DTDD: DIVERGENCE-TRIGGERED DYNAMIC DISTILLATION FOR RELIABLE ON-POLICY SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) provides dense token-level supervision on student rollouts, but this supervision can become unreliable when the student and teacher diverge, either because of a large initial policy gap or compounding errors in long-horizon reasoning and tool use. We introduce Divergence-Triggered Dynamic Distillation (DTDD), an adaptive OPD framework that routes supervision between standard OPD and teacher-generated recovery using segment-level student–teacher divergence. When divergence triggers an intervention during reasoning, the student prefix continues to receive standard OPD supervision. The teacher then generates a recovery suffix, which the student learns through behavior cloning, an online SFT-style loss. We evaluate DTDD in three regimes: long-horizon reasoning, distillation from a base initialization without offline SFT, and multi-turn agents. We also compare alternative trigger designs and teacher-intervention rates. DTDD reduces the post-warmup peak gradient norm by $8.6\times$ and, from a base initialization, reaches 21.3% held-out math accuracy in $2.6\times$ fewer optimizer updates than OPD; both methods finish near 22%. It improves best-observed performance by 4.8 and 4.1 percentage points on ALFWorld and agentic code, respectively. Ablations show that segment-level triggers outperform token-level triggers. Behavior cloning is also more effective than the standard OPD objective for training teacher-generated recoveries. Across 23 experimental arms, takeover tends to help when the teacher substantially outperforms the OPD-trained student, provided that the teacher can recover from student-visited states, the student can learn those recoveries, and teacher intervention is neither excessive nor insufficient.

1 INTRODUCTION

Knowledge distillation transfers capabilities from a strong teacher to a smaller student (Hinton et al., 2015). Training on fixed teacher-generated sequences creates a distribution mismatch: the student learns from teacher-generated prefixes during training but must condition on its own prefixes at inference (Kim & Rush, 2016; Agarwal et al., 2024). One motivation for on-policy distillation (OPD) is to reduce this state-distribution mismatch by sampling student trajectories and querying the teacher at student-visited states (Agarwal et al., 2024). OPD therefore combines student-state coverage with dense teacher supervision. The quality of this signal depends on whether the teacher provides learnable guidance along the states the student visits.

Reliable on-policy supervision can fail for two broad reasons: the teacher signal may be poorly matched to the student at initialization, or it may deteriorate as errors accumulate along a rollout. In reasoning, incompatible student–teacher thinking patterns and limited teacher information gain can make early OPD ineffective, motivating SFT cold starts and teacher-aligned prompts (Li et al., 2026b). Even when initial learning is viable, longer reasoning rollouts can amplify deviations and destabilize training (Luo et al., 2026). In multi-turn tool use, an incorrect tool call additionally changes subsequent observations and contexts, causing errors and student–teacher divergence to compound across interactions (Zhong et al., 2026; Wang et al., 2026). Figure 1 illustrates these cases. Despite their different origins, each can move the student into prefixes where local reverse-KL feedback is no longer informative (Liu et al., 2026; Fu et al., 2026). OPD ensures coverage of student-visited states, but effective learning still requires a reliable and learnable signal at those states.

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

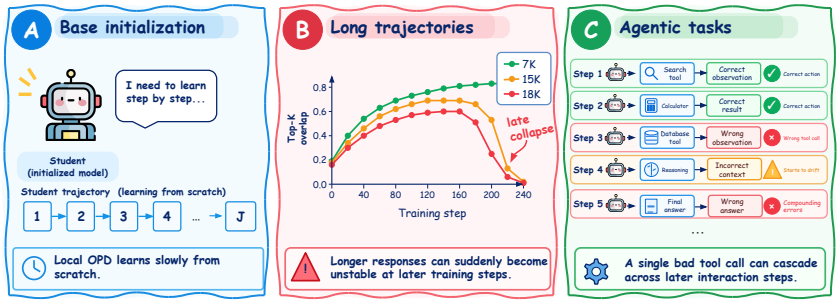

Figure 1: Representative failures of local OPD supervision: **(A)** slow learning from a base initialization, **(B)** late instability under long generation budgets, and **(C)** compounding errors after an incorrect action in multi-turn interaction. Appendix A reports measured examples of (A) and (B).

Prior work addresses unreliable OPD supervision through two broad strategies. The first modifies the token-level learning signal while retaining the student trajectory: StableOPD constrains policy drift and mixes reference rollouts, TGPO introduces directional teacher guidance, and SOD attenuates distillation in high-divergence regions (Luo et al., 2026; Liu et al., 2026; Zhong et al., 2026). These methods protect or redirect optimization within the student trajectory. The second strategy introduces coherent teacher-generated corrections, including curriculum-controlled trajectory segments, expert suffixes, rewritten trajectories, per-step teacher targets, and temporary teacher relays (Wang et al., 2026; Lauffer et al., 2025; Li et al., 2026a; Jiang et al., 2026; Sun et al., 2026; Xu et al., 2026; Peng et al., 2025). Existing approaches typically either reshape supervision while retaining the student continuation, or introduce teacher corrections without combining state-dependent intervention with source-specific training. This leaves open how to preserve OPD where it remains effective and switch to direct teacher supervision where it does not.

To address this gap, we propose **DTDD**, a **D**ivergence-**T**riggered **D**ynamic **D**istillation framework built around a simple teacher-takeover rule. DTDD partitions each student rollout into short token segments and measures student–teacher divergence within each segment. When a segment exhibits high divergence, DTDD rewinds to its beginning and lets the teacher complete the trajectory from the retained student prefix. The prefix continues to receive standard OPD supervision, while the teacher-generated suffix is trained with behavior cloning, an online SFT-style loss. Without a trigger, DTDD reduces to standard OPD. The same mechanism can intervene sparsely to stabilize an already capable student or more frequently to provide coherent supervision from a base initialization. In multi-turn settings, takeover instead replaces the current reasoning-and-tool-call action before interaction continues.

We evaluate DTDD on long-horizon math, base-model distillation without offline SFT, and multi-turn agents. With 15K- and 18K-token response budgets, it preserves held-out accuracy while reducing post-warmup peak gradient norms by $8.6\times$ and $7.0\times$. From Qwen3-1.7B-Base, high-dose DTDD reaches 21.3% held-out accuracy after roughly 70 updates and stays near that level; OPD reaches it around update 180, and both finish near 22%. At comparable accuracy, DTDD requires 22.0 serial hours and 43M teacher-written tokens; offline SFT followed by OPD requires 42.8 hours and 95M. Action-level takeover improves the best ALFWorld and agentic-code results by 4.8% and 4.1%. Ablations show that segment-averaged divergence outperforms individual token spikes and behavior cloning outperforms OPD training on teacher suffixes. Across 23 arms in 13 settings, 19 follow a common pattern. DTDD tends to help when the teacher is stronger than the OPD student, can recover from student prefixes, and provides continuations the student can learn; the intervention rate must provide enough guidance without replacing on-policy training.

2 Related Work

On-policy distillation. Knowledge distillation transfers teacher knowledge through output-level supervision (Hinton et al., 2015). Conventional off-policy approaches either match teacher token distributions under fixed contexts or train on complete teacher-generated sequences (Kim & Rush, 2016). On-policy distillation instead computes teacher supervision along student-generated trajectories (Gu et al., 2024; Agarwal et al., 2024). Recent analyses identify thinking-pattern incompatibility, limited teacher information gain, rollout-length effects, and unreliable sampled-token feedback

as important failure sources (Li et al., 2026b; Fu et al., 2026). To reduce initial pattern mismatch, Li et al. (2026b) first fine-tune the student on offline teacher rollouts before applying OPD. StableOPD (Luo et al., 2026) combines a reference-based divergence constraint with auxiliary SFT on fixed, high-quality "golden" demonstrations to counter repetition-driven length inflation. Other methods reshape supervision within the student rollout: SOD (Zhong et al., 2026) reweights interaction steps using step-level divergence, while Prune-OPD (Yang et al., 2026) downweights and truncates a rollout after detecting prefix drift. These approaches improve initialization, augment the training distribution, or modify the original learning signal. DTDD complements them by conditionally regenerating an unreliable region of the current student trajectory.

Teacher-generated trajectory intervention. A complementary line of work changes the training trajectory itself by introducing teacher-generated corrections. These methods differ primarily in how the intervention boundary is chosen. TRD (Jiang et al., 2026) rewrites complete student rollouts, whereas SOPD (Sun et al., 2026) requests a coherent teacher continuation from every visited prefix. Other methods allocate trajectory control independently of the current student–teacher discrepancy. OEC (Lauffer et al., 2025) uses a random one-way expert handoff; TCOD (Wang et al., 2026) schedules the horizon controlled by the student, with its backward-to-forward curriculum replaying prefixes from pre-collected successful teacher trajectories. More closely related to our setting, Relay-OPD (Xu et al., 2026) and DASH-OPD (Xia et al., 2026) use state-dependent signals to insert temporary teacher segments. AdaSwitch (Peng et al., 2025) uses a context-relative next-token divergence spike to trigger a one-way teacher takeover and optimizes the resulting hybrid sequence with a unified KD objective. Compared with AdaSwitch, DTDD makes two design changes. First, it averages divergence over short token segments and triggers at the first segment whose average exceeds the threshold, rather than reacting to an individual token spike. Second, in reasoning tasks, DTDD rewinds to that segment, retains the preceding student prefix, and asks the teacher to regenerate the remaining suffix. The student prefix is trained with OPD, while the teacher-generated suffix is trained with behavior cloning. In multi-turn tasks, the same trigger replaces the current action or transfers the remaining episode to the teacher, depending on the interactive realization.

3 Divergence-Triggered Dynamic Distillation

We first formulate the OPD objective, then present DTDD and formalize its segment-level trigger, teacher takeover, and training objective.

3.1 On-Policy Distillation Preliminaries

Let $x \sim \mathcal{D}_x$ be a prompt, π_θ the student, and π_T a frozen teacher. The student samples an autoregressive rollout $y^S = (y_1^S, \ldots, y_L^S) \sim \pi_\theta(\cdot \mid x)$ under a maximum generation budget R, where $L \le R$. At position t, let $h_t = (x, y_{<t}^S)$ denote the student-visited history, with $p_t(v) = \pi_\theta(v \mid h_t)$ and $q_t(v) = \pi_T(v \mid h_t)$ denoting the student and teacher next-token distributions. On-policy distillation evaluates the teacher along trajectories sampled from the current student (Agarwal et al., 2024). We use a reverse-KL objective (Gu et al., 2024), writing $\ell_t^{\mathrm{RKL}}(\theta) = D_{\mathrm{KL}}(p_t \| q_t)$ for its per-position loss:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{\substack{x \sim \mathcal{D}_x, \\ y^S \sim \pi_\theta(\cdot|x)}} \left[\frac{1}{L} \sum_{t=1}^{L} \ell_t^{\mathrm{RKL}}(\theta) \right]. \tag{1}$$

To obtain multi-token supervision without evaluating the full vocabulary, we retain the student's K highest-probability next-token candidates. Let $\mathcal{V}_t = \mathrm{TopK}(p_t, K)$ and $\bar{p}_t(v) = p_t(v) / \sum_{u \in \mathcal{V}_t} p_t(u)$ for $v \in \mathcal{V}_t$. For each candidate, we construct the token-level signal

$$A_{t,v}^{\mathrm{OPD}} = \bar{p}_t(v) \big(\log q_t(v) - \log p_t(v) \big). \tag{2}$$

A positive signal increases the probability of a candidate favored more strongly by the teacher, whereas a negative signal suppresses a candidate over-preferred by the student. We treat these signals as detached token-level advantages during optimization. The corresponding unclipped surrogate is

$$\widehat{\ell}_t^{\mathrm{OPD}}(\theta) = - \sum_{v \in \mathcal{V}_t} \mathrm{sg}\big[A_{t,v}^{\mathrm{OPD}} \big] \log \pi_\theta(v \mid h_t), \tag{3}$$

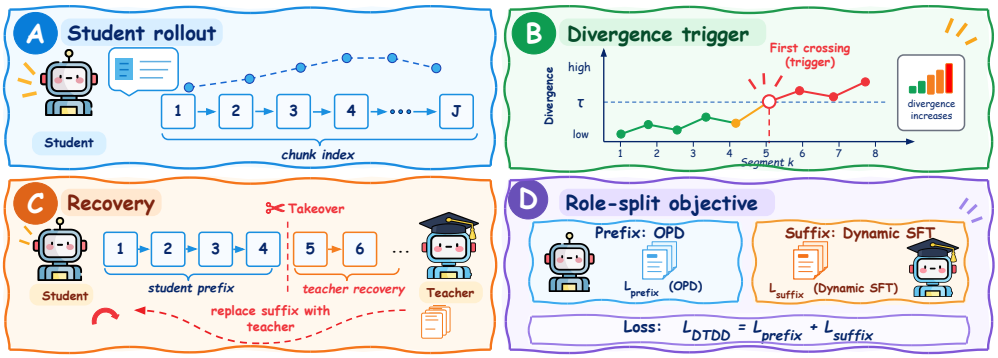

Figure 2: Reasoning realization of DTDD. (A) The student generates a segmented rollout. (B) The first threshold crossing selects the rewind boundary. (C) The teacher regenerates the suffix from the retained student prefix. (D) The prefix is trained with OPD and the suffix with behavior cloning; without a crossing, DTDD reduces to OPD.

where sg denotes stop-gradient. This is the policy-gradient surrogate used to optimize the top-K approximation of reverse KL. We use it for reasoning OPD, while Section 3.3 defines the separate nonnegative discrepancy used for triggering.

