PASS THE BATON: TRAJECTORY-RELAYED ON-POLICY DISTILLATION

Haolei Xu[1,2]*, **Xiaowen Xu**[2]*, **Haiwen Hong**[2*†], **Zixuan Ni**[1],
Hongxing Li[1,2], **Yiwen Qiu**[1], **Weiming Lu**[1‡], **Yongliang Shen**[1]

[1]Zhejiang University, [2]Yuvion Team, Alibaba Group
{xuhaolei, luwm, syl}@zju.edu.cn honghaiwen.hhw@alibaba-inc.com

GitHub: https://github.com/zju-real/Relay-OPD
Project: https://zju-real.github.io/Relay-OPD

ABSTRACT

On-policy distillation (OPD) grounds token-level supervision in the student's own trajectory, yet suffers from prefix failure: once the student commits to a wrong reasoning direction, all subsequent generation builds on this deviation, producing misdirected continuations that elicit unreliable supervision and waste compute. We identify a teacher–student continuation asymmetry on failed prefixes, where the teacher tends to redirect while the student continues along the original direction, and convert it into a label-free handoff trigger in Relay On-Policy Distillation (Relay-OPD). During training, Relay-OPD constructs relay trajectories by letting the teacher briefly take over at detected trigger points to produce a teacher leg, after which the student resumes and is optimized on the resulting trajectory. A limited relay budget concentrates intervention on critical early positions while limiting departure from the student policy. With a Qwen3-4B-Instruct-2507 teacher and Qwen3-0.6B/1.7B-Non-Thinking students on eight mathematical reasoning benchmarks, Relay-OPD achieves the best or second-best results on every benchmark, outperforming standard OPD by +5.73% and the strongest baseline FastOPD by +1.49% on average for 1.7B, with consistent gains at 0.6B. Training trajectory length is reduced by over 50%.

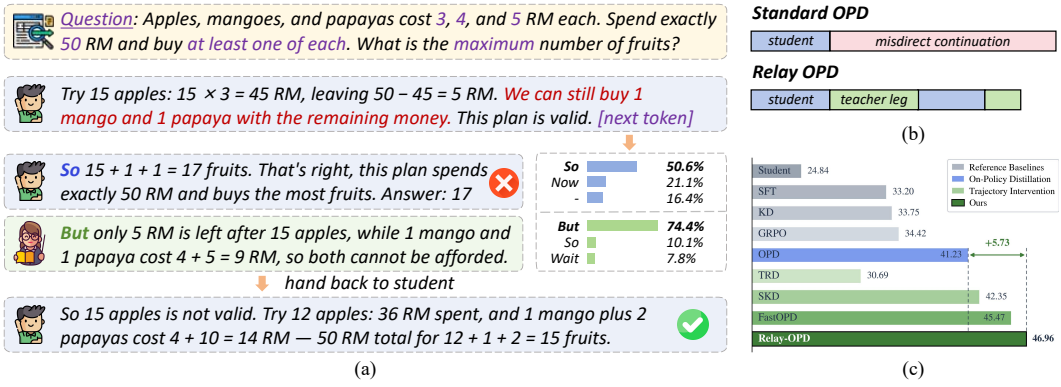


Figure 1: **(a)** A Relay-OPD case: at the detected handoff trigger, the teacher prefers the reflection token (`But`, 74.4%) while the student would continue along the current direction (`So`, 50.6%); a brief teacher leg corrects the reasoning and the student resumes to the correct answer. **(b)** Difference between OPD and Relay-OPD. **(c)** Overall performance.

* Equal contribution.
† Project leader.
‡ Corresponding author.

1 INTRODUCTION

Effectively transferring capabilities from strong models to weaker ones has become a central challenge in large language model post-training (Yang et al., 2025a; Xu et al., 2026; Zeng et al., 2026; Xiao et al., 2026). Unlike supervised fine-tuning and offline knowledge distillation (Hinton et al., 2015; Kim & Rush, 2016), which both rely on teacher-generated data, on-policy distillation (OPD) (Agarwal et al., 2024; Lu & Lab, 2025) lets the student generate trajectories from its own policy. The teacher then supplies dense token-level guidance on the prefixes the student actually visits. Because supervision is always grounded in the student's own state distribution, OPD effectively mitigates the train–inference distribution shift, and has shown clear gains in strong-to-weak distillation.

Yet the student's own trajectories inevitably include its failures. In long-chain reasoning (Jaech et al., 2024; Guo et al., 2025), this leads to **prefix failure** (Li et al., 2026; Fu et al., 2026; Xie et al., 2026): once the student commits to a wrong direction early on, all subsequent generation builds on this deviation (Figure 1(b)). The resulting long misdirected continuations elicit unreliable and potentially harmful supervision, and they waste substantial training compute.

Prior work addresses prefix failure from several angles, each with a structural limitation. Fixed-length truncation (ESR, FastOPD) (Ziheng et al., 2026; Zhang et al., 2026) cuts the rollout at a rigid position regardless of where the reasoning actually fails. Offline rewriting (TRD) (Jiang et al., 2026) repairs trajectories only after the rollout completes, and the repairs often leave visible artifacts. Token-level mixing (SKD) (Xu et al., 2025) switches between teacher and student on generic distributional disagreement, not on an explicit signal that the reasoning direction has failed. What is missing is a mechanism that acts *online*, correcting failure as it emerges, and that decides *where* to intervene from the reasoning state itself.

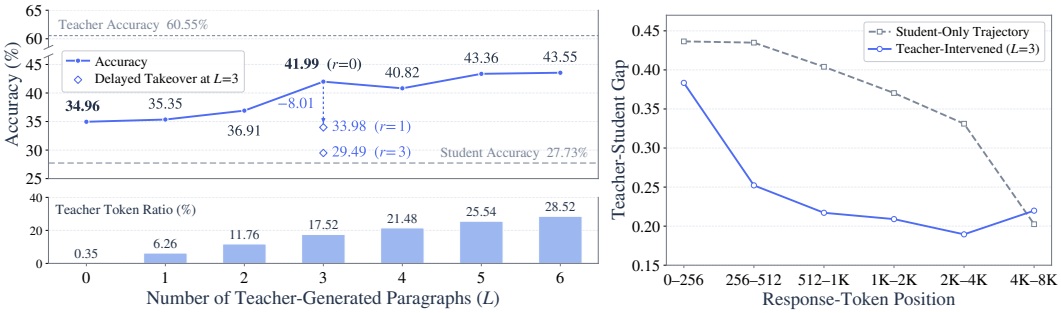
On a failed student prefix, the teacher and student diverge in how they continue (Figure 1(a); the complete case appears in Appendix E.1): the teacher tends to stop, re-examine the reasoning, and redirect, whereas the student tends to press on in the same wrong direction. This divergence is observable during generation, without any label, verifier, or reward model, and it marks where the student's reasoning has gone wrong. We call such a position a **handoff trigger**. To find out how a trigger should be handled, we run trajectory intervention experiments: at each trigger the teacher takes over briefly, and we vary the duration and timing of the intervention.

Two properties of this intervention stand out (Figure 2). First, correction can be remarkably *local*: replacing only the single reflection token at each trigger, so that teacher tokens are merely 0.35% of all generated tokens, already lifts accuracy from 27.73 to 34.96 (+7.23%; Figure 2(a)), and teacher takeover likewise reduces the teacher–student gap (Figure 2(b)). Second, the intervention's value is *front-loaded*: holding the intervention length fixed but shifting it to later triggers instead of the earliest ones drops accuracy from 41.99 to 33.98 and then 29.49. The teacher–student gap explains why timing dominates. It narrows as generation proceeds, whether the student runs alone or the teacher finishes the trajectory. As the prefix grows, the teacher is pulled along by the student's context, so a late takeover cannot redirect it.

A clear design principle emerges: prefix failure can be addressed through early, local teacher intervention at detected failure points, before the wrong direction becomes entrenched in the trajectory. This leaves two questions for a training method: *when* should the teacher take over, and how do we keep these takeovers from pulling the trajectory too far from the student's own policy?

