.413

Analysis

This instance requires a longer dependency chain than Case A: the agent must search multiple receptacles, remember two object locations, transport each object, and revisit the target desk. The early trajectory invokes the teacher at nine decisions distributed across these phases. In the later trajectory, the student reproduces the useful search and transport actions—including navigating to shelves 5 and 6, collecting both pens, and returning to the desk—without any teacher-generated turn. The simultaneous reductions in teacher interventions ($9 \rightarrow 0$), ineffective actions ($4 \rightarrow 0$), and trajectory length ($24 \rightarrow 21$) indicate that the improvement is not explained by longer exploration. Rather, the student executes the same multi-stage task autonomously and more efficiently.

Case C.3: Compositional Transfer — The Full Skill Chain Becomes Student-Controlled

Task: Place a cooled bowl on the dining table. **Dataset instance:** 176
Early trajectory: batch 6; 19 turns, 7 teacher interventions, and 2 ineffective actions.
Later trajectory: batch 117; 8 turns, 0 teacher interventions, and 0 ineffective actions.
Outcome: Both trajectories receive reward 1 but the later student executes the complete manipulation chain directly.

Phase	Executed action sequence (abbreviated)	Controller	s_t
Early training — the student detours and the teacher supplies the critical manipulation chain			
Acquire object	Student opens cabinet; teacher executes `take bowl 1` and `go fridge 1`	Mixed	0.834, 0.754 at interventions
Detour	Student returns to the table, moves the bowl back, and repeatedly examines the scene	Student	0.427–1.076
Recover object	`take bowl 1 from diningtable 1`	**Teacher**	0.748
Cool object	`open fridge 1; cool bowl 1 with fridge 1`	**Teacher**	0.417, 0.490
Complete task	Teacher navigates to the table and finally executes `move bowl 1 to diningtable 1` → reward 1	Mixed	0.601, 0.603 at interventions
Later training — the student executes a direct valid chain without intervention			
Acquire object	`go cabinet 1; open cabinet 1; take bowl 1`	**Student**	0.290–0.323
Cool object	`go fridge 1; open fridge 1; cool bowl 1 with fridge 1`	**Student**	0.308–0.491
Complete task	`go diningtable 1; move bowl 1 to diningtable 1` → reward 1	**Student**	0.342–0.533

Analysis

The early trajectory exposes a compositional failure rather than a single invalid action: after acquiring the bowl, the student returns it to the dining table before cooling it and then stalls on repeated examination. The teacher must re-acquire the object and supply the state-changing actions that open the fridge, cool the bowl, and complete the placement. On the later occurrence, the student executes precisely the required dependency order—acquire, navigate, open, cool, return, and place—in eight valid actions. Teacher use falls from seven turns to zero while trajectory length falls by more than half. Together with Case C.2, this demonstrates full-trajectory student control in two distinct task families, complementing the local transfer illustrated by Case C.1.

D Limitations and Future Work

While STI-OPD offers practical benefits, it also comes with a few limitations that point to interesting future work. All experiments use the text-only LLMs, two student scales, two teacher scales, and two families of text-based interactive tasks. We do not test multimodal or embodied settings, or environments with irreversible external side effects. Whether discrepancy-calibrated intervention transfers to those settings is open. The routing score in Equation (5) needs the teacher's per-token

probabilities on student-proposed tokens, and the correction in Equation (10) needs them again on teacher-generated tokens. STI-OPD is therefore inapplicable to teachers exposed only through a sampling API, which excludes many of the strongest available models. When $z_t = 1$ the teacher regenerates the entire turn. This preserves coherence between the reasoning, the action, and the environment transition, but it spends teacher supervision on the whole response even when discrepancy is concentrated in a single decision, and it offers no mechanism for correcting part of a turn.