3.2 DTDD Overview

DTDD routes supervision in three stages (Figure 2). First, the student generates an on-policy rollout partitioned into fixed-length segments. A prespecified controller, configured for the desired intervention regime, selects the first segment whose aggregated student–teacher divergence crosses its threshold. Second, DTDD rewinds to the preceding boundary and asks the frozen teacher to regenerate the remaining suffix from the retained student prefix. Third, the student prefix receives OPD supervision, while the teacher suffix receives behavior cloning. If no crossing occurs, the entire rollout receives standard OPD.

The controller setting determines the intervention dose. Frequent takeover supplies coherent trajectory-level guidance when a base student has not acquired a teacher-compatible reasoning pattern, whereas sparse takeover repairs localized failures in otherwise effective OPD. In multi-turn tasks, the same routing logic operates at action boundaries, where the teacher generates from the triggering state before interaction continues.

3.3 Segment-Level Trigger and Recovery

We now define the segment discrepancy, intervention boundary, and recovery trajectory. Algorithm 1 summarizes the reasoning realization.

We construct the trigger from a nonnegative token-level discrepancy that captures the magnitude of local disagreement rather than its direction. In our reasoning experiments, it is computed from the same student top-K comparisons as the sparse OPD estimator:

$$\delta_t = \sum_{v \in \mathcal{V}_t} \bar{p}_t(v) \left| \log p_t(v) - \log q_t(v) \right| . \tag{4}$$

Let $(W_1, \ldots, W_J)$ be an ordered partition of the model-generated positions into *segments*. We define the discrepancy of segment j as

$$d_j = \frac{1}{|W_j|} \sum_{t \in W_j} \delta_t . \tag{5}$$

For a reasoning rollout, we use fixed-length segments of C tokens: $J = \lceil L/C \rceil$ and $W_j = \{(j-1)C + 1, \ldots, \min(jC, L)\}$. For an agentic trajectory, each W_j instead contains the tokens of the j-th model-generated action; environment observations provide context but are not included in the discrepancy.

Algorithm 1 DTDD Training for Reasoning Rollouts

Require: Student π_θ; frozen teacher π_T; threshold controller H; rollout parameters $R, C, j_{\min}$

for $s = 1, \ldots, S$ **do**
  Roll out $\mathcal{B} = \{(x_i, y_i^S)\}$ with $y_i^S \sim \pi_\theta(\cdot \mid x_i)$
  Compute $d_{i,1:J_i}$ by Equations 4–5; $\tau_s \leftarrow H_s(\{d_{i,j}\}_{i,j})$
  for each $(x_i, y_i^S) \in \mathcal{B}$ **do**
    $j_i^\star \leftarrow$ first $j \geq j_{\min}$ with $d_{i,j} > \tau_s$, if one exists
    if no such $j_i^\star$ exists **then**
      $(c_i, \widetilde{y}_i) \leftarrow (L_i, y_i^S)$ {standard OPD}
    else
      $c_i \leftarrow (j_i^\star - 1)C$ {rewind}
      Sample $y_i^T \sim \pi_T(\cdot \mid x_i, y_{i,1:c_i}^S)$; $\widetilde{y}_i \leftarrow [y_{i,1:c_i}^S; y_i^T]$ {teacher recovery}
    end if
  end for
  Update θ using $\mathcal{L}_{\mathrm{DTDD}}$: OPD on student prefixes and behavior cloning on teacher suffixes
end for

Let $j_{\min}$ be the first eligible segment and τ_s the threshold at optimization step s. DTDD selects the first eligible threshold crossing, $j^\star = \min\{j \geq j_{\min} : d_j > \tau_s\}$. If no eligible segment crosses the threshold, DTDD leaves the trajectory unchanged and applies standard OPD. Averaging divergence within a segment prevents isolated token-level spikes from causing intervention.

We instantiate τ_s with either a fixed or rate-controlled threshold. Both variants use divergence to select which rollouts trigger and the first crossing to determine the intervention boundary. Appendix B.1.1 gives the controller and recovery details.

In reasoning tasks, when a trigger occurs, we define the rewind position $c = (j^\star - 1)C$ as the boundary immediately preceding the triggering segment. DTDD then replaces the student suffix $y_{c+1:L}^S$ with a continuation $y_{c+1:\widetilde{L}}^T$ generated by the frozen teacher conditioned on $(x, y_{1:c}^S)$ within the remaining generation budget. The resulting hybrid trajectory is $\widetilde{y} = \left[y_{1:c}^S;\, y_{c+1:\widetilde{L}}^T\right]$. Thus the teacher replaces the divergent segment itself, rather than inheriting it and correcting only later tokens. Once control is transferred, the teacher completes the suffix without returning generation to the student. If no trigger occurs, we set $c = L$ and $\widetilde{y} = y^S$, so DTDD reduces exactly to OPD.

The recovery horizon follows how errors propagate. In agentic code, later tool calls depend on earlier outputs, so errors can compound. When a student action triggers intervention, that action is discarded and the teacher takes over from that point, generating the replacement action and the rest of the episode. In ALFWorld and WebShop, the teacher replaces only the triggering action, and the student resumes at the next turn from the resulting environment state. We use this shorter recovery horizon because each executed action yields a new observation that provides a natural point for returning control to the student. Appendix B.2 gives the corresponding training objectives.

3.4 Role-Split Training Objective

For a recovered rollout, DTDD applies OPD to the retained student prefix and behavior cloning to the teacher-generated suffix. Behavior cloning is standard token-level cross-entropy: each teacher-written token is the target, conditioned on the preceding hybrid trajectory. Let $\widehat{\ell}_t^{\mathrm{OPD}}(\theta)$ be the sparse OPD surrogate from Equation 3. The suffix loss is $\ell_t^{\mathrm{BC}}(\theta) = -\log \pi_\theta(\widetilde{y}_t \mid x, \widetilde{y}_{<t})$, and the combined objective is

$$\mathcal{L}_{\mathrm{DTDD}}(\theta) = \mathbb{E}\left[\frac{1}{\widetilde{L}}\left(\sum_{t=1}^{c} \widehat{\ell}_t^{\mathrm{OPD}}(\theta) + \lambda_{\mathrm{BC}} \sum_{t=c+1}^{\widetilde{L}} \ell_t^{\mathrm{BC}}(\theta)\right)\right], \tag{6}$$

where λ_{BC} weights suffix supervision. If no trigger occurs, $c = L$ and $\widetilde{y} = y^S$, so the objective reduces to OPD. We use $\lambda_{\mathrm{BC}} = 1$ in the main experiments. Section 4.1 (Table 2) varies this weight from 0.1 to 3 and also evaluates a soft cross-entropy target.

Table 1: Design controls on DAPO-Math-17K (average@16, %). OPD/DTDD: two-seed mean$_{\pm \mathrm{SD}}$ (sample SD); others: one run.

Method / trigger	Loss	Optimizer step			
		60	80	100	120
OPD	–	$12.84_{\pm 0.68}$	$14.00_{\pm 0.01}$	$13.68_{\pm 0.30}$	$14.54_{\pm 0.16}$
DTDD (segment)	BC	$20.32_{\pm 0.23}$	$21.47_{\pm 0.18}$	$21.08_{\pm 0.25}$	$20.83_{\pm 0.10}$
Token spike	KD	12.47	12.14	12.81	13.51
Token spike	BC	13.19	13.07	13.26	16.79
Segment	KD	5.78	7.80	8.04	10.23
AdaSwitch	KD	18.16	19.57	18.68	19.95
SFT + OPD	–	21.56	22.81	22.37	22.03

Table 2: Suffix-loss ablation (Table 1 setup). Hard: one-hot; Soft: soft cross-entropy.

Target	λ_{BC}	Optimizer step			
		60	80	100	120
Hard	1	20.16	21.34	21.25	20.76
Hard	3	21.15	21.48	19.91	19.10
Hard	0.1	20.94	20.31	22.40	20.78
Soft	1	21.27	22.01	19.95	20.59

4 EXPERIMENTS

We evaluate DTDD across long-form reasoning and multi-turn agents, asking whether takeover improves long-horizon stability (**Q1**), base-model learning efficiency (**Q2**), transfer to interactive tasks (**Q3**), and which student–teacher regimes support intervention (**Q4**).

4.1 EXPERIMENTAL SETUP

Evaluation settings. We evaluate DTDD in two complementary regimes. The reasoning suite tests long-horizon stability by distilling DeepSeek-R1-Distill-Qwen-1.5B from JustRL-DeepSeek-1.5B on DAPO-Math-17K (DeepSeek-AI, 2025; Yu et al., 2025a), and base-initialization efficiency by pairing Qwen3-1.7B-Base with Qwen3-4B-Base-GRPO on DAPO-Math-17K and with a Qwen3-4B non-thinking teacher on OpenThoughts3 (Yang et al., 2025; Guha et al., 2025). Both math settings follow the experimental configurations in Li et al. (2026b). The interactive suite tests action-level takeover on ALFWorld (Shridhar et al., 2020) with a Qwen2.5-3B-Instruct student (Yang et al., 2024) and GiGPO-7B teacher (Wang et al., 2026), and on the code subset of Open-AgentRL-30K with a Qwen3-1.7B-SFT student and DemyAgent-GRPO-4B teacher.

Comparison and metrics. Vanilla OPD is the direct baseline. We also evaluate AdaSwitch (Peng et al., 2025) in the base-student setting (Table 1; Appendix D.3). Within each setting, paired OPD and DTDD runs share the student initialization, teacher, prompts, batch order, generation budget, and optimizer, providing a controlled comparison of the intervention controller. For training-dynamics and ablation claims, we compare fixed shared updates or the full learning curves. Table 3 reports the best observed checkpoint for each method, selected by Average within its block. Appendix Tables 6 and 7 report shared endpoints and matched-curve summaries, respectively. We report average@16 accuracy on AIME 2024, AIME 2025, and AMC 2023 for math; success on the seen and unseen splits for ALFWorld; and average@8 execution accuracy on LiveCodeBench (Jain et al., 2025) and held-out in-distribution problems for agentic code. Appendix B specifies the full training, sampling, checkpoint-selection, and reporting protocols.