We propose **Relay On-Policy Distillation (Relay-OPD)**, which interleaves student generation with brief teacher takeovers at the points where reasoning first goes wrong. As the student rolls out its trajectory, Relay-OPD monitors each prefix for a handoff trigger: a state where the teacher would redirect the reasoning while the student would continue on its current course. When a trigger fires, the teacher generates a short *teacher leg* that steers the reasoning back and returns control to the student (Figure 1(b)). A *relay budget* caps the number and length of these legs, concentrating correction at the early positions where prefix failure originates while keeping the trajectory close to the student's own policy. The student is then distilled on the resulting relay trajectory, exactly as in standard on-policy distillation.

Our contributions are as follows: (1) We identify a **teacher–student continuation asymmetry** on failed reasoning prefixes and show, through trajectory intervention experiments, that correcting prefix failure needs only early and local teacher intervention, with teacher tokens as low as 0.35%.



(a) (b)

Figure 2: **Preliminary Trajectory Intervention Experiments.** Qwen3-4B-Instruct-2507 (teacher) and Qwen3-1.7B-Non-Thinking (student) on 128 DAPO-Math-17K English samples; the handoff criterion matches the main experiments. **(a)** Accuracy (mean@4) and teacher token ratio as the number of teacher-generated paragraphs L varies; a periodic delay of r skips r consecutive valid triggers before the next takeover. **(b)** Teacher–student gap by response-token position under the $L{=}3$ intervention, computed as the absolute value of the mean token-level advantage (defined in §3.1) within each position interval.

(2) We propose **Relay-OPD**, which turns this asymmetry into a label-free handoff trigger and a budgeted relay rollout, and unifies student and teacher generation in a single speculative-decoding engine. (3) Across eight math reasoning benchmarks and two student scales, Relay-OPD outperforms standard OPD by $+5.73\%$ and FastOPD by $+1.49\%$ for the 1.7B student, while cutting average training trajectory length by over 50%.

2 Related Work

2.1 On-Policy Distillation

On-policy distillation samples trajectories from the student, with the teacher scoring each token on the prefixes the student actually visits (Agarwal et al., 2024; Gu et al., 2024); compared with offline training on teacher-generated data (Hinton et al., 2015; Kim & Rush, 2016), it grounds supervision in the student's own state distribution, yields faster and stronger transfer, and is now standard in post-training pipelines (Yang et al., 2025a; Lu & Lab, 2025). Recent extensions cover distillation across model families (Patiño et al., 2025), black-box teachers without logit access (Ye et al., 2025), and self-distillation of a model's own in-context or privileged knowledge (Yang et al., 2025b; Hübotter et al., 2026; Shenfeld et al., 2026; Zhao et al., 2026; Penaloza et al., 2026).

2.2 Prefix Failure

Grounding supervision entirely in student trajectories also imports the student's failures. Distillation gains shrink when teacher and student reasoning patterns are incompatible (Li et al., 2026), and as student prefixes drift from teacher-supported states, subsequent supervision grows unreliable (Fu et al., 2026; Xie et al., 2026). These effects are most pronounced in long-chain reasoning, where early directional deviations compound autoregressively into extended misdirected continuations. Existing remedies intervene on the trajectory itself: ESR and FastOPD (Ziheng et al., 2026; Zhang et al., 2026) truncate rollouts at a fixed length to discard late, low-value supervision, TRD (Jiang et al., 2026) rewrites student trajectories offline via the teacher, and SKD (Xu et al., 2025) mixes teacher and student generation token by token according to distributional agreement.

3 Method

This section first reviews standard on-policy distillation (§3.1), then introduces the handoff trigger, relay trajectory construction, and optimization objective of Relay-OPD (§3.2), and finally describes its efficient implementation (§3.3).

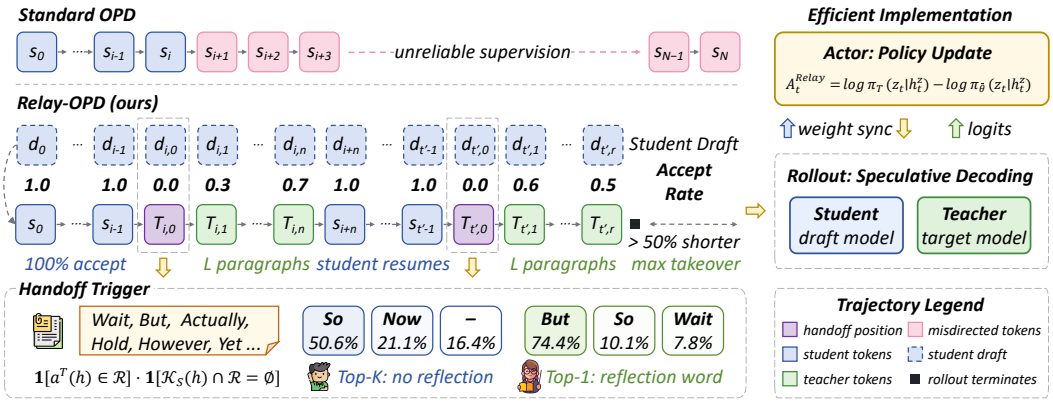


Figure 3: **Overview of Relay-OPD.** Unlike standard OPD (top), which trains on student-only trajectories that continue misdirected after prefix failure, Relay-OPD constructs relay trajectories (middle): when the handoff trigger (bottom) detects that the teacher would redirect the reasoning while the student would not, the teacher briefly takes over before the student resumes. The entire rollout runs in a single speculative decoding engine (right; §3.3).

3.1 On-Policy Distillation

Let $x \sim \mathcal{D}$ denote a training prompt, and let π_θ, $\pi_{\bar{\theta}}$, and π_T denote the current student policy, old student policy, and teacher policy. At each training iteration, standard OPD first generates an on-policy trajectory using the old student policy:

$$y = (y_1, \ldots, y_N) \sim \pi_{\bar{\theta}}(\cdot \mid x). \tag{1}$$

Let $h_t = (x, y_{<t})$ denote the student prefix at position t. Standard OPD typically employs the reverse KL divergence. Taking a single-sample estimate (Lu & Lab, 2025) at student-sampled token y_t:

$$\widehat{D}^{(t)}_{\mathrm{RKL}} = \log \pi_{\bar{\theta}}(y_t \mid h_t) - \log \pi_T(y_t \mid h_t). \tag{2}$$

Its negation serves as the advantage:

$$A^{\mathrm{OPD}}_t = -\widehat{D}^{(t)}_{\mathrm{RKL}} = \log \pi_T(y_t \mid h_t) - \log \pi_{\bar{\theta}}(y_t \mid h_t). \tag{3}$$

A positive advantage encourages the student to increase the probability of the sampled token; a negative one decreases it. Since training is entirely grounded in student-visited prefixes, once prefix failure occurs early in generation, the resulting long misdirected continuation is still included in training despite receiving increasingly unreliable supervision.

3.2 Relay-OPD

Relay-OPD introduces state-driven teacher takeover into the student trajectories of standard OPD. The method first detects handoff triggers based on teacher–student continuation tendencies on the current prefix, then constructs a relay trajectory comprising student legs and teacher legs, and performs distillation on the resulting trajectory. Figure 3 provides an overview.

Handoff trigger. Identifying prefix failure online requires a signal that depends on neither external verifiers nor process labels. Rather than relying on generic teacher–student distributional differences, we focus on divergences in reasoning direction: takeover is triggered when the teacher tends to redirect while the student tends to continue along the current direction (Figure 3, bottom).

Let $\mathcal{R} \subseteq \mathcal{V}$ denote a set of reflection tokens used for redirecting reasoning (Guo et al., 2025; Muennighoff et al., 2025), such as Wait, But, and However, together with their case and leading-space

variants; the complete list is provided in Appendix A.1. Given a student prefix h, define the teacher's most probable next token as

$$a^T(h) = \arg\max_{v \in \mathcal{V}} \pi_T(v \mid h), \tag{4}$$

and the student's top-K support set on this prefix as

$$\mathcal{K}_S(h) = \mathrm{TopK}_K(\pi_{\bar{\theta}}(\cdot \mid h)). \tag{5}$$

When the teacher's preferred token belongs to $\mathcal{R}$ while the student's top-K support set contains no token from $\mathcal{R}$, we define the handoff criterion:

$$\phi(h) = \mathbf{1}[a^T(h) \in \mathcal{R}] \cdot \mathbf{1}[\mathcal{K}_S(h) \cap \mathcal{R} = \varnothing]. \tag{6}$$

K controls the divergence threshold required for triggering and thus the sensitivity of the criterion.

Relay trajectory construction. Two findings from the preliminary experiments shape the design. First, the benefit of extending teacher takeover saturates (Figure 2(a)): from $L = 3$ to $L = 6$, accuracy plateaus around 41–44 despite the teacher token ratio rising from 17.52% to 28.52%, motivating a limited teacher leg. Second, valuable supervision concentrates early in the trajectory (Figure 2(b)), motivating a relay budget that caps the total number of takeovers and focuses intervention on earlier positions.

The relay budget (M, L) specifies the maximum number of teacher takeovers M and the number of additional paragraphs L generated after the reflection token in each teacher leg, with paragraphs delimited by `\n\n`. We measure teacher legs in paragraphs rather than a fixed number of tokens so that each leg ends at a structurally complete reasoning unit instead of breaking off mid-thought; in our setting a paragraph contains 23.2 tokens on average. The student generates the student leg autoregressively according to $\pi_{\bar{\theta}}$, computing $\phi(h)$ at each position. When $\phi(h) = 1$ and the takeover count has not reached M, the teacher takes over: $a^T(h)$ from the trigger criterion serves as the starting token of the teacher leg, and generation continues for L paragraphs. When $L = 0$, the segment consists solely of the reflection token. After the teacher leg, if budget remains, the student resumes generation from the extended prefix; when the M-th teacher leg ends, the current rollout terminates. This yields the relay trajectory $z = (z_1, \ldots, z_N)$.

Unlike fixed-length truncation, both the intervention and termination positions of Relay-OPD are determined by the current reasoning state.

Optimization objective. Figure 2(a) shows that although teacher takeover significantly reduces prefix failure, even at $L = 6$ with a teacher token ratio of 28.52%, accuracy reaches only 43.55 compared with 60.55 for the teacher generating independently. We therefore posit that the training signal the teacher provides on relay trajectories differs from its ideal supervision on its own trajectories. Accordingly, we adopt a reverse-KL-style single-sample objective that optimizes directly on the observed tokens in the relay trajectory, enabling the student to selectively absorb the teacher's corrective signals rather than fully fitting the teacher distribution via forward KL.

At position t, let $h_t^z = (x, z_{<t})$ denote the relay prefix. The advantage for the actually generated token z_t is

$$A_t^{\mathrm{Relay}} = \log \pi_T(z_t \mid h_t^z) - \log \pi_{\bar{\theta}}(z_t \mid h_t^z). \tag{7}$$

The update ratio of the current student relative to the old student is

$$\rho_t(\theta) = \frac{\pi_\theta(z_t \mid h_t^z)}{\pi_{\bar{\theta}}(z_t \mid h_t^z)}. \tag{8}$$

Relay-OPD optimizes:

$$\mathcal{L}_{\text{Relay}} = -\mathbb{E}_{x,\, z}\left[\frac{1}{N}\sum_{t=1}^{N}\min\left(\rho_t A_t^{\text{Relay}}, \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon)A_t^{\text{Relay}}\right)\right]. \tag{9}$$

This objective uses the actually generated tokens z_t across the entire relay trajectory. Teacher legs both provide corrected context for subsequent student legs and directly participate in optimization through their generated tokens.

When $\phi(h) \equiv 0$, the relay trajectory degenerates to a standard student trajectory; when $L = 0$, the teacher leg contains only the replacement token at the handoff trigger position, corresponding to minimal prefix correction. The complete algorithm is given in Appendix C.1.

3.3 Efficient Implementation

Relay-OPD requires continuously determining whether to trigger teacher takeover during generation and alternating between student and teacher legs. Maintaining two independent generation pipelines would necessitate repeated coordination between teacher and student models for takeover timing, plus switching generation engines at takeover and recovery points, introducing substantial scheduling and communication overhead. We instead unify the entire Relay-OPD generation process within a single speculative decoding engine (Leviathan et al., 2023; Chen et al., 2023): the student serves as draft model and the teacher as target model (Figure 3, right).

Relay as a state-switched decoding process. The engine produces the relay trajectory z of §3.2 one position at a time: at each position t, it first determines a decoding state $s_t \in \{\mathsf{S}, \mathsf{T}, \perp\}$ from the current prefix $h_t^z = (x, z_{<t})$, and then emits z_t in a student leg (S) or a teacher leg (T), or terminates ($\perp$). Let j_t denote the number of teacher takeovers and ℓ_t the number of completed paragraphs (delimiter `\n\n`) in the current teacher leg, both up to position t. Starting from $s_1 = \mathsf{S}$, the state for the next position is determined by

$$s_{t+1} = \begin{cases} \mathsf{T}, & \text{if } s_t = \mathsf{S},\ \phi(h_{t+1}^z) = 1,\ \text{and } j_t < M, \\ \mathsf{S}, & \text{if } s_t = \mathsf{T},\ \ell_t = L,\ \text{and } j_t < M, \\ \perp, & \text{if } s_t = \mathsf{T},\ \ell_t = L,\ \text{and } j_t = M, \\ s_t, & \text{otherwise}, \end{cases} \tag{10}$$

and additionally enters the absorbing state $\perp$ whenever z_t is the end-of-sequence token or the length limit is reached. The first case realizes the handoff trigger: a trigger detected on the prefix h_{t+1}^z switches the state to T, so the teacher takes over at position $t+1$. When $L = 0$, the exit condition $\ell_t = L$ already holds at the leg-initial position, so the teacher leg consists of exactly the reflection token, recovering the minimal-correction case of §3.2.

Unified token generation. Every position is generated by speculative decoding against a state-dependent target policy: $\pi_t^{\text{tgt}} = \pi_{\bar{\theta}}$ when $s_t = \mathsf{S}$, and $\pi_t^{\text{tgt}} = \pi_T$ when $s_t = \mathsf{T}$. In student legs every draft is accepted and generation reduces to ordinary student decoding, while teacher legs perform standard speculative decoding against the teacher. The single exception is the leg-initial position of each teacher leg, where the engine directly emits the trigger token $z_t = a^T(h_t^z)$ without verification. At every other position, the student drafts $a_t^S \sim \pi_{\bar{\theta}}(\cdot \mid h_t^z)$, which is accepted with probability

$$\alpha_t = \min\left(1,\ \frac{\pi_t^{\text{tgt}}(a_t^S \mid h_t^z)}{\pi_{\bar{\theta}}(a_t^S \mid h_t^z)}\right); \tag{11}$$

if an i.i.d. draw $u_t \sim \mathcal{U}(0,1)$ exceeds α_t, the draft is rejected and z_t is instead sampled from the residual distribution

$$q_t(v) = \frac{\left[\pi_t^{\text{tgt}}(v \mid h_t^z) - \pi_{\bar{\theta}}(v \mid h_t^z)\right]_+}{\sum_{w\in\mathcal{V}}\left[\pi_t^{\text{tgt}}(w \mid h_t^z) - \pi_{\bar{\theta}}(w \mid h_t^z)\right]_+}, \qquad [x]_+ = \max(x, 0). \tag{12}$$

In student legs, substituting $\pi_t^{\text{tgt}} = \pi_{\bar{\theta}}$ into Eq. equation 11 gives $\alpha_t \equiv 1$, so drafts are accepted unconditionally; in teacher legs, Eqs. equation 11–equation 12 are exactly standard speculative rejection sampling against π_T.