Design controls. Table 1 ablates two choices in the base-student setting: the triggering mechanism and the teacher-suffix loss. We cross segment-averaged divergence and token-spike triggers with BC (cross-entropy on teacher tokens) and KD (the top-K OPD surrogate on both prefix and suffix), yielding four configurations. AdaSwitch (Peng et al., 2025) pairs a relative token-level trigger with unified KD; we reproduce its moving-average threshold with a multiplier of 3. At step 120, DTDD reaches 20.83%, compared with 10.23–16.79% for the other three and 19.95% for AdaSwitch. The suffix-loss variants retain the early accuracy gain across weights and target types (Table 2). Offline SFT followed by OPD provides a complementary warm-up baseline. Appendix D.3 reports the individual seeds, trigger details, and controller costs. Overall, DTDD provides the best-performing online trigger–loss combination across all four checkpoints.

4.2 Q1: DOES DTDD STABILIZE LONG-HORIZON OPD?

Long generation budgets expose a late OPD instability that DTDD suppresses. We hold the student, teacher, data, sampling, and optimizer fixed while varying the maximum rollout length over 7,168,

15,360, and 18,432 tokens. We quantify late-training instability by the maximum gradient norm from update 51 onward, after the short-lived optimization transient at the start of training.

The 7K control remains stable under both methods, with peak gradient norms of 0.54 for OPD and 0.96 for DTDD. At 15K, OPD enters a sharp late event: its gradient norm reaches 12.02 and its truncated-rollout ratio reaches 52%. DTDD limits the same quantities to 1.40 and 32%, reducing the gradient peak by $8.6\times$. The 18K run reproduces the pattern, with gradient peaks of 12.81 for OPD and 1.84 for DTDD ($7.0\times$). Held-out accuracy remains within one point across all three budgets. Thus, DTDD removes the long-budget optimization failure while preserving the performance level reached by OPD. Appendix C provides the complete entropy, truncation, overlap, reward, and accuracy trajectories.

Clipping the per-token OPD signal to $[-1, 1]$ also achieves comparable stabilization without teacher continuations: peak gradient norms fall below one at both long budgets, while held-out accuracy remains within 1.1 points of the reproduced OPD baseline. Appendix C reports these controls and the weaker $[-5, 5]$ clamp, which leaves the spikes largely intact.

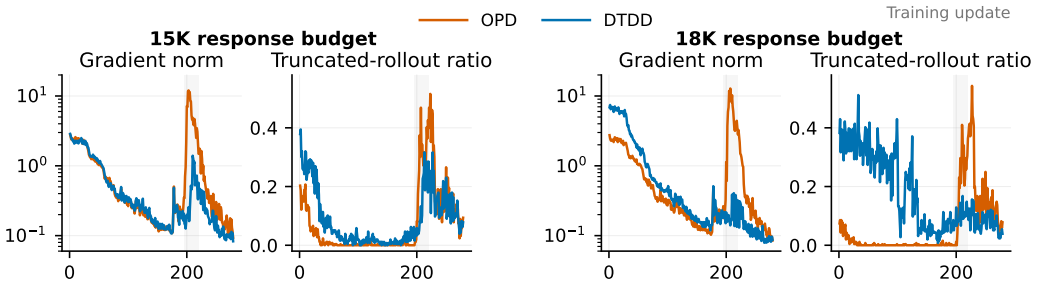

Figure 3: DTDD suppresses late OPD spikes on DAPO-Math-17K at 15K/18K response budgets.

4.3 Q2: Can Takeover Warm Up a Base Student Online?

DTDD can warm up a base student online, without first preparing an SFT dataset and warm-starting the student on it. Figure 4 compares a Qwen3-1.7B-Base student with a Qwen3-4B-Base-GRPO teacher on DAPO-Math-17K. High-dose DTDD reaches the 21.3% held-out reference level near update 70, whereas OPD reaches it near update 180. The trajectories later converge near 22% average accuracy. Appendix D reports the complete dose, window-size, and cost analyses.

We measure intervention frequency by the *takeover rate*: the fraction of sampled rollouts in which divergence triggers a teacher takeover. *Rewind* replaces the divergent suffix; *restart* replaces the entire rollout. Fixed thresholds allow takeover rates to vary with divergence, whereas target-rate controllers adjust τ to maintain a prescribed rate.

Figure 4 compares OPD with four rewind configurations. Three form a window-size ablation: $C \in \{64, 128, 256\}$, with fixed thresholds $\tau \in \{0.475, 0.362, 0.296\}$ calibrated to match an initial takeover rate of approximately 90%. The fourth fixes $\tau = 0.2$ at $C = 256$ and makes the third segment ($j_{\min} = 3$) the first eligible for takeover. This control starts near 25% takeover, providing limited teacher supervision early in training. As model rollouts become longer, more of them satisfy the trigger; the resulting increase in takeover rate supplies additional teacher supervision.

At update 120, the three high-initial-takeover arms reach 21.32% ($C = 64$), 20.90% ($C = 128$), and 21.23% ($C = 256$), whereas the fixed-$\tau = 0.2$ control remains close to OPD (14.81% versus 14.65%). A separate rate-controlled restart setting targeting 45% takeover reaches 20.49%. By update 240, all four rewind configurations reach 21.04–22.40%, comparable to OPD at 22.29%. Thus, for a base student without prior SFT warm-up, sufficient coherent teacher continuations at a high initial dose accelerate online warm-up; segment-level divergence routes them online and lets fixed-threshold intervention recede with alignment.

The three high-coverage DAPO arms—excluding the low-initial-takeover $\tau = 0.2$ control—show decreasing segment divergence and teacher-written token share, together with later takeover positions (Appendix Figure 11). Figure 5 illustrates these dynamics on OpenThoughts3.

With $\tau = 0.87$, the OpenThoughts3 takeover rate falls from 88.3% to below 5% by update 87 and to zero by update 150, while the median trigger moves from window 1 to 10.5. Among rollouts

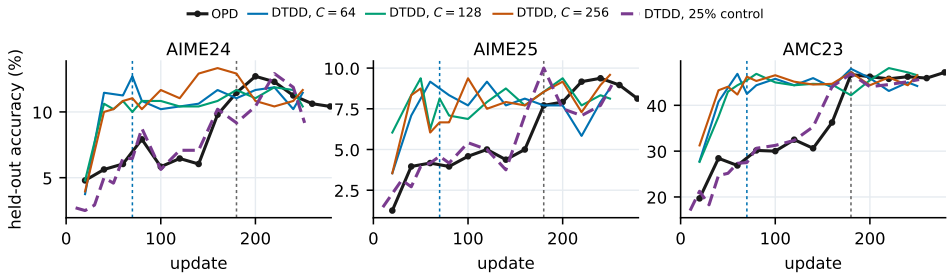

Figure 4: Held-out accuracy for Qwen3-1.7B-Base ← Qwen3-4B-Base-GRPO on DAPO-Math-17K. The length-gated control starts near 25% takeover and adapts thereafter.

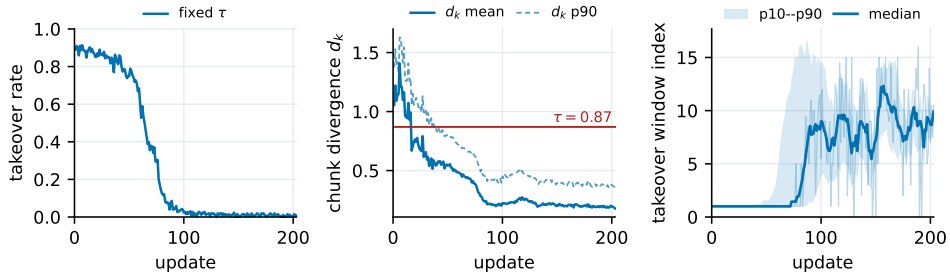

Figure 5: Fixed-threshold dynamics ($\tau = 0.87$) for Qwen3-1.7B-Base ← Qwen3-4B on OpenThoughts3.

that still trigger, the later position retains a longer student prefix before teacher continuation. At matched update 200, the three OpenThoughts3 controllers reach 14.36–15.75% accuracy, compared with 12.85% for OPD (Appendix Table 6).

On DAPO-Math-17K, DTDD reaches 21.34% held-out accuracy in 22.0 hours; offline SFT followed by OPD reaches 21.55% in approximately 42.8 serial hours, including teacher-data generation and SFT. DTDD uses 43M versus 95M teacher-written tokens, allocated online within one training stage. Appendix D.5 reports the checkpoint-level timing and hardware details.

4.4 Q3: Does Divergence-Guided Routing Transfer to Agentic Tasks?

In multi-turn tasks, DTDD triggers teacher recovery at action boundaries, where each action contains a reasoning span and tool call. Takeover is used only during training; evaluation uses the student alone.

On ALFWorld, we pair a LoRA-adapted Qwen2.5-3B-Instruct student with the ALFWorld-specific GiGPO-7B teacher and set $\tau = 0.7$. As in the reasoning experiments, training begins with high teacher coverage (90.2% takeover). As the student improves over training, takeover falls to near zero, automatically reducing teacher supervision. Near update 180, a transient rise in discrepancy reactivates takeover despite its previously near-zero rate (Appendix Figure 12). DTDD therefore also acts as an online monitor and recovery mechanism: it remains mostly inactive after alignment but resumes corrective teacher supervision when instability emerges. With checkpoints evaluated every 20 updates, DTDD leads OPD at all ten matched checkpoints after update 200 in this run. Best-observed average success increases from 76.49% to 81.32%, with the largest gain on the unseen split (Table 3). Appendix E.1 reports the full trajectory and adaptation controls, including the matched update-400 comparison.

The agentic code experiment examines intervention dose for an SFT-initialized policy, pairing Qwen3-1.7B-SFT with DemyAgent-GRPO-4B on the code-only subset of Open-AgentRL-30K (Yu et al., 2025b). Our main setting uses $\tau = 0.50$, yielding a sparse 2.8% teacher-token dose. At their best-observed checkpoints, sparse takeover improves all four evaluation splits and raises aggregate accuracy from 36.49% to 40.59% (Table 3). This result complements the base-student study: high coverage supplies a missing reasoning pattern, whereas sparse correction preserves the on-policy interaction of an already viable agent. Appendix E.2 reports matched endpoints and the dose ablation.

Table 3: Best-observed results (%; selection details in Appendix B.3). Subscripts show percentage-point changes from OPD; bold marks the best trained method.

Math reasoning: student Qwen3-1.7B-Base; teacher Qwen3-4B (non-thinking); data OpenThoughts3

Method	AIME24	AIME25	AMC23	Average ↑
Teacher	*21.04*	*20.62*	*67.66*	*36.44*
OPD	5.62	5.42	27.50	12.85
DTDD fixed $\tau = 0.87$	6.88 $_{+1.26}$	**7.08** $_{+1.66}$	31.41 $_{+3.91}$	15.12 $_{+2.27}$
DTDD rewind, target $90\% \rightarrow 5\%$	7.29 $_{+1.67}$	4.79 $_{-0.63}$	32.19 $_{+4.69}$	14.76 $_{+1.91}$
DTDD restart, target $90\% \rightarrow 5\%$	**8.75** $_{+3.13}$	5.21 $_{-0.21}$	**33.28** $_{+5.78}$	**15.75** $_{+2.90}$

Agentic interaction: student Qwen2.5-3B-Instruct (LoRA); teacher GiGPO-7B; data ALFWorld

Method	Seen	Unseen	Average ↑
Teacher	*85.71*	*88.06*	*86.89*
OPD	83.57	69.40	76.49
DTDD	**84.29** $_{+0.72}$	**78.36** $_{+8.96}$	**81.32** $_{+4.83}$

Agentic code: student Qwen3-1.7B-SFT; teacher DemyAgent-GRPO-4B; data Open-AgentRL-30K (code only)

Method	HO-200 full	HO-200 sub	LCB v1	LCB v6	Average ↑
Teacher	*62.12*	*64.19*	*68.78*	*47.43*	*60.63*
OPD	31.00	30.56	50.91	33.48	36.49
DTDD	**35.81** $_{+4.81}$	**37.62** $_{+7.06}$	**53.65** $_{+2.74}$	**35.27** $_{+1.79}$	**40.59** $_{+4.10}$

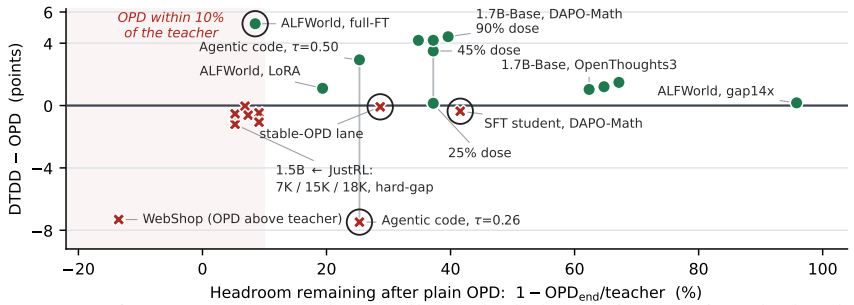

Figure 6: DTDD minus OPD, averaged over all matched checkpoints, against the headroom plain OPD leaves below its teacher. Points left of zero are settings where OPD ends above the teacher's own score. Ringed markers identify arms analyzed by the validity, learnability, and dose controls in Appendix F.

4.5 Q4: What Determines Whether Teacher Takeover Helps?

Headroom is an empirical diagnostic of the observed regimes. We define it as $h = 1 - \mathrm{OPD}_{\mathrm{end}}/\mathrm{teacher}$, with both scores measured on the same evaluation harness. Across 23 arms from 13 settings, 19 follow a common pattern: gains concentrate in regimes where OPD retains more headroom below the teacher (Figure 6).

Three controls refine this diagnostic. In the rebuilt WebShop environment, the task-specific teacher falls below OPD, so takeover routes training toward an invalid supervision source. Conversely, the 0.5B ALFWorld student retains substantial teacher headroom but does not reliably learn the recovery, exposing a student-learnability constraint. Finally, the agentic-code dose ablation shows that excessive intervention can replace too much on-policy interaction. Appendix F reports the complete settings, per-arm results, and TCOD comparison.

5 Conclusion and Limitations

DTDD uses segment-level student–teacher divergence to route supervision between standard OPD and teacher-generated recovery. Across paired experiments, sparse takeover suppresses long-budget optimization extremes, high initial teacher coverage accelerates learning from a base initialization, and action-level takeover improves ALFWorld and agentic code. Controlled comparisons favor segment-level triggering with behavior cloning. Relative to offline SFT followed by OPD, this single-stage approach reaches comparable accuracy while cutting measured serial time by a factor of 1.9 and using 55% fewer teacher-written tokens. DTDD uses manually specified thresholds or target takeover rates. Its divergence trigger does not assess whether teacher recovery is useful for the student. Automatic dose calibration and recovery-quality assessment are directions for future work.

Reproducibility Statement

Section 4.1 and Appendix B specify the models, datasets, generation budgets, decoding parameters, objectives, evaluation pipeline, and checkpoint-selection rules. Paired OPD and DTDD runs share the student initialization, teacher, prompts, generation budget, batch order, and optimizer configuration. Each training curve corresponds to one run; average@k reports evaluation-time sampling variation from repeated generations.

AI Use Statement

Generative AI tools, including Anthropic Claude and OpenAI Codex, were used to provide feedback on method formulation and experimental design; assist with code development, debugging, and analysis of experiment logs; support literature search and citation verification; and draft, edit, and format portions of the manuscript, tables, and figures. The authors reviewed the AI-assisted outputs and checked the resulting code, numerical claims, citations, and manuscript text against the source code, training logs, evaluation outputs, and cited papers. The authors take responsibility for the final content of this work.

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 *The Twelfth International Conference on Learning Representations*, 2024.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Etash Guha et al. OpenThoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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 *The Thirteenth International Conference on Learning Representations*, 2025.

Li Jiang, Haoran Xu, Yichuan Ding, and Amy Zhang. Trajectory-refined distillation. *arXiv preprint arXiv:2606.08432*, 2026.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016.

Niklas Lauffer, Xiang Deng, Srivatsa Kundurthy, Brad Kenstler, and Jeff Da. Imitation learning for multi-turn LM agents via on-policy expert corrections. *arXiv preprint arXiv:2512.14895*, 2025.

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. *arXiv preprint arXiv:2606.15912*, 2026a.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Ning Ding, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026b.

Xinyu Liu, Kechen Jiao, Chunyang Xiao, Runsong Zhao, Junhao Ruan, Bei Li, Jiahao Liu, Qifan Wang, Xin Chen, Jingang Wang, Chenglong Wang, Tong Xiao, and Jingbo Zhu. Teacher-guided policy optimization for on-policy reasoning distillation under large policy divergence. *arXiv preprint arXiv:2605.13230*, 2026.

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. Demystifying OPD: Length inflation and stabilization strategies for large language models. *arXiv preprint arXiv:2604.08527*, 2026.

Jingyu Peng, Maolin Wang, Hengyi Cai, Yuchen Li, Kai Zhang, Shuaiqiang Wang, Dawei Yin, and Xiangyu Zhao. AdaSwitch: Balancing exploration and guidance in knowledge distillation via adaptive switching. *arXiv preprint arXiv:2510.07842*, 2025.

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

Changhui Sun, Lanbo Liu, Hang Lei, Tong Ling, Jiahang Xie, Zhiyong Zheng, Yujia Wang, Hao Liu, Feng Xiao, Lu Liu, Yanlong Du, Zifeng Cheng, Ziwei Jiang, and Qing Gu. Step-level on-policy distillation: Interpolating between on-policy distillation and supervised fine-tuning. *arXiv preprint arXiv:2608.16333*, 2026.

Jiaqi Wang, Wenhao Zhang, Weijie Shi, Yaliang Li, and James Cheng. TCOD: Exploring temporal curriculum in on-policy distillation for multi-turn autonomous agents. *arXiv preprint arXiv:2604.24005*, 2026.

Yuchen Xia, Qianguo Sun, Chao Song, Junlong Wu, Yiyan Qi, and Yunjian Xu. DASH-OPD: Discrepancy-aware switching with hysteresis for on-policy distillation. *arXiv preprint arXiv:2607.29078*, 2026.

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. *arXiv preprint arXiv:2607.26057*, 2026.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhicheng Yang, Zhijiang Guo, Yifan Song, Minrui Xu, Yongxin Wang, Yiwei Wang, Xiaodan Liang, and Jing Tang. Prune-OPD: Efficient and reliable on-policy distillation for long-horizon reasoning. *arXiv preprint arXiv:2605.07804*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025a.

Zhaochen Yu, Ling Yang, Jiaru Zou, Shuicheng Yan, and Mengdi Wang. Demystifying reinforcement learning in agentic reasoning. *arXiv preprint arXiv:2510.11701*, 2025b.

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. *arXiv preprint arXiv:2605.07725*, 2026.

APPENDIX

A Measured Limitations of Vanilla OPD 12
B Full Experimental Details 12
B.1 Reasoning Training Protocol 12
B.1.1 Controller and Recovery Details 12
B.2 Interactive Training Protocol 15
B.3 Evaluation and Reporting 15
C Additional Long-Horizon Stability Results 17
D Additional Base-Initialization Results 19
D.1 Dose Response on DAPO-Math-17K 19
D.2 Window Size and Intervention Form 20
D.3 Controller and Objective Controls 20
D.4 Independent OpenThoughts3 Setting 21
D.5 Update and Wall-Clock Cost 22
E Additional Agentic Results 23
E.1 ALFWorld Action-Level Intervention 23
E.2 Agentic Code Dose 24
F Boundary Cases and Controller Interpretation 24

A MEASURED LIMITATIONS OF VANILLA OPD

Figure 7 provides the measurements underlying the slow base-model warm-up and late long-horizon instability illustrated in panels (A) and (B) of Figure 1.

B FULL EXPERIMENTAL DETAILS

B.1 REASONING TRAINING PROTOCOL

Within each reasoning comparison, OPD and DTDD share one implementation and differ only by the controller in Table 5. Each update samples 64 prompts with four rollouts per prompt. Student and teacher-logit temperatures are both 1.0, and each rollout batch receives one policy update. Optimization uses AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01, and global gradient clipping at 1.0. The learning rate is held constant at 10^{-6} without warm-up. Distillation uses the student top-K set with $K = 16$, student-probability weighting, and token-mean aggregation. The teacher is frozen; no sequence-level outcome reward enters the objective, and teacher-written tokens use $\lambda_{\mathrm{BC}} = 1$. The length sweep sets $j_{\min} = 3$, while the base-initialization studies set $j_{\min} = 1$. The 256-token continuation-budget check does not filter any reported trigger.

Both terms in Equation 6 share the same optimization path. The prefix uses the clipped policy-gradient form of $\widehat{\ell}_t^{\mathrm{OPD}}$; the suffix uses a one-hot target, which gives the standard behavior-cloning gradient when clipping is inactive.

Reasoning takeover is retrospective. The student first generates a full rollout, after which DTDD computes segment discrepancies and requests a teacher continuation for each triggered example. The update retains the student prefix and excludes the replaced student suffix. This pipeline changes the training sequence without reducing student rollout compute, so Section 4.3 measures optimization-step efficiency.

B.1.1 CONTROLLER AND RECOVERY DETAILS

For rollout-level rewind, we suppress triggers too close to the maximum generation budget to leave at least $B_{\min}$ tokens for teacher continuation. We instantiate τ_s with either a fixed or rate-controlled threshold. The fixed controller holds $\tau_s = \tau$, allowing takeover to recede as student–teacher discrepancy falls. The rate-controlled variant sets τ_s to the empirical $(1 - r_s)$-quantile of the maximum eligible segment discrepancy in each rollout, where the target takeover rate r_s may be fixed or scheduled. Both variants use divergence to select which rollouts trigger and the first crossing to determine the intervention boundary. In addition, our implementation triggers a fraction $\epsilon = 0.05$ of rollouts

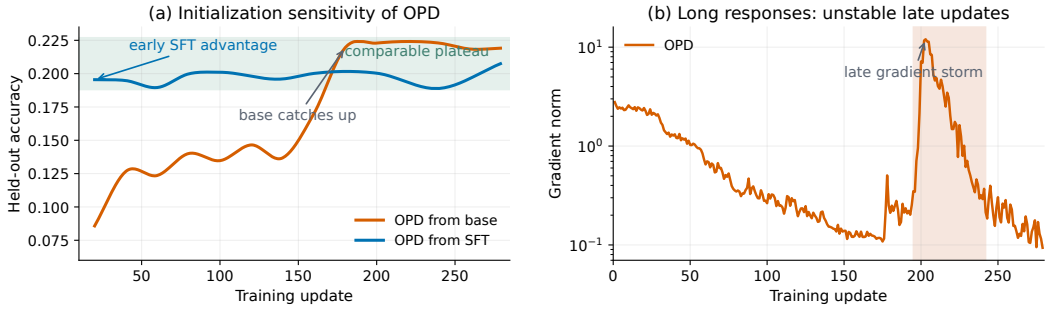

Figure 7: Two representative limitations of vanilla OPD. *Left:* Qwen3-1.7B-Base and Qwen3-1.7B-SFT are distilled from the same Qwen3-4B-Base-GRPO teacher on DAPO-Math-17K with a 7,168-token generation budget. The teacher, data, optimizer, and generation budget are held fixed; the SFT initialization has a large early advantage, while the base initialization eventually reaches a comparable plateau. Held-out accuracy is the mean avg@16 score on AIME 2024, AIME 2025, and AMC 2023. *Right:* DeepSeek-R1-Distill-Qwen-1.5B is distilled from JustRL-DeepSeek-1.5B on DAPO-Math-17K with a 15,360-token generation budget. OPD develops a large late gradient spike despite otherwise decreasing gradient norms.