Exactness and efficiency. By the correctness of speculative sampling (Leviathan et al., 2023), every verified position satisfies $z_t \sim \pi_t^{\mathrm{tgt}}(\cdot \mid h_t^z)$. Hence, conditioned on the prefix and the deterministic leg-initial token, teacher legs are distributionally identical to the teacher continuing generation directly from the extended prefix, and student legs to ordinary student sampling: the single-engine implementation reproduces the two-model relay process of §3.2 exactly, without switching between two generation pipelines. Efficiency-wise, teacher legs batch-verify student drafts instead of decoding serially, and the teacher logits computed during verification simultaneously provide $a^T(h_t^z)$ and the trigger criterion $\phi(h_t^z)$ at no additional cost. In practice, Eq. equation 10 is applied during block verification: draft tokens following the first transition point within a block are discarded, so the per-token semantics above are preserved exactly.

After each parameter update, the latest student weights are synchronized to the draft model; teacher parameters remain frozen throughout training.

4 Experiments

4.1 Experimental Setup

Models, data, and benchmarks. We use Qwen3-4B-Instruct-2507 (Yang et al., 2025a) as teacher and Qwen3-0.6B-Non-Thinking and Qwen3-1.7B-Non-Thinking as students. Training data is the English subset of DAPO-Math-17K (Yu et al., 2026). Evaluation covers eight mathematical reasoning benchmarks: AIME 2024 (AI-MO, 2024a), AIME 2025, AIME 2026, MATH500 (Hendrycks et al., 2021; Lightman et al., 2024), AMC 2023 (AI-MO, 2024b), OlympiadBench (He et al., 2024), HMMT February 2026 (Balunović et al., 2025), and HMMT November 2025. At evaluation time, we set temperature to 1.0 and top-p to 1.0, with a maximum generation length of 32,768 tokens. Each problem in AIME, AMC, and HMMT is sampled 32 times; each problem in MATH500 and OlympiadBench is sampled 4 times. We report mean accuracy.

Implementation details. All methods are implemented on verl (Sheng et al., 2025) and vLLM 0.21.0 (Kwon et al., 2023), trained on 8 H100 GPUs. Online distillation methods train for 1 epoch with a maximum response length of 16,384 tokens. Trajectory sampling uses temperature 1.0 and top-$p = 1.0$. Relay-OPD sets $K = 5$ and $(M, L) = (2, 3)$, without using any external verifier, process supervision labels, or answer correctness labels. Complete hyperparameters are provided in Appendix A.2, and the training and inference prompt template in Appendix B.1.

Baselines. We compare Relay-OPD against three categories of methods.

1. **Reference baselines.** The untrained initial student, supervised fine-tuning (SFT), token-level knowledge distillation (KD) (Hinton et al., 2015), and the outcome-reward reinforcement learning method GRPO (Shao et al., 2024).
2. **On-policy distillation.** Standard OPD (Lu & Lab, 2025), which performs token-level distillation on student-generated trajectories.
3. **Trajectory intervention methods.** TRD (Jiang et al., 2026), which rewrites student trajectories offline via the teacher; FastOPD (Ziheng et al., 2026; Zhang et al., 2026), which shortens the rollout budget; and SKD (Xu et al., 2025), which mixes teacher and student generation through speculative decoding. For FastOPD, we evaluate fixed truncation lengths in $\{1024, 2048, 4096, 8192\}$; Table 1 reports the best configuration at 4,096 tokens, with complete results in Appendix D.2.

Implementation details for each method are provided in Appendix C, with a structured overview in Table 5.

4.2 Main Results

Overall performance. As summarized in Table 1, Relay-OPD achieves the best or second-best results across all eight benchmarks for both student models. For Qwen3-1.7B-Non-Thinking, it attains an average accuracy of 46.96, outperforming standard OPD by +5.73% and the strongest trajectory intervention baseline FastOPD by +1.49%; AIME 2025 and AIME 2026 improve over OPD by +7.29% and +7.19%. Qwen3-0.6B-Non-Thinking exhibits a consistent trend: average accuracy

Table 1: **Main Experimental Results.** Mean accuracy is reported; bold and underline denote the best and second-best results for each student model. Subscripts in the Avg column indicate the training step of the best checkpoint. Train Len denotes the average rollout response length from the start of training to the best checkpoint.

Method	AIME24	AIME25	AIME26	MATH	AMC23	Olymp.	HMMT Feb26	HMMT Nov25	Avg	Train Len
Teacher	60.42	46.04	52.19	94.20	93.83	70.62	31.25	41.88	61.30	—
Student: Qwen3-0.6B-Non-Thinking										
Student	1.77	2.40	0.73	44.10	24.45	16.36	0.76	3.85	11.80	—
SFT	4.90	7.60	4.06	59.45	34.92	26.74	1.89	3.02	$17.82_{@110}$	4262
KD	4.17	7.19	4.79	57.75	35.23	27.15	2.94	2.71	$17.74_{@110}$	4262
GRPO	8.23	15.21	10.00	68.60	46.56	35.42	7.86	6.04	$24.74_{@110}$	3379
OPD	13.44	17.92	11.98	75.30	51.25	41.32	7.39	5.62	$28.03_{@110}$	6900
TRD	4.58	8.54	3.96	57.60	36.33	26.93	3.79	2.81	$18.07_{@20}$	3275
FastOPD	15.83	20.10	13.54	75.30	53.67	44.81	**11.55**	8.54	$30.42_{@100}$	3302
SKD	8.44	16.98	9.79	66.65	43.67	35.76	8.14	5.62	$24.38_{@20}$	5800
Relay-OPD	**15.94**	**20.94**	**14.06**	**76.80**	**55.55**	**45.03**	11.17	**8.85**	$\mathbf{31.04}_{@75}$	**2490**
Δ vs OPD	+2.50	+3.02	+2.08	+1.50	+4.30	+3.71	+3.78	+3.23	+3.01	-63.9%
Student: Qwen3-1.7B-Non-Thinking										
Student	12.60	9.58	7.40	71.95	47.89	38.54	6.34	4.38	24.84	—
SFT	23.33	19.48	16.15	81.40	59.45	46.62	12.59	6.56	$33.20_{@110}$	4262
KD	23.54	21.15	15.31	81.45	60.23	48.07	12.78	7.50	$33.75_{@110}$	4262
GRPO	24.58	22.08	15.62	80.35	60.16	48.74	14.49	9.38	$34.42_{@105}$	2558
OPD	35.83	25.52	23.33	85.70	70.08	55.27	20.08	14.06	$41.23_{@55}$	4658
TRD	19.27	19.69	12.71	77.70	55.47	44.18	11.93	4.58	$30.69_{@40}$	2785
FastOPD	42.29	30.42	26.35	87.95	74.30	58.16	23.58	**20.73**	$45.47_{@45}$	2709
SKD	33.12	30.73	28.85	87.35	72.42	54.41	20.08	11.88	$42.35_{@35}$	4753
Relay-OPD	**42.71**	**32.81**	**30.52**	**89.50**	**76.88**	**58.79**	**24.72**	19.79	$\mathbf{46.96}_{@35}$	**2296**
Δ vs OPD	+6.88	+7.29	+7.19	+3.80	+6.80	+3.52	+4.64	+5.73	+5.73	-50.7%

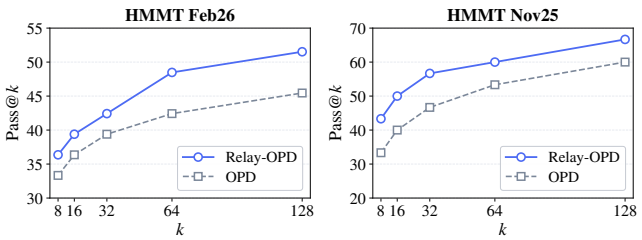


Figure 4: Pass@k performance of Relay-OPD and OPD on HMMT Feb26 and HMMT Nov25.

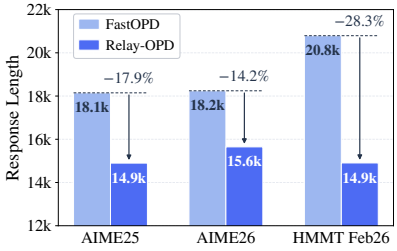


Figure 5: Inference response length of Relay-OPD vs. FastOPD.

improves by +3.01% over OPD and surpasses FastOPD by +0.62%. Relay-OPD further achieves significantly higher pass@k than standard OPD across different sampling budgets (Figure 4).