Table 4: Experimental settings. The budget specifies the maximum generated tokens for reasoning tasks and the maximum number of environment or tool turns for interactive tasks.

Setting	Student	Teacher	Training data	Budget	Evaluation
Rollout length	R1-Distill-Qwen-1.5B	JustRL-DeepSeek-1.5B	DAPO-Math-17K	7,168 / 15,360 / 18,432 tokens	AIME24/25, AMC23
Base student	Qwen3-1.7B-Base	Qwen3-4B-Base-GRPO	DAPO-Math-17K	7,168 tokens	AIME24/25, AMC23
Base student	Qwen3-1.7B-Base	Qwen3-4B (non-thinking)	OpenThoughts3	7,168 tokens	AIME24/25, AMC23
ALFWorld	Qwen2.5-3B-Instruct	GiGPO-7B	ALFWorld train	30 turns; 512 tokens/action	Seen / unseen success
WebShop	Qwen2.5-3B-Instruct	GiGPO-7B (WebShop)	Rebuilt 120K catalog	15 turns; 512 tokens/action	Success
Agentic code	Qwen3-1.7B-SFT	DemyAgent-GRPO-4B	Open-AgentRL-30K (code only)	8 turns / 12,288 tokens	LCB v1/v6, HO-200 full/sub

at a uniformly sampled candidate boundary as an exploration floor. Once segment discrepancy contracts below τ, this floor rather than the divergence signal accounts for most interventions; Section C reports where each length-sweep arm sits relative to the floor. Unless otherwise specified, DTDD uses rewind; we consider full-rollout restart ($c = 0$) only as an ablation.

Table 5: Reasoning configurations. Target rate is a rollout-level control parameter; the realized rate and teacher-token dose are measured from the generated batches. A dash denotes vanilla OPD.

Study	Budget	Segment	Mode	Controller	Initialization / data
Length sweep	7,168 / 15,360 / 18,432	256 tokens	rewind	fixed $\tau = 0.7$, $\epsilon = 0.05$	DeepSeek-R1-Distill-Qwen-1.5B; DAPO-Math-17K
DAPO base, low initial coverage	7,168	256 tokens	rewind	fixed $\tau = 0.2$; initial takeover $\approx 25\%$	Qwen3-1.7B-Base; DAPO-Math-17K
DAPO base, medium coverage	7,168	256 tokens	restart	target rate 45%	Qwen3-1.7B-Base; DAPO-Math-17K
DAPO base, high initial coverage	7,168	64 / 128 / 256 tokens	rewind	fixed $\tau = .475/.362/.296$; initial takeover $\approx 90\%$	Qwen3-1.7B-Base; DAPO-Math-17K
OpenThoughts3 base	7,168	256 tokens	rewind / restart	target rate $90\% \rightarrow 5\%$, or fixed $\tau = 0.87$	Qwen3-1.7B-Base; OpenThoughts3

B.2 Interactive Training Protocol

Interactive experiments apply discrepancy checks at action boundaries. Let $z_j \in \{S, T\}$ denote whether the student or teacher writes action j, let W_j contain its generated-token positions, and let N_S and N_T count student- and teacher-written tokens. Let $a_{j,t}$ be token t of the action executed at turn j, and let π_{old} denote the student that collected the interaction. For both student- and teacher-written actions, the multi-turn trainer uses the detached token signal

$$A^{\text{act}}_{j,t} = \log \pi_T(a_{j,t} \mid h_{j,t}) - \log \pi_{\text{old}}(a_{j,t} \mid h_{j,t}). \tag{7}$$

For student-written actions, define the unclipped sampled-token OPD term $\widehat{\ell}^{\text{OPD}}_{j,t} = -\operatorname{sg}[A^{\text{act}}_{j,t}] \log \pi_\theta(a_{j,t} \mid h_{j,t})$. Agentic code uses the role-split objective

$$\mathcal{L}_{\text{code}}(\theta) = \mathbb{E}\left[\frac{1}{N_S} \sum_{j:z_j=S} \sum_{t \in W_j} \widehat{\ell}^{\text{OPD}}_{j,t}(\theta) - \frac{\lambda_{\text{BC}}}{N_T} \sum_{j:z_j=T} \sum_{t \in W_j} \log \pi_\theta(a_{j,t} \mid h_{j,t})\right], \tag{8}$$

where an empty role-specific term is omitted and the OPD term uses its clipped implementation. For ALFWorld and WebShop, let $N = N_S + N_T$; both action types pass through the sampled-token multi-turn OPD surrogate,

$$\widehat{\mathcal{L}}_{\text{env}}(\theta) = -\mathbb{E}\left[\frac{1}{N} \sum_j \sum_{t \in W_j} \operatorname{sg}\big[A^{\text{act}}_{j,t}\big] \log \pi_\theta(a_{j,t} \mid h_{j,t})\right], \tag{9}$$

implemented with the clipped surrogate. Student actions are sampled from the rollout policy; teacher actions are sampled from the teacher, making the latter term log-ratio-weighted imitation. On teacher-written actions, this is gap-weighted imitation evaluated on teacher tokens rather than unit-weight cross-entropy. Because teacher-written tokens are sampled from the teacher, we treat this term as a practical surrogate rather than an unbiased KL estimator.

ALFWorld pairs a Qwen2.5-3B-Instruct student with GiGPO-7B for episodes of at most 30 turns and 512 generated tokens per action. DTDD computes action-level discrepancy under the current interaction history. When the discrepancy exceeds the threshold, the teacher generates a replacement turn, including its reasoning and action, from the same history. The environment executes the teacher action, and its observation conditions the student's next turn. We test both LoRA and full-model optimization. The LoRA setting uses rank and scale 32 on all linear layers with learning rate 10^{-4}; full fine-tuning uses learning rate 10^{-6}. Both use a training batch size of 32, training temperature 1.0, and evaluation temperature 0.4. WebShop uses the same student, LoRA configuration, and WebShop-specific GiGPO-7B teacher on the rebuilt 120K-product catalog, with at most 15 turns, 512 generated tokens per action, and $\tau = 0.25$.

Our agentic-code implementation builds on the released SOD training stack (Zhong et al., 2026). Initial SOD experiments used the full Open-AgentRL-30K dataset with task rewards and step-wise distillation. The code comparisons reported here use its code-only subset and the revised execution-based evaluation protocol in Appendix B.3.

Agentic code pairs Qwen3-1.7B-SFT with DemyAgent-GRPO-4B on the code-only portion of Open-AgentRL-30K. Each episode allows eight model turns and 12,288 generated tokens. Takeover occurs online at an action boundary, so no future student tokens are generated after a trigger; the teacher generates all remaining turns. Fixed thresholds 0.50 and 0.26 produce the sparse- and high-dose conditions. This implementation applies OPD only to student-written tokens and unit-weight BC to teacher-written action tokens. It uses learning rate 10^{-6}, a training batch of 32 prompts with eight rollouts each, a policy minibatch of 16, and $\lambda_{\text{BC}} = 1$. Training and evaluation use temperature 1.0; evaluation uses top-$p = 0.6$ and eight samples per problem.

All interactive runs use AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01, a constant learning-rate schedule without warm-up, and global gradient clipping at 1.0.

B.3 Evaluation and Reporting

Table 8 summarizes held-out evaluation. Math checkpoints use a 16,384-token generation budget and 16 samples per problem at temperature 0.7 and top-$p = 0.95$; accuracy averages over sam-

Table 6: Matched-checkpoint results (%). All trained methods within each block are evaluated at the same update: 200 for math, 400 for ALFWorld, and 130 for agentic code. Subscripts show the absolute percentage-point change from OPD; bold marks the best trained method. For code, Held-out averages the two HO-200 splits and LCB averages LiveCodeBench v1 and v6.

Math reasoning: student Qwen3-1.7B-Base; teacher Qwen3-4B (non-thinking); data OpenThoughts3

Method	AIME24	AIME25	AMC23	Average ↑
Teacher	*21.04*	*20.62*	*67.66*	*36.44*
OPD	5.62	**5.42**	27.50	12.85
DTDD fixed $\tau = 0.87$	6.88 $_{+1.25}$	**5.42** $_{+0.00}$	31.25 $_{+3.75}$	14.51 $_{+1.67}$
DTDD rewind, target $90\% \to 5\%$	**8.96** $_{+3.33}$	4.58 $_{-0.83}$	29.53 $_{+2.03}$	14.36 $_{+1.51}$
DTDD restart, target $90\% \to 5\%$	8.75 $_{+3.13}$	5.21 $_{-0.21}$	**33.28** $_{+5.78}$	**15.75** $_{+2.90}$

Agentic interaction: student Qwen2.5-3B-Instruct (LoRA); teacher GiGPO-7B; data ALFWorld

Method	Seen	Unseen	Average ↑
Teacher	*85.71*	*88.06*	*86.89*
OPD	70.71	69.40	70.06
DTDD	**84.29** $_{+13.57}$	**70.90** $_{+1.49}$	**77.59** $_{+7.53}$

Agentic code: student Qwen3-1.7B-SFT; teacher DemyAgent-GRPO-4B; data Open-AgentRL-30K (code only)

Method	Held-out	LCB	Average ↑
Teacher	*63.16*	*58.11*	*60.63*
OPD	25.60	**42.26**	33.93
DTDD (2.8% dose)	**31.50** $_{+5.90}$	38.43 $_{-3.83}$	**34.96** $_{+1.03}$
DTDD ($\sim$ 50% dose)	19.78 $_{-5.82}$	36.48 $_{-5.78}$	28.13 $_{-5.80}$

Table 7: Reasoning results across matched reporting rules. All deltas are signed percentage-point differences (DTDD minus OPD) in held-out average@16 accuracy. N is the number of jointly evaluated checkpoints; OPD span is computed over the final five.

Setting	DTDD controller	N	Last	Last 5	All	OPD span
15K generation budget	Fixed $\tau = 0.7$	10	-0.16	-0.73	-0.62	0.43
18K generation budget	Fixed $\tau = 0.7$	14	-0.80	-0.20	-1.20	2.07
DAPO-Math, base start	Rewind, $C = 128$	12	$+0.10$	$+0.56$	$+4.18$	5.40
OpenThoughts3, base start	Fixed $\tau = 0.87$	10	$+1.67$	$+1.68$	$+1.49$	1.46
OpenThoughts3, base start	Rewind, target $90\% \to 5\%$	10	$+1.51$	$+1.47$	$+1.03$	1.46
OpenThoughts3, base start	Restart, target $90\% \to 5\%$	10	$+2.90$	$+1.53$	$+1.21$	1.46

ples and problems. ALFWorld evaluates the student alone, with no teacher intervention. All code methods use the same execution sandbox and decoding parameters.

Table 3 reports the performance envelope obtained by selecting each arm's checkpoint by the Average within its block, considering only checkpoints with results for every benchmark in that block. For OpenThoughts3, the selected updates are 200 for OPD, 180 for fixed-threshold DTDD, 180 for scheduled rewind, and 200 for scheduled restart. ALFWorld selects updates 280 and 140 for OPD and DTDD, respectively; agentic code selects updates 80 and 60.

Appendix Table 6 evaluates every trained method within a block at one shared update: 200 for OpenThoughts3, 400 for ALFWorld, and 130 for agentic code. Stability and boundary analyses use matched checkpoints, and efficiency analyses use the first evaluated checkpoint that reaches the specified performance level. Every curve represents one training run; average@k refers to k evaluation generations. Checkpoints from the same run are correlated and are not treated as independent training replicates; average@k measures evaluation-time sampling variation rather than training-seed uncertainty.

Table 7 summarizes the reasoning comparisons under three matched-checkpoint rules. Last-five Δ averages the DTDD–OPD difference over the final five checkpoints shared by each pair; all-shared Δ averages the same difference over the complete shared training curve. The OPD late span is its maximum minus minimum over the final five shared checkpoints and provides a reference for within-run checkpoint variation. The last shared updates are 279 for the 15K and 18K settings, 240 for DAPO-Math, and 200 for OpenThoughts3.

The OpenThoughts3 gains remain positive at the last checkpoint, over the final five checkpoints, and across the full shared curve. On DAPO-Math, the larger all-shared gain and smaller late-window

Table 8: Held-out evaluation protocol.

Domain	Evaluation set	Sampling / metric
Math	AIME24, AIME25, AMC23	16 samples, $T = 0.7$, top-$p = 0.95$; average@16
ALFWorld	140 seen, 134 unseen tasks	Student-only environment success
Code	LCB v1/v6, HO-200 full/sub	8 samples, $T = 1.0$, top-$p = 0.6$; execution accuracy

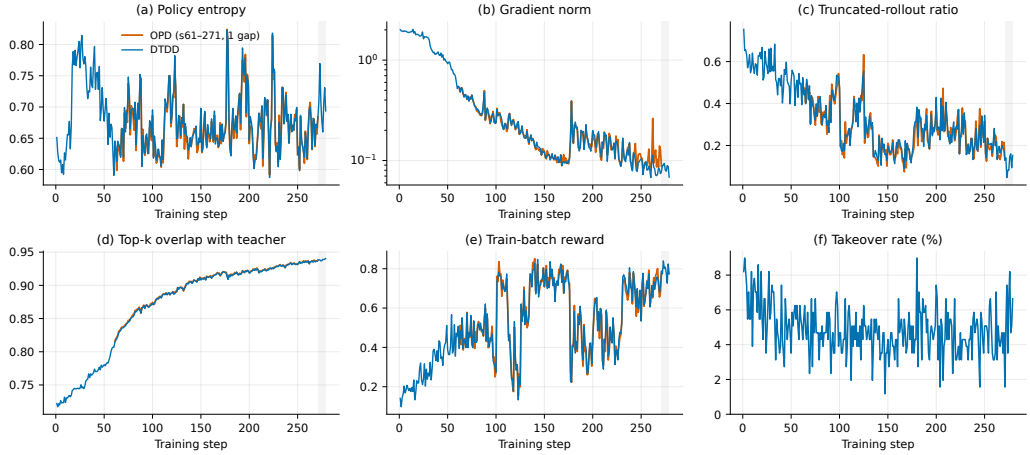

Figure 8: Complete 7K training dynamics for DeepSeek-R1-Distill-Qwen-1.5B distilled from JustRL-DeepSeek-1.5B on DAPO-Math-17K. This setting serves as the stable control. The OPD telemetry begins at update 61; held-out comparisons use checkpoints available for both methods.

gain reflect faster early learning rather than a higher final plateau. The 15K and 18K comparisons do not improve late-window accuracy, consistent with their role as stability interventions.

Stability statistics use the maximum gradient norm after update 50, applying one cutoff across generation budgets. Truncation ratio measures the fraction of rollouts that reach the maximum budget. Takeover rate measures how often an intervention occurs, while teacher-token dose measures how much of the training sequence the teacher writes. Mean takeover position is reported only over triggered examples and is used to locate intervention within a trajectory.

C Additional Long-Horizon Stability Results

The rollout-length study changes only the maximum generation budget. At 7K, OPD and DTDD follow the same entropy, gradient, truncation, teacher-overlap, and train-batch reward trajectories (Figure 8). Both post-warmup gradient peaks remain below one, establishing a stable control for the two longer budgets.

At 15K, OPD destabilizes near update 200: gradient norm and entropy rise, truncation increases, and teacher top-K overlap drops together (Figure 9). DTDD reduces the gradient peak from 12.02 to 1.40 and the truncation peak from 52% to 32%. The matched held-out averages differ by 0.1 point, separating optimization stability from endpoint accuracy.

The 18K run reproduces the coupled event (Figure 10). OPD reaches a gradient norm of 12.81; DTDD remains below 1.84. The takeover rate falls from above 80% to single digits as discrepancy contracts, then rises around the late instability. The matched endpoint accuracy differs by 0.8 point.

Clipped-OPD control. We reproduce plain OPD at both long budgets and clamp its per-token distillation signal before aggregation. Table 10 shows that the tight $[-1, 1]$ clamp removes the optimization event, whereas the $[-5, 5]$ clamp leaves the peak and the number of high-gradient updates nearly unchanged. At update 279, the tight clamp is within 0.9 point of the original OPD

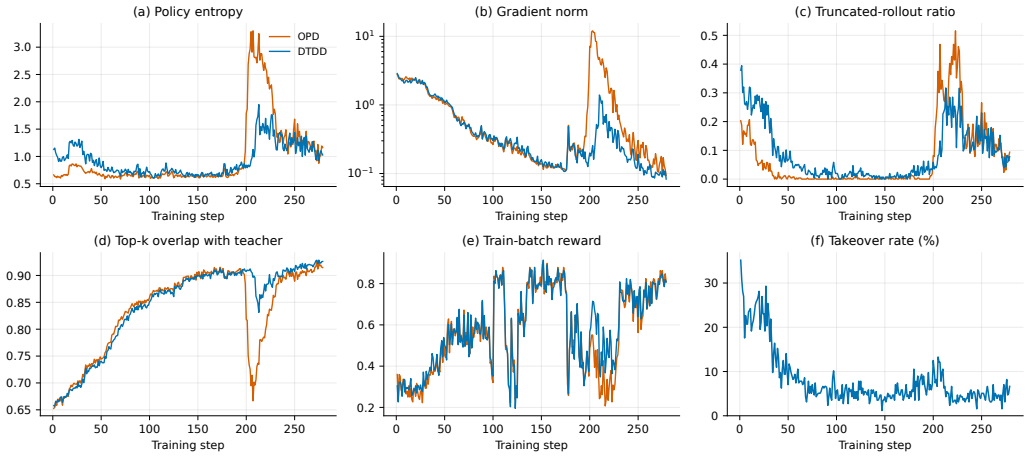

Figure 9: Complete 15K training dynamics. The OPD gradient and entropy spikes coincide with increasing truncation and decreasing teacher overlap. DTDD limits each component of this event while preserving held-out accuracy.

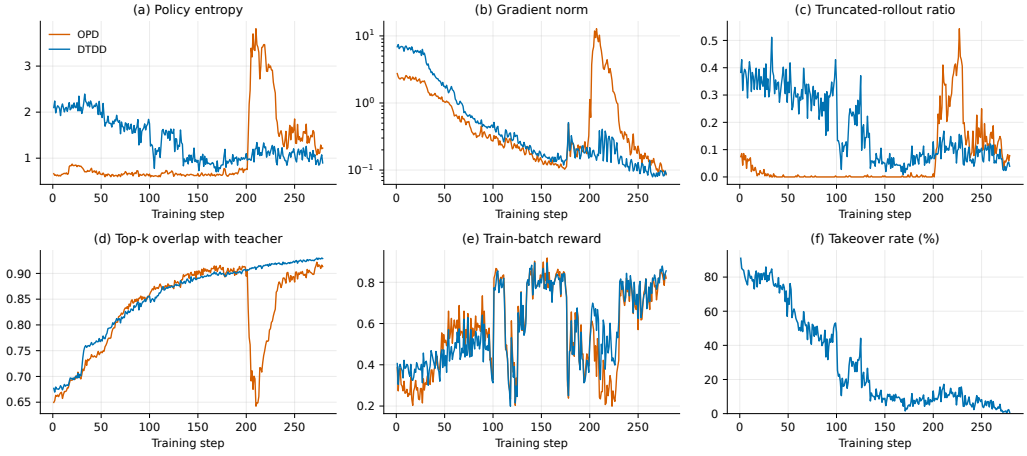

Figure 10: Complete 18K training dynamics. DTDD suppresses the late OPD instability in gradient norm, entropy, truncation, and teacher overlap. Its intervention rate decreases over training and remains responsive to the late discrepancy increase.

Table 9: Response-length summary at the last common checkpoint. Accuracy is average@16 over AIME 2024, AIME 2025, and AMC 2023. Gradient peaks are computed over updates greater than 50.

Budget	Accuracy (%)		Peak grad.		OPD / DTDD
	OPD	DTDD	OPD	DTDD	ratio
7K	54.4	**54.8**	0.54	0.96	0.56×
15K	**54.1**	54.0	12.02	1.40	8.6×
18K	**55.4**	54.6	12.81	1.84	7.0×

run at 15K and 1.8 points at 18K; the original and reproduced OPD runs themselves differ by 0.35 and 0.76 point. This control establishes that teacher takeover is not required to stabilize these two runs.

The shadow runs also log the trigger without applying it. At 18K, its would-trigger rate stays at 8–10% through update 198, rises to 25% at update 201, and the gradient spike begins at update

Table 10: Clipped-OPD control at update 279. Accuracy is held-out average@16 (%); gradient statistics exclude the first 50 updates.

Budget	OPD signal	Peak gradient	Updates > 2	Accuracy
15K	Unclipped reproduction	14.33	22	54.50
15K	Clamp $[-5, 5]$	12.48	21	55.26
15K	Clamp $[-1, 1]$	**0.95**	**0**	55.05
18K	Unclipped reproduction	12.42	21	54.62
18K	Clamp $[-5, 5]$	11.28	21	53.72
18K	Clamp $[-1, 1]$	**0.96**	**0**	53.54

Table 11: Base-initialization dose comparison on DAPO-Math-17K (average@16, %). Updates 120 and 240 provide common early and late checkpoints for all reported arms.

Method	Window C	Controller	Update 120	Update 240
OPD	–	–	14.65	22.29
DTDD rewind	256	fixed $\tau = 0.2$, length-gated (initial takeover $\approx 25\%$)	14.81	21.58
DTDD restart	–	target rate 45%	20.49	21.67
DTDD rewind	64	fixed $\tau = 0.475$ (initial takeover $\approx 90\%$)	**21.32**	21.04
DTDD rewind	128	fixed $\tau = 0.362$ (initial takeover $\approx 90\%$)	20.90	**22.40**
DTDD rewind	256	fixed $\tau = 0.296$ (initial takeover $\approx 90\%$)	21.23	21.39

204. Policy entropy leaves its preceding range at update 192. In this run, divergence reacts near the optimization event rather than providing an earlier warning than entropy.

Controller regimes. The three length-sweep arms share $\tau = 0.7$ and a 5% random exploration floor, but the divergence branch activates differently across budgets. At 7K, the realized takeover rate remains between 2% and 8%, making this run a no-harm control for sparse intervention. At 15K, takeover starts at 35%, recedes to the floor by update 100, and reactivates around the late instability. At 18K, takeover starts at 91% and remains divergence-driven for most of training before responding again to the late discrepancy increase. The 15K and 18K runs therefore test adaptive triggering, while 7K establishes the stable control.

Random-intervention ablation. The length controller includes a 5% random-intervention floor. We fork the 18K run at update 170 and remove this branch. Across six evaluated checkpoints, the matched held-out accuracy changes by only -0.05 point. The random floor therefore has negligible effect on student-only accuracy in this regime.

D Additional Base-Initialization Results

D.1 Dose Response on DAPO-Math-17K

Table 11 expands the DAPO-Math-17K result from Section 4.3. All arms use Qwen3-1.7B-Base, Qwen3-4B-Base-GRPO, and a 7,168-token generation budget. The three high-initial-coverage rewind controllers separate from OPD within 40–70 updates and reach the 21.3% reference level near update 70; OPD reaches the same level near update 180. At update 240, these arms span 21.04–22.40%, compared with 22.29% for OPD. The evidence therefore supports faster early learning rather than a higher final plateau. The length-gated fixed-threshold control begins near 25% takeover and tracks OPD early, while the 45% restart arm provides an intermediate dose.