Comparison with trajectory intervention baselines. TRD exhibits marked performance degradation relative to standard OPD on both student models: 30.69 vs. 41.23 on 1.7B and 18.07 vs. 28.03 on 0.6B. We observe that its rewritten trajectories often carry visible rewriting artifacts rather than resembling naturally unfolding problem-solving; examples appear in Appendix E.2. SKD only marginally outperforms OPD on 1.7B (42.35 vs. 41.23) and drops to 24.38 on 0.6B; it struggles to break established repetitive generation patterns (Appendix E.3). FastOPD concentrates training signals at the sequence front via fixed truncation but cannot provide demonstrations of how to recover from failed prefixes. In contrast, Figure 5 shows that Relay-OPD reduces mean response length by 17.9%, 14.2%, and 28.3% on AIME 2025, AIME 2026, and HMMT February 2026, respectively, compared to FastOPD, while simultaneously improving accuracy by +2.39%, +4.17%, and +1.14%.

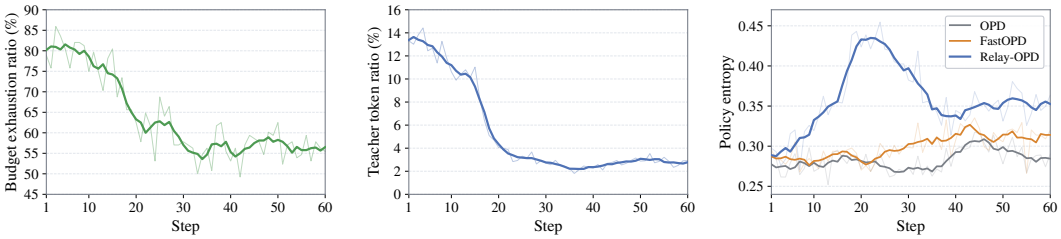


Figure 6: **Training Dynamics of Relay-OPD.** Using Qwen3-4B-Instruct-2507 as teacher and Qwen3-1.7B-Non-Thinking as student. **Left:** fraction of trajectories exhausting the relay budget; **Middle:** teacher token ratio relative to all effective response tokens; **Right:** policy entropy of Relay-OPD, OPD, and FastOPD.

Relay-OPD teaches the student to redirect reasoning before deviations compound further, yielding more accurate answers with shorter reasoning processes.

Training token efficiency. Relay-OPD uses shorter training trajectories and reaches the best checkpoint in fewer update steps. For the 1.7B student, Relay-OPD reaches its optimum at step 35, earlier than OPD at step 55 and FastOPD at step 45; its average rollout response length is 2,296 tokens, a 50.7% reduction from OPD's 4,658 and shorter than FastOPD's 2,709. A consistent trend holds for the 0.6B student: Relay-OPD's average training trajectory length is 2,490 tokens, a 63.9% reduction from OPD.

4.3 Training Dynamics

To elucidate how Relay-OPD adjusts teacher intervention as the student policy evolves, we track three metrics over the first 60 training steps in Figure 6. The left panel shows that the fraction of trajectories exhausting the relay budget gradually decreases, from roughly 75%–85% early on to roughly 50%–60%. The middle panel shows that teacher token ratio starts at roughly 13%, then drops rapidly and stabilizes at 2%–3% after about 20 steps. Both trends indicate that the teacher–student gap and the prefix failures requiring intervention shrink during training.

The right panel compares policy entropy across the three methods. Relay-OPD maintains consistently higher entropy than OPD and FastOPD over the first 60 steps. We attribute this to teacher intervention altering the prefixes the student visits and the subsequent generation directions, thereby increasing student exploration.

4.4 Ablation Studies

Role of the teacher leg. The benefit of the teacher leg stems from corrected context and local reasoning demonstrations, not merely early termination. To isolate this contribution, we set the maximum number of takeovers to $M = 1$ for both variants: Trigger-stop terminates immediately at the first trigger without generating any teacher tokens; Relay-OPD generates a teacher leg of length $L = 3$ before terminating.

As shown in Table 2, adding the teacher leg improves average accuracy from 43.48 to 46.25 (+2.77% over Trigger-stop), confirming that corrected context and local reasoning demonstrations yield benefits beyond dynamic truncation alone.

Training objective for the teacher leg. We compare two alternative objectives for the teacher leg, keeping the student leg under the standard OPD objective throughout. To support this ablation, the engine (§3.3) saves position markers for teacher legs together with the student's draft token a_t^S on prefix h_t^z at each teacher-leg position, before the teacher verifies or replaces it. The Student draft token variant uses the engine-saved a_t^S and replaces the default advantage with

$$A_t^{\text{draft}} = \log \pi_T(a_t^S \mid h_t^z) - \log \pi_{\bar{\theta}}(a_t^S \mid h_t^z). \tag{13}$$

That is, the policy gradient loss and parameter update at teacher-leg positions are computed with respect to the student's draft token a_t^S rather than the actually generated relay token z_t.

Table 2: **Teacher-Leg Ablations.** Using Qwen3-1.7B-Non-Thinking as student; all other settings are identical to the main experiments. **Top:** teacher leg vs. trigger-stop, where both variants set $M = 1$. **Bottom:** training objective for the teacher leg.

Variant	AIME24	AIME25	AIME26	MATH	AMC23	Olymp.	HMMT Feb26	HMMT Nov25	Avg
Teacher leg vs. trigger-stop									
Trigger-Stop ($M = 1$, No Teacher Leg)	39.48	28.75	24.48	87.00	71.56	55.27	23.11	18.23	43.48
Relay-OPD ($M = 1$, $L = 3$)	**42.40**	**31.98**	**29.69**	**88.55**	**75.47**	**58.38**	**23.30**	**20.21**	**46.25**
Training objective for the teacher leg									
Student Draft Token	38.54	27.60	26.25	88.10	75.94	57.42	23.67	18.96	44.56
Teacher FKL ($k = 128$)	41.77	28.96	26.56	88.20	72.73	55.68	21.69	17.08	44.08
Relay Token (Ours)	**42.71**	**32.81**	**30.52**	**89.50**	**76.88**	**58.79**	**24.72**	19.79	**46.96**

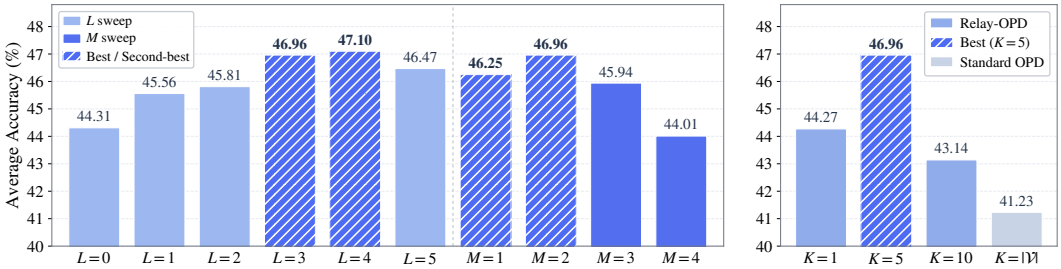


Figure 7: **Hyperparameter Sensitivity of Relay-OPD.** Using Qwen3-4B-Instruct-2507 as teacher and Qwen3-1.7B-Non-Thinking as student. **Left:** relay budget (M, L); **Right:** handoff top-K, where $|\mathcal{V}|$ denotes the vocabulary size.