The controller labels distinguish a rate-controlled target from an offline-calibrated initial rate. Under its fixed threshold and length gate, the low-initial-coverage arm rises from approximately 25% to 94% rollout-level takeover as rollouts lengthen. The high-coverage rewind arms use fixed thresholds calibrated only to start near 90%, whereas the 45% rate-controlled restart arm realizes approximately 37% after eligibility filtering. Thus, the configured label identifies the controller, while the training curves report its realized coverage.

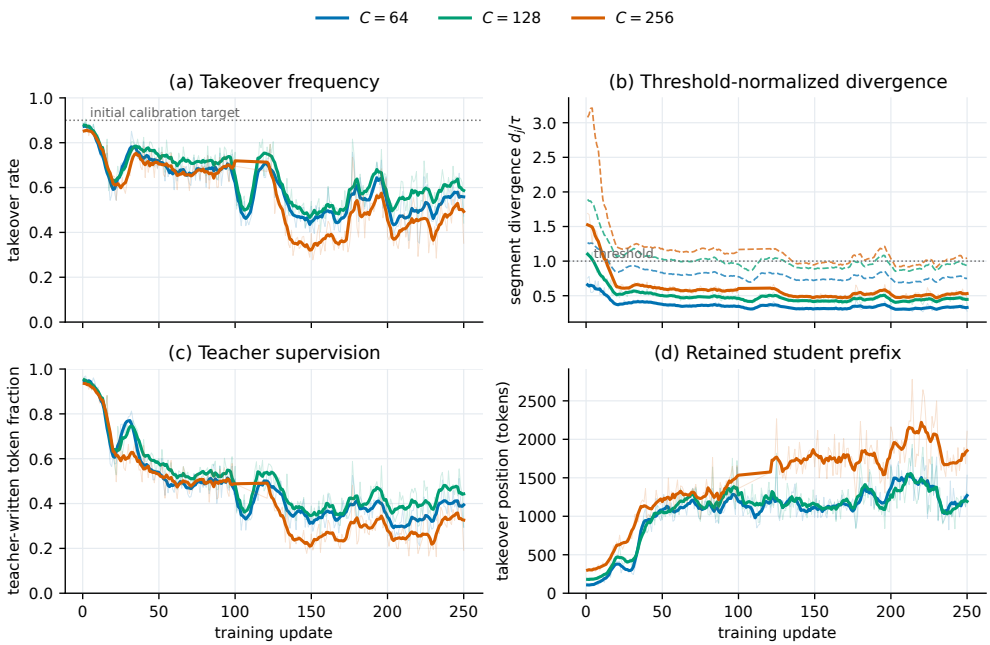

Figure 11: Controller dynamics for the three high-coverage DAPO-Math rewind arms. Faint curves show per-update values and solid curves show seven-update moving averages. **(a)** Rollout-level takeover rate; the dotted line is the initial calibration target. **(b)** Segment divergence normalized by each arm's fixed threshold (solid: mean; dashed: 90th percentile). **(c)** Fraction of generated training tokens written by the teacher. **(d)** Mean takeover position, conditional on triggering and converted from window index to token position.

D.2 Window Size and Intervention Form

Rewind remains effective across a fourfold window-size range. The $C \in \{64, 128, 256\}$ arms follow similar early trajectories and reach 21.0–22.4% late-checkpoint accuracy. Figure 11 compares their controller dynamics directly. All three begin near the calibrated 90% takeover rate. As segment discrepancy contracts, takeover frequency and the teacher-written token fraction decrease, while the remaining takeovers preserve longer student prefixes.

Restart supplies a complete teacher rollout after each trigger, while rewind preserves the student prefix and replaces the detected suffix. Both accelerate the base student at sufficient dose. Restart contributes more teacher tokens per triggered example; rewind retains more on-policy context.

D.3 Controller and Objective Controls

Table 12 separates the trigger, teacher-suffix objective, and intervention dose at fixed updates. It reports the individual OPD and DTDD seeds summarized in Table 1; the main-table standard deviations use the sample estimator with $n - 1$ in the denominator. The two DTDD seeds follow nearly identical trajectories: their mean accuracies over updates 60–120 are 21.0% and 20.9%.

The token-spike and unified-KD variants perform worse. In the segment-plus-unified-KD row, the reverse-KL surrogate is evaluated on the student's top-K support, which can exclude the token written by the teacher. This arm therefore measures support mismatch as well as the choice between a unified and role-split objective. The controllers also differ in realized intervention position and teacher-token dose (Table 13). We therefore interpret Table 12 as a comparison of complete controller designs, not an isolated causal estimate of trigger granularity or suffix objective. The token-spike ablations use a history-normalized z-score trigger. The AdaSwitch row instead uses its published rule: a token triggers when its divergence exceeds three times the mean over the preceding ten tokens. This AdaSwitch controller (Peng et al., 2025) performs better than the other token-spike variants but remains below role-split DTDD at all four reported checkpoints. In the

Table 12: Base-student controller and objective controls on DAPO-Math-17K (held-out average@16, %). A dash denotes an unscheduled evaluation; *TBD* denotes an evaluation in progress.

Trigger / method	Teacher-suffix objective	Optimizer step					
		20	40	60	80	100	120
OPD (seed 0)	–	8.58	12.67	12.36	14.01	13.47	14.65
Segment, DTDD (seed 0)	BC	12.83	18.82	20.49	21.60	20.90	20.90
OPD (seed 1)	–	–	10.40	13.32	13.99	13.89	14.43
Segment, DTDD (seed 1)	BC	12.47	19.65	20.16	21.34	21.25	20.76
Token spike	Unified KD	–	–	12.47	12.14	12.81	13.51
Token spike	BC	8.61	*TBD*	13.19	13.07	13.26	16.79
Segment	Unified KD	6.90	*TBD*	5.78	7.80	8.04	10.23
AdaSwitch ($K = 3$)	Unified KD	16.40	18.90	18.16	19.57	18.68	19.95

Table 13: Later-training controller and runtime measurements for the base-student controls. Trigger position is reported in each controller's native granularity. Runtime follows each completed job's allocation; GPU models are shown where recorded, and times are not hardware-normalized.

Controller / objective	Median trigger	Trigger rate	Teacher-written tokens/update	Time/update
DTDD (segment + BC)	segment 5	71%	0.52M	15.1 min (4×A100); 19.5 min (4×L40S)
Token spike + unified KD	token 91	79%	0.40M	10.2 min
Token spike + BC	token 88	82%	0.43M	13.0 min
Segment + unified KD	segment 1	80%	0.67M	14.2 min
AdaSwitch ($K = 3$)	token 14	100%	0.86M	31.3 min
OPD	–	–	0	5.4 min (2×A100)

Table 14: Fixed-update comparison of direct OPD, online takeover, and offline SFT warm start (held-out average@16, %). A dash denotes an unscheduled checkpoint.

Training route	Optimizer step								
	20	40	60	80	100	120	160	200	240
OPD from base	8.58	12.67	12.36	14.01	13.47	14.65	17.00	22.30	22.30
DTDD from base	12.83	18.82	20.49	21.60	20.90	20.90	21.00	22.00	22.40
SFT → OPD	21.55	21.27	21.56	22.81	22.37	22.03	22.30	21.70	21.80

later-training measurements summarized in Table 13, AdaSwitch triggers on every rollout and uses 0.86M teacher-written tokens per update, compared with 0.52M for DTDD.

Table 14 compares the same online runs with an offline SFT warm start followed by ordinary OPD. SFT→OPD reaches 21.55% at its first evaluation, update 20, and remains near 22% through update 240. Most of its observed gain is therefore present by the first evaluated OPD checkpoint. The two routes reach comparable accuracy; DTDD provides a single online stage whose teacher-token demand is determined during training.

The suffix objective shows limited sensitivity across a $30\times$ range of BC weights and under a soft cross-entropy target (Table 2 in Section 4.1). At updates 60 and 80, the variants span 20.31–22.01%. The heavier $\lambda_{\mathrm{BC}} = 3$ arm subsequently declines to 19.10% at update 120, whereas $\lambda_{\mathrm{BC}} = 0.1$ and the soft cross-entropy target reach 20.78% and 20.59%, respectively.

D.4 Independent OpenThoughts3 Setting

OpenThoughts3 provides an independent base-initialization setting with a Qwen3-4B non-thinking teacher. Scheduled restart, scheduled rewind, and fixed-threshold rewind all improve the late-checkpoint average over OPD (Table 15). The gain is shared across controller forms and intervention modes.

Table 15: Per-benchmark base-initialization results (average@16, %). Rows report the final common comparison checkpoint for each setting.

Method	AIME24	AIME25	AMC23	Average
Qwen3-1.7B-Base ← Qwen3-4B-Base-GRPO, DAPO-Math-17K, 7,168 tokens				
Teacher: Qwen3-4B-Base-GRPO	*24.17*	*18.12*	*64.22*	*35.50*
OPD	11.25	**9.38**	46.25	22.29
DTDD restart, target rate 45%	11.25	7.50	46.25	21.67
DTDD rewind, initial takeover $\approx 90\%$, $C = 128$	11.67	8.33	**47.19**	**22.40**
DTDD fixed τ=0.2, initial takeover $\approx 25\%$	**11.88**	7.71	45.16	21.58
Qwen3-1.7B-Base ← Qwen3-4B-Base (non-thinking), OpenThoughts3, 7,168 tokens				
Teacher: Qwen3-4B-Base	*21.04*	*20.62*	*67.66*	*36.44*
OPD	5.62	**5.42**	27.50	12.85
DTDD rewind, target rate $90\% \to 5\%$	**8.96**	4.58	29.53	14.36
DTDD restart, target rate $90\% \to 5\%$	8.75	5.21	**33.28**	**15.75**
DTDD fixed τ=0.87	6.88	**5.42**	31.25	14.51

Table 16: Serial cost to comparable observed checkpoints. Training time sums logged step times. Stage times are rounded; totals use unrounded measurements. Accelerator types differ across routes.

Route	Measured stages	Checkpoint (accuracy)	Serial time	GPU-h	Teacher-written tokens
Online DTDD (seed 1)	online training (22.0 h)	Step 80 (21.34)	22.0 h	88	43M
SFT → OPD	teacher data (26.9 h) + SFT (11.7 h) + merge (0.1 h) + OPD (4.0 h)	Step 20 (21.55)	≈42.8 h	≈66	95M

The fixed-threshold arm additionally demonstrates autonomous withdrawal. With $\tau = 0.87$, its takeover rate moves from 88.3% at the first update to 2.3% at update 100 and zero by update 150. The scheduled arms instead follow a prescribed 90%-to-5% linear controller.

D.5 Update and Wall-Clock Cost

Against OPD trained directly from the same base model, high-dose DTDD reaches the 21.3% reference level in approximately $2.6\times$ fewer optimizer updates. This reduction does not translate into lower wall-clock time than direct OPD under the current implementations: later-training updates take 5.4 minutes for OPD on two A100 GPUs, compared with 15.1 minutes for DTDD on four A100 GPUs and 19.5 minutes on four L40S GPUs. These measurements use different resource allocations and are reported as observed rather than hardware-normalized costs.

The offline SFT baseline answers a different operational question. Its teacher-data generation (26.9 hours), SFT (11.7 hours), checkpoint merge (0.1 hour), and first 20 OPD updates (4.0 hours) form a serial pipeline that takes approximately 42.8 hours before the first evaluated checkpoint reaches 21.55%; online DTDD reaches 21.34% at update 80 after 22.0 hours, a $1.9\times$ reduction in end-to-end latency. The corresponding teacher-written token counts are 95M and 43M, a 55% reduction. Measured accelerator use is approximately 66 versus 88 GPU-hours. Generation, SFT, and checkpoint merging use completed-job elapsed times; OPD and DTDD use cumulative logged step times through updates 20 and 80, respectively. These costs exclude queue delays and failed launches. Totals use unrounded measurements. These wall-clock values use the hardware assigned to each completed job—A40 and A100 GPUs for the SFT route, and four L40S GPUs for DTDD—and are not normalized across device types. The comparison supports a shorter single-stage pipeline and lower demand for teacher-written tokens; it is not a hardware-normalized or token-budget-matched ranking of algorithmic efficiency. Unlike an offline corpus whose size must be chosen before SFT, DTDD requests teacher continuations only when its online controller triggers.