The Teacher FKL variant takes the teacher's top-k support set $\mathcal{K}_t = \mathrm{TopK}_k(\pi_T(\cdot \mid h_t^z))$ with $k = 128$. The teacher probability is renormalized over $\mathcal{K}_t$ as

$$q_T^k(v \mid h_t^z) = \frac{\pi_T(v \mid h_t^z)}{\sum_{u \in \mathcal{K}_t} \pi_T(u \mid h_t^z)}, \qquad v \in \mathcal{K}_t. \tag{14}$$

The corresponding Teacher FKL objective is

$$\mathcal{L}_{\mathrm{FKL}}^{\mathrm{teacher}}(t) = \sum_{v \in \mathcal{K}_t} q_T^k(v \mid h_t^z) \left[\log q_T^k(v \mid h_t^z) - \log \pi_\theta(v \mid h_t^z)\right]. \tag{15}$$

As shown in Table 2, the relay token objective (46.96) substantially outperforms Student draft token (44.56) and Teacher FKL (44.08). The decline with Teacher FKL aligns perfectly with the design rationale in §3.2: the reverse-KL-style single-sample objective is inherently mode-seeking, enabling the student to selectively absorb the teacher's corrective signals on failed prefixes. In contrast, the mode-covering nature of Teacher FKL forces the student to match the teacher's full distribution, indiscriminately incorporating unreliable guidance. Student draft token predominantly suppresses tokens the student originally tended to generate, whereas using the actually generated tokens after takeover provides a more explicit learning signal.

4.5 Sensitivity Analysis

Moderate intervention balances teacher correction against the on-policy property of the trajectory. Figure 7 summarizes average accuracy when varying one hyperparameter at a time; complete per-benchmark results appear in Appendix D.1.

Relay budget (M, L)**.** As shown in the left panel, average accuracy improves from 44.31 to 47.10 as L increases from 0 to 4, then drops to 46.47 at $L = 5$; both $L = 3$ and $L = 4$ achieve strong performance. For M, $M = 1$ and $M = 2$ achieve 46.25 and 46.96, whereas $M = 4$ drops

to 44.01. Figure 2(a) already showed that extending teacher takeover yields diminishing returns; training results here further confirm that overly long or frequent teacher takeovers push the trajectory too far from the student's current policy, undermining the advantages of on-policy distillation.

Handoff top-K. As shown in the right panel, $K = 1$, $K = 5$, and $K = 10$ achieve average accuracies of 44.27, 46.96, and 43.14—all exceeding the 41.23 of standard OPD (equivalent to $K = |\mathcal{V}|$), validating state-driven selective intervention. $K = 5$ achieves the best result, indicating that the divergence threshold K should be neither too small (over-triggering on minor differences) nor too large (missing genuine divergences).

5 Conclusion

We introduced Relay-OPD, which addresses the prefix failure problem in on-policy distillation by detecting teacher–student reasoning-direction divergences online and letting the teacher intervene locally at detected failure points. A limited relay budget concentrates intervention on critical early positions while limiting departure from the student policy. Across two Qwen3 student models and eight mathematical reasoning benchmarks, Relay-OPD outperforms standard OPD by +5.73% and the strongest trajectory intervention baseline by +1.49% on average, while reducing training trajectory length by over 50%. Further ablations and sensitivity analyses validate the contribution of each design component.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

AI-MO. Aime 2024. https://huggingface.co/datasets/AI-MO/aimo-validation-aime, 2024a.

AI-MO. Amc 2023. https://huggingface.co/datasets/AI-MO/aimo-validation-amc, 2024b.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions, February 2025. URL https://matharena.ai/.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

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 *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

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

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

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.

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 *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pp. 19274–19286. PMLR, 2023.

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

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

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

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20286–20332, 2025.

Carlos Miguel Patiño, Kashif Rasul, Quentin Gallouédec, Ben Burtenshaw, Sergio Paniego, Vaibhav Srivastav, Thibaud Frere, Ed Beeching, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Unlocking on-policy distillation for any model family. https://huggingface.co/spaces/HuggingFaceH4/on-policy-distillation, 2025.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026.

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

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Yan Xie, Sijie Zhu, Tiansheng Wen, Bo Chen, and Yifei Wang. On the position bias of on-policy distillation. *arXiv preprint arXiv:2606.22600*, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, volume 2025, pp. 64616–64646, 2025.

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*, 2025a.

Wenkai Yang, Yankai Lin, Jie Zhou, and Ji-Rong Wen. Distilling rule-based knowledge into large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 913–932, 2025b.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025.

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. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D Lyng, Sanjit Singh Batra, and Robert E Tillman. Fast and effective on-policy distillation from reasoning prefixes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25553–25569, 2026.

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. *arXiv preprint arXiv:2601.18734*, 2026.

Zhou Ziheng, Jiaqi Li, Huacong Tang, Ying Nian Wu, and Demetri Terzopoulos. Less is more: Early stopping rollout for on-policy distillation. *arXiv preprint arXiv:2605.27028*, 2026.

A Training Configuration

A.1 Handoff Trigger Word List

The reflection token set $\mathcal{R}$ used in the handoff criterion consists of the following base words together with their case and leading-space variants:

Handoff Trigger Word List

```
Wait, But, Hmm, Actually, Hold, However, Yet, Oh,
Alternatively, No, Ah, Oops, Well
```

A.2 Training Hyperparameters

The complete training hyperparameters for the online distillation methods and for GRPO are listed in Table 3 and Table 4, respectively.

Table 3: Training hyperparameters for online distillation methods.

Hyperparameter	Value
Max Prompt Length	2048
Max Response Length	16384
Rollout Temperature	1.0
Rollout Top-p	1.0
Rollout Number per Prompt (n)	1
Global Batch Size	128
PPO Mini-Batch Size	128
PPO Epochs	1
PPO Clipping Range	0.2
Learning Rate	1×10^{-6}
Learning-Rate Schedule	constant
Training Epochs	1

Table 4: Training hyperparameters for GRPO.

Hyperparameter	Value
Max Prompt Length	2048
Max Response Length	16384
Rollout Temperature	1.0
Rollout Top-p	1.0
Rollout Number per Prompt (n)	8
Global Batch Size	128
PPO Mini-Batch Size	128
PPO Epochs	1
PPO Clipping Range	0.2
Learning Rate	1×10^{-6}
Learning-Rate Schedule	constant
Training Epochs	1

B Prompts

B.1 Training and Inference Prompt

Both training and inference use the following template:

Training and Inference Prompt

< |im_start| >system
Please reason step by step, and put your final answer within \boxed{}.< |im_end| >
< |im_start| >user
{problem}< |im_end| >
< |im_start| >assistant
<think>

</think>

B.2 TRD Rewrite Prompt

We adopt the rewrite prompt from the original TRD paper (Jiang et al., 2026), with user content as follows:

TRD Rewrite Prompt

Your task is to rewrite your mathematical solution.

Problem:
{problem}

Your Initial Solution:
{initial_response}

Instructions:
1. Preserve the overall structure and reasoning path of your original solution
2. Identify and fix errors in computation or logic
3. Keep correct intermediate steps and meaningful work
4. Output ONLY the rewritten solution

C Algorithm and Baselines

C.1 Relay-OPD Algorithm

The relay rollout construction procedure of Relay-OPD is given in Algorithm 1.

Algorithm 1 Relay Rollout Construction

Require: prompt x, student policy $\pi_{\tilde{\theta}}$, teacher policy π_T, reflection-token set $\mathcal{R}$, handoff top-K, relay budget (M, L)
Ensure: relay trajectory z

Initialize $z \leftarrow ()$, takeover count $j \leftarrow 0$, state $s \leftarrow \mathsf{S}$
while neither EOS nor the maximum length is reached **do** ▷ either event enters the terminal state $\perp$
  // Student drafting and handoff detection
  Set prefix $h \leftarrow (x, z)$ and sample student draft token $a^S \sim \pi_{\tilde{\theta}}(\cdot \mid h)$
  Compute teacher argmax $a^T \leftarrow \arg\max_v \pi_T(v \mid h)$ and student top-K set $\mathcal{K}_S(h) \leftarrow \mathrm{TopK}_K(\pi_{\tilde{\theta}}(\cdot \mid h))$
  Evaluate the handoff criterion $\phi(h) \leftarrow \mathbf{1}[a^T \in \mathcal{R}] \cdot \mathbf{1}[\mathcal{K}_S(h) \cap \mathcal{R} = \varnothing]$
  if $s = \mathsf{S}$ **and** $\phi(h) = 1$ **and** $j < M$ **then** ▷ $\mathsf{S} \to \mathsf{T}$ transition
    // Teacher leg
    $j \leftarrow j + 1$, $s \leftarrow \mathsf{T}$; append a^T to z ▷ reflection token opens the teacher leg
    for $\ell = 1, \ldots, L$ **do**
      Generate one teacher paragraph via speculative decoding (§3.3) and append it to z
    end for
    if $j = M$ **then**
      break ▷ relay budget exhausted; enter the terminal state $\perp$
    else
      $s \leftarrow \mathsf{S}$ ▷ $\mathsf{T} \to \mathsf{S}$ transition: the student resumes
    end if
  else
    Append a^S to z ▷ student leg, including $\phi(h) = 1$ with $j = M$
  end if
end while

C.2 Method Comparison Overview

We summarize all compared methods in Table 5 along policy type, trajectory source, training loss, and method-specific settings.

Table 5: **Method Comparison Overview.** Original FKL-based papers use full vocabulary or varying top-k settings; we uniformly adopt top-k=128. The original SKD paper uses acceptance top-k=25, but in our teacher–student configuration top-25 barely filters any student draft tokens, so we adopt top-k=5.

Method	Policy	Trajectory Source	Loss	Method-Specific Settings
SFT	Off-policy	Teacher trajectory	CE	—
KD	Off-policy	Teacher trajectory	FKL	—
TRD	Off-policy	Rewritten student trajectory	FKL	rewrite prompt (B.2)
GRPO	On-policy	Student rollout	Outcome RL	8 rollouts per prompt
OPD	On-policy	Student rollout	RKL	max response = 16384
FastOPD	On-policy	Student rollout	RKL	fixed truncation $\in \{1024, 2048, 4096, 8192\}$ (C.5)
SKD	On-policy	Speculative mixed rollout	FKL	acceptance top-K=5 (C.4)
Relay-OPD	On-policy	Relay rollout	RKL	handoff top-K; relay budget (M, L)

C.3 TRD

TRD (Jiang et al., 2026) constructs training data offline in two stages. Stage one lets the student generate an initial trajectory y_o from the original problem x; stage two provides x and y_o to the teacher within a rewrite prompt, producing a rewritten trajectory y_r. We adopt the rewrite prompt from the original TRD paper; the complete prompt is given in Appendix B.2.

During TRD training, teacher and student use different contexts. For the t-th token in the rewritten trajectory, the student context is $h_t^S = (x, y_{r,<t})$, while the teacher context additionally includes the student's initial trajectory: $h_t^T = (x, y_o, y_{r,<t})$. On teacher context h_t^T, the top-k support set is $\mathcal{K}_t^T := \mathrm{TopK}_k\big(\pi_T(\cdot \mid h_t^T)\big)$ with $k = 128$, and teacher probability is renormalized over this support set as

$$q_T^k(v \mid h_t^T) := \frac{\pi_T(v \mid h_t^T)}{\sum_{u \in \mathcal{K}_t^T} \pi_T(u \mid h_t^T)}, \qquad v \in \mathcal{K}_t^T. \tag{16}$$

The corresponding top-k FKL objective is

$$\mathcal{L}_{\mathrm{TRD}}(t) = \sum_{v \in \mathcal{K}_t^T} q_T^k(v \mid h_t^T) \left[\log q_T^k(v \mid h_t^T) - \log \pi_\theta(v \mid h_t^S)\right]. \tag{17}$$

This asymmetry means that the teacher provides supervision conditioned on y_o, while the student cannot access y_o at either training or inference time.

C.4 SKD

SKD (Xu et al., 2025) constructs teacher–student mixed trajectories via speculative decoding. The student serves as draft model and the teacher as target model; throughout this subsection, z denotes SKD's mixed trajectory and $h_t = (x, z_{<t})$ its prefix. Given prefix h_t and student draft token a_t^S, the teacher's top-K support set is $\mathcal{K}_T(h_t) = \mathrm{TopK}_K(\pi_T(\cdot \mid h_t))$ with $K = 5$, and the draft is accepted iff $a_t^S \in \mathcal{K}_T(h_t)$; if accepted, $z_t = a_t^S$. If rejected, a replacement token is sampled from the teacher's distribution renormalized over $\mathcal{K}_T(h_t) \setminus \{a_t^S\}$:

$$z_t \sim \widetilde{\pi}_T(\cdot \mid h_t), \qquad \mathrm{supp}(\widetilde{\pi}_T) = \mathcal{K}_T(h_t) \setminus \{a_t^S\}. \tag{18}$$

Replaced positions are marked as teacher-owned. The training objective uses top-k FKL on the mixed trajectory, with teacher and student sharing the same prefix h_t.

C.5 FASTOPD

FastOPD (Ziheng et al., 2026; Zhang et al., 2026) truncates the student rollout at a fixed length. Given student-generated trajectory $y = (y_1, \ldots, y_N)$ and truncation length B, the training trajectory is $y^{(B)} = (y_1, \ldots, y_{N_B})$ with $N_B = \min(N, B)$.

FastOPD then optimizes the standard OPD objective (§3.1) on $y^{(B)}$, replacing N and y with N_B and $y^{(B)}$. Thus FastOPD changes only the rollout length participating in training, not the token-level learning signal of standard OPD.

D SUPPLEMENTARY RESULTS

D.1 COMPLETE SENSITIVITY RESULTS

Complete per-benchmark results for the relay budget and handoff top-K sweeps of §4.5 are reported in Table 6.

Table 6: **Complete Sensitivity Results.** Using Qwen3-4B-Instruct-2507 as teacher and Qwen3-1.7B-Non-Thinking as student, varying one hyperparameter at a time. Standard OPD corresponds to $K = |\mathcal{V}|$; bold denotes the best result in each column.

Variant	AIME24	AIME25	AIME26	MATH	AMC23	Olymp.	HMMT Feb26	HMMT Nov25	Avg
$L = 0$	40.42	29.06	27.19	88.15	73.83	56.31	23.48	16.04	44.31
$L = 1$	41.46	31.46	26.88	88.75	76.64	57.68	22.73	18.85	45.56
$L = 2$	40.52	32.50	28.23	88.55	78.20	57.23	22.16	19.06	45.81
$L = 4$	42.60	32.40	**32.71**	**89.50**	78.20	58.49	23.39	19.48	**47.10**
$L = 5$	40.00	**32.81**	30.10	89.35	**78.28**	57.94	23.58	19.69	46.47
$M = 1$	42.40	31.98	29.69	88.55	75.47	58.38	23.30	20.21	46.25
$M = 3$	**43.12**	31.46	26.88	88.85	76.48	57.46	22.82	**20.42**	45.94
$M = 4$	37.60	32.08	26.88	86.90	73.52	56.34	21.59	17.19	44.01
$K = 1$	40.00	29.90	24.38	87.55	74.22	56.86	22.73	18.54	44.27
$K = 10$	36.46	29.27	26.04	88.15	73.05	56.19	20.36	15.62	43.14
$K = \|\mathcal{V}\|$	35.83	25.52	23.33	85.70	70.08	55.27	20.08	14.06	41.23
Default ($K = 5$, $M = 2$, $L = 3$)	42.71	**32.81**	30.52	**89.50**	76.88	**58.79**	**24.72**	19.79	46.96

D.2 FASTOPD TRUNCATION LENGTH SWEEP

Complete results across fixed truncation lengths are reported in Table 7. Both student models achieve the highest average accuracy at 4,096 tokens; Table 1 therefore reports this configuration.

Table 7: **FastOPD Fixed Truncation Length Comparison.** Subscripts in the Avg column indicate the training step of the best checkpoint; Train Len denotes average rollout response length from training start to that checkpoint.