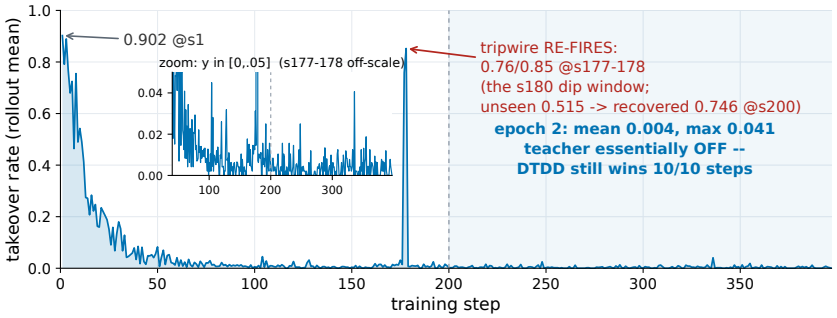

Figure 12: ALFWorld action-level takeover (Qwen2.5-3B-Instruct ← GiGPO-7B). The controller withdraws as discrepancy contracts and reactivates around the update-180 disruption.

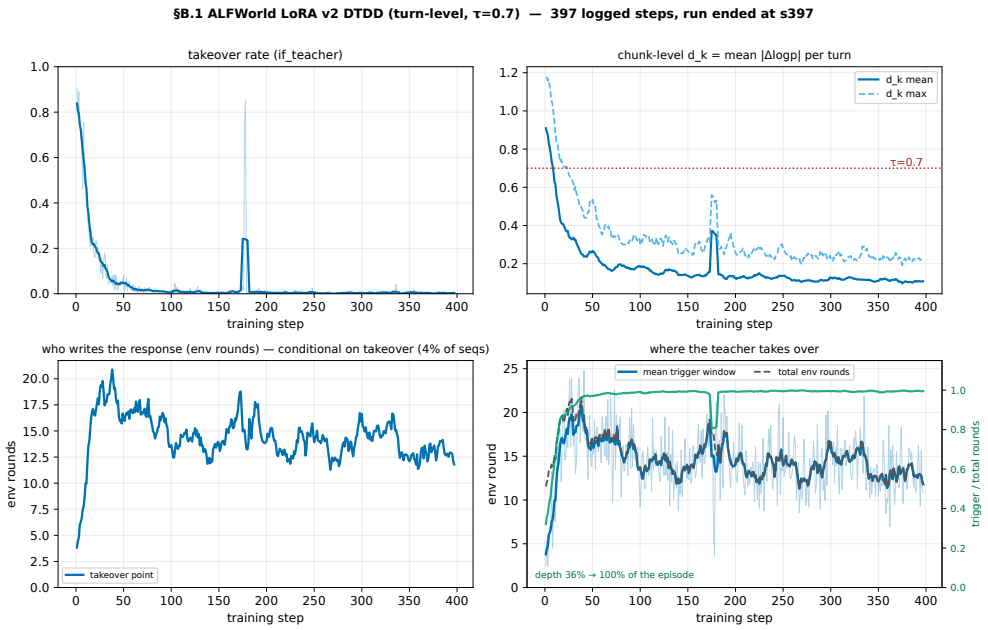

Figure 13: ALFWorld action-level controller dynamics. Top: takeover rate and action discrepancy. Bottom: intervention turn and its depth within the episode, conditional on takeover. The student is Qwen2.5-3B-Instruct, the teacher is GiGPO-7B, and the maximum episode length is 30 turns.

E Additional Agentic Results

E.1 ALFWorld Action-Level Intervention

Figure 12 shows the action-level takeover rate for the ALFWorld LoRA run. Takeover rapidly becomes sparse, briefly reactivates around the update-180 disruption, and returns below 5% thereafter.

Figure 13 decomposes the same run. Action discrepancy and takeover rate contract together during the first 50 updates. Trigger position and episode depth show where the remaining sparse interventions occur. The controller therefore withdraws and reactivates within one run without a temporal schedule.

At update 400, LoRA improves average success from 70.06% to 77.59%, driven primarily by the seen split. Full-model training raises OPD to 79.50% and narrows the DTDD advantage to about 0.4 point. Parameter-efficient adaptation therefore leaves a larger recoverable gap, while full-model OPD closes most of that gap on its own.

Appendix Table 6 reports the $+7.53$-point matched update-400 comparison; Table 3 reports the $+4.83$-point best-observed gap. Across the last five matched checkpoints, DTDD leads by 7.22

Table 17: ALFWorld success rate (%) at update 400, the last common checkpoint. The LoRA comparison matches Appendix Table 6; full-model training provides a capacity control.

Method	Seen	Unseen	Average
Qwen2.5-3B-Instruct ← GiGPO-7B, LoRA			
OPD	70.71	69.40	70.06
DTDD	**84.29**	**70.90**	**77.59**
Qwen2.5-3B-Instruct ← GiGPO-7B, full fine-tuning			
OPD	**82.14**	76.87	79.50
DTDD	**82.14**	**77.61**	**79.88**

Table 18: Agentic code execution accuracy (average@8, %) at update 130. Held-out averages the two in-distribution splits; LCB averages LiveCodeBench v1 and v6.

Method	Dose	Held-out	LCB	Avg.
Qwen3-1.7B-SFT ← DemyAgent-GRPO-4B, code interaction				
Teacher	–	63.16	58.11	60.63
OPD	0%	25.60	**42.26**	33.93
DTDD ($\tau = 0.50$)	2.8%	**31.50**	38.43	**34.96**
DTDD ($\tau = 0.26$)	∼50%	19.78	36.48	28.13

points on average. The OPD arm varies by 10.2 points over the same window, so repeated runs are needed to separate the intervention effect from late-training variation.

E.2 Agentic Code Dose

The code study distills Qwen3-1.7B-SFT from DemyAgent-GRPO-4B. At the shared update-130 endpoint, takeover with $\tau = 0.50$ remains 1.03 points above OPD in aggregate because of gains on the two in-distribution splits, although both LiveCodeBench scores trail OPD (Appendix Table 6). The best-observed checkpoints in Table 3 capture the earlier period in which sparse takeover leads on all four evaluation splits.

The two thresholds deliver 2.8% and approximately 50% teacher-token dose. Increasing the dose to approximately 50% reverses the gain over OPD, reducing aggregate accuracy from 33.93% to 28.13% at update 130. The teacher itself reaches 60.63% on the same aggregate evaluation, so teacher quality leaves ample headroom. The contrasting outcome points to excessive replacement of a warm-start student's on-policy interaction.

F Boundary Cases and Controller Interpretation

The boundary analysis uses headroom as an empirical diagnostic, then examines teacher validity at the intervention state, student learnability, and the amount of teacher behavior introduced by the controller. Table 19 collects the settings that isolate these factors.

Headroom analysis. Figure 6 in Section 4.5 uses $h = 1 - \mathrm{OPD}_{\mathrm{end}}/\mathrm{teacher}$, with both scores measured on the same evaluation harness. Negative h indicates that OPD exceeds the teacher. Each vertical coordinate summarizes the DTDD–OPD difference averaged over the matched checkpoints available for that arm. This differs from the best-observed protocol of Table 3: the ALFWorld LoRA arm, for example, has 19.4% headroom and a $+1.11$-point matched-checkpoint mean here, against a $+4.83$-point best-observed gap in Table 3. The analysis is descriptive: arms that share a setting are not independent, and the 10% boundary summarizes the observed split rather than defining a fitted decision threshold. The split summarizes the observed sign rather than the magnitude. Six DAPO-Math arms share one headroom value but range from $+0.11$ to $+4.42$ points as dose and intervention form change.

Four arms fall outside the common split. ALFWorld full fine-tuning averages a 5.25-point gain across 20 matched checkpoints but finishes only about 0.4 point above OPD, indicating transient acceleration. Both axes for the code points use LiveCodeBench v1. The $\tau = 0.26$ arm loses 7.48

Table 19: Boundary cases. Scores use the native evaluation metric of each environment. The final column gives the controller response suggested by the observed regime.

Condition	Setting	Observation	Result	Action
Little remaining gap	7K/15K/18K distilled math student	OPD finishes within 5–9% of teacher	Stability gain; accuracy tie	Adaptive repair
Valid, learnable gap	Base math student	OPD learns slowly; teacher continuation is learnable	High dose reaches plateau earlier	High dose
Warm, healthy student	Agentic code, $\rho_T = 2.8\%$	Small in-distribution gain	$+1.03$ points at update 130	Sparse repair
Excessive dose	Agentic code, $\rho_T \approx 50\%$	Teacher behavior replaces on-policy interaction	-5.80 aggregate points at update 130	Reduce dose
Invalid teacher	Rebuilt WebShop	Teacher 18.7% $<$ OPD 19.7%	DTDD 9.1%	Abstain
Limited learnability	ALFWorld, 0.5B $\leftarrow$ 7B	Matched mean: OPD 0.79%, DTDD 0.96%	$+0.17$ points; both remain near zero	Abstain

points when its LiveCodeBench-v1 gap is averaged over 12 matched checkpoints. The -5.80-point result in Table 18 instead compares OPD and DTDD at their shared update-130 endpoint after averaging HO-200 full, HO-200 subset, LiveCodeBench v1, and LiveCodeBench v6. The same arm has an 85.4% rollout-level takeover rate (approximately 50% teacher-token dose), showing that excessive replacement can erase otherwise useful headroom. The "SFT student" point is Qwen3-1.7B-SFT $\leftarrow$ Qwen3-4B-Base-GRPO on DAPO-Math-17K; the "stable-OPD lane" is DeepSeek-R1-Distill-Qwen-1.5B $\leftarrow$ DeepSeek-R1-Distill-Qwen-7B on DAPO-Math-17K. Their matched-checkpoint means differ from OPD by -0.36 and -0.08 point, respectively.

Teacher validity. The rebuilt WebShop environment uses a 120K-product catalog that retains all goal-bearing products and adds random distractors. It therefore differs from the catalog and observation format on which the WebShop-specific GiGPO-7B teacher was trained. On this environment, that teacher reaches 18.7% success and OPD reaches 19.7%. This negative-headroom regime yields a DTDD score of 9.1%: takeover routes training toward a teacher that is weaker on the current environment.

Student learnability. The small-student ALFWorld experiment pairs Qwen2.5-0.5B-Instruct with the ALFWorld-specific GiGPO-7B teacher, creating a roughly $14\times$ parameter gap. Across the six matched evaluations from updates 20 through 120, OPD and DTDD average 0.79% and 0.96% success, respectively. At update 120, the scores are 3.65% and 1.80%; the additional DTDD evaluation at update 140 reaches 4.71%. The teacher supplies valid behavior, but takeover does not produce reliable student-only success under this training configuration.

Available headroom and dose. The student in the rollout-length study already reaches within 5–9% of its teacher under OPD, so takeover improves stability without raising endpoint accuracy. The base math student retains more headroom, and high-dose continuation accelerates learning. The warm-start code setting isolates dose within one student–teacher pair: increasing teacher-token dose from 2.8% to approximately 50% changes an aggregate gain into a 5.8-point loss.

Independent comparison. As an external check, applying the same headroom calculation to the main TCOD results (Wang et al., 2026) separates its two student scales. The 1.7B OPD baselines retain over 99% headroom, and all 18 takeover cells improve by 9.2–24.7 points. The 4B baselines finish within 6.9–13.4% of their teachers; mean gains fall to 0.3–1.5 points, with four negative cells. This independent result matches the regime split in Figure 6.