Truncation Length	AIME24	AIME25	AIME26	MATH	AMC23	Olymp.	HMMT Feb26	HMMT Nov25	Avg	Train Len
Student: Qwen3-0.6B-Non-Thinking										
1,024	15.94	20.42	12.40	74.65	54.69	43.95	10.51	9.58	$30.27_{@105}$	987
2,048	15.83	19.38	15.42	76.90	51.48	44.77	10.51	8.33	$30.33_{@95}$	1,827
4,096	15.83	20.10	13.54	75.30	53.67	44.81	11.55	8.54	$\mathbf{30.42}_{@100}$	3,302
8,192	14.79	18.44	11.88	75.35	51.80	43.36	9.66	8.65	$29.24_{@60}$	5,055
Student: Qwen3-1.7B-Non-Thinking										
1,024	39.06	26.46	25.21	86.00	70.70	57.08	23.20	19.17	$43.36_{@105}$	983
2,048	43.23	29.58	28.33	88.35	72.11	57.42	23.67	19.17	$45.23_{@75}$	1,753
4,096	42.29	30.42	26.35	87.95	74.30	58.16	23.58	20.73	$\mathbf{45.47}_{@45}$	2,709
8,192	40.52	27.60	26.56	86.70	72.81	56.27	21.88	20.31	$44.08_{@50}$	4,004

E CASE STUDIES

E.1 RELAY-OPD TAKEOVER CASE

In this case, the student has only 5 RM remaining after purchasing 15 apples yet incorrectly concludes it can still afford both a mango and a papaya. Without teacher intervention, it rationalizes this plan into a total of 17 fruits. Faced with the same prefix, the teacher selects `But` with 74.4084% probability (Table 8), whereas the student favors `So` or `Now` to extend the current conclusion. The teacher leg points out that the two fruits would cost at least 9 RM, and the student accordingly rejects the original plan and finds the feasible combination of 12 apples, 1 mango, and 2 papayas, arriving at the correct answer of 15. This case illustrates two points: the handoff trigger correctly identifies the moment the student would continue along a wrong direction, and a brief teacher takeover suffices to correct the subsequent reasoning.

Problem

Danial went to a fruit stall that sells apples, mangoes, and papayas. Each apple costs 3 RM, each mango costs 4 RM, and each papaya costs 5 RM. He bought at least one of each fruit and paid exactly 50 RM. What is the maximum number of fruits that he could have bought?

Ground truth: `15`

Student Trajectory (w/o Intervention)

...
Now, try **15 apples**:

- $15 \times 3 = 45$
- Remaining: $50 - 45 = 5$

We can buy **1 papaya** (5 RM), and at least 1 mango (4 RM). **So that's valid.**
[Handoff trigger detected; the following is the student's original continuation]
So:

- 15 apples
- 1 mango
- 1 papaya

Total fruits: $15 + 1 + 1 = 17$
Is this the **maximum** possible? Let's check.
Try **14 apples**: $14 \times 3 = 42$, remaining $50 - 42 = 8$; we can buy 1 mango (4 RM) and 1 papaya (5 RM). Total: $14 + 1 + 1 = 16$ fruits. Smaller than 17.
...
Final Answer:

$$\boxed{17}$$

Table 8: Teacher and student distributions.

Rank	Teacher distribution	Student distribution
1	`But` 74.4084%	`So` 50.6288%
2	`So` 10.0701%	`Now` 21.1052%
3	`Wait` 7.8426%	– 16.4368%
4	`We` 4.1978%	`Total` 5.3362%
5	– 1.2027%	`Then` 2.8563%
6	`Can` 0.5014%	`That` 0.5624%
7	`Now` 0.2368%	`F` 0.4963%
8	`Remaining` 0.2368%	`This` 0.4380%
9	`Then` 0.2090%	`We` 0.4380%
10	`Try` 0.1268%	`Let` 0.3866%

Relay Trajectory

...
Now, try **15 apples**:

- $15 \times 3 = 45$
- Remaining: $50 - 45 = 5$

We can buy **1 papaya** (5 RM), and at least 1 mango (4 RM). **So that's valid.**
[Handoff trigger detected; the teacher takes over]
But we have only 5 RM left after 15 apples — we would have to spend 4 RM on mango and 5 on papaya — total 9, which exceeds 5.
Wait — no: we have 5 RM left, and we must spend at least 4 RM for mango and 5 RM for papaya, but only 5 RM is available. So both cannot be bought.
We need at least one mango and one papaya.
[Teacher leg ends; generation returns to the student]
So 15 apples is not valid.
...
Try **12 apples**:

- $12 \times 3 = 36$
- Remaining: $50 - 36 = 14$

Try:

- 1 mango = 4, 2 papayas = 10 $\rightarrow$ total = 14 $\rightarrow$ **valid!**

So:

- 12 apples
- 1 mango
- 2 papayas

Total = $12 + 1 + 2 = \mathbf{15}$ fruits
...

$\boxed{15}$

E.2 TRD Rewritten Text

We matched references to the original solution (e.g., `initial/original solution`), direct descriptions of the rewriting task (e.g., `rewrite/revision`), and reviewer-style openings (e.g., `after review/reevaluation`); 18.96% of trajectories contain at least one such expression, confirming that these rewriting artifacts are not isolated. The following representative fragments are extracted directly from teacher-generated rewritten trajectories; ellipses indicate truncated content.

TRD Rewritten Fragments

Given the complexity, and that **the original solution incorrectly assumed a simple parity, we must revise.** ... a common variant of this game has the property that the first player wins iff ($N \bmod 3 = 1$). Even though our manual simulation showed ($N = 6$) ... **perhaps we made a mistake.**

This may be incorrect based on simulation; however, without complete DP simulation ... **this is the most reasonable periodic answer.** ... **we go with** $\boxed{674}$**.**

E.3 SKD Repetition Patterns

SKD accepts a student draft token whenever it falls within the teacher's top-K support set (Appendix C.4). Because acceptance is driven by this generic distributional agreement, SKD struggles to break a repetitive generation pattern once the student has established it: the tokens that continue the repetition typically remain inside the teacher's top-K candidates, even when the teacher itself would terminate or redirect. The following fragment from an SKD mixed trajectory is representative: the student repeatedly emits a completed final-answer block.

SKD Mixed Trajectory

...
Final Answer:

$$\boxed{\text{No integer } n}$$

Final Answer:

$$\boxed{\text{No integer } n}$$

...

Table 9: Teacher and student distributions at the end of a repeated answer block.

Rank	Teacher distribution	Student distribution
1	<|im_end|> 73.7035%	\n\n 78.5176%
2	\n\n 16.4455%	<|im_end|> 19.8524%
3	Ġâl' 6.8555%	\n 1.2691%
4	\n 0.7226%	or 0.1516%
5	** 0.5627%	Ġâl' 0.0558%

Table 9 shows the teacher and student next-token distributions on the prefix ending at a completed answer block.[1] The teacher intends to terminate generation: its top-1 token is the end-of-sequence token <|im_end|> with probability 73.7035%. The student instead prefers to open yet another paragraph, placing 78.5176% on \n\n. Under the acceptance criterion $a_t^S \in \mathcal{K}_T(h_t)$ with $K = 5$, however, \n\n still ranks second in the teacher's support set (16.4455%), so the student draft is accepted and the repetition continues—and because the same distributions recur on the extended prefix, the trajectory can degenerate into an **infinite loop** of final-answer blocks. Acceptance driven by generic distributional agreement thus provides no explicit signal that the established generation pattern itself has gone wrong, which underlies the repetition-related degradation of SKD discussed in §4.2.

F LIMITATIONS

Our evaluation focuses on mathematical reasoning with Qwen3 teacher–student pairs, the strong-to-weak setting targeted by prior on-policy distillation work. The relay mechanism itself is task-agnostic, and we leave its application to other domains, such as code generation or agentic tool use, to future work; the reflection-token set in Appendix A.1 may need adjustment when switching to a different model family. Relay-OPD also presumes a teacher whose continuations redirect failed prefixes more reliably than the student's, so its benefit is expected to diminish as the teacher–student capability gap narrows. Finally, the relay budget was tuned on the 1.7B student and reused unchanged for the 0.6B student; the sensitivity analysis in §4.5 indicates robustness across moderate budget ranges, though new model pairs may benefit from re-tuning.

[1] Token strings appear in the tokenizer's byte-level representation, which maps every byte to a printable character: Ġ denotes a leading space, and âl' are the first two UTF-8 bytes of a check-mark symbol (e.g., ✓), whose final byte would be completed by a subsequent token.