MISMATCH MATTERS: ON-POLICY DISTILLATION BEYOND TOKEN AGREEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) has emerged as a core component of modern LLM post-training pipelines, yet we reveal a failure mode: *degenerate agreement*, where students exploit repetitive loops to achieve near-perfect token agreement with the teacher despite globally flawed responses [1]. We therefore shift our focus from agreement to teacher–student mismatch and find that the mismatch tokens can be mainly categorized to two types: student-excess tokens and student-deficit tokens. Specifically, student-excess tokens are those generated by the student but assigned near-zero probability by the teacher; their log-ratio corrections grow unbounded and destabilize the update. Student-deficit tokens, in contrast, are preferred by the teacher but rarely sampled by the student; their absence blocks the transfer of the teacher's reasoning patterns. To tackle these mismatch directions, we propose **TIDE** (Token-level Independent Deficit–Excess correction), which applies bounded Hellinger shaping to suppress the most severe sampled excesses and an analytic teacher top-K injection to restore deficient probability mass without requiring deficit tokens to be sampled. Across mathematical reasoning benchmarks with multiple Qwen3 teacher–student pairs, TIDE consistently outperforms standard OPD and recent token-selection and reward-shaping baselines. Besides, the gains of TIDE are more pronounced under strong teacher–student mismatch, where it improves Avg@8 from 6.9% to 20.3%, reduces average response length by a factor of 3.6, and substantially reduces formatting failures. Our code is available at https://anonymous.4open.science/r/TIDE-B1B2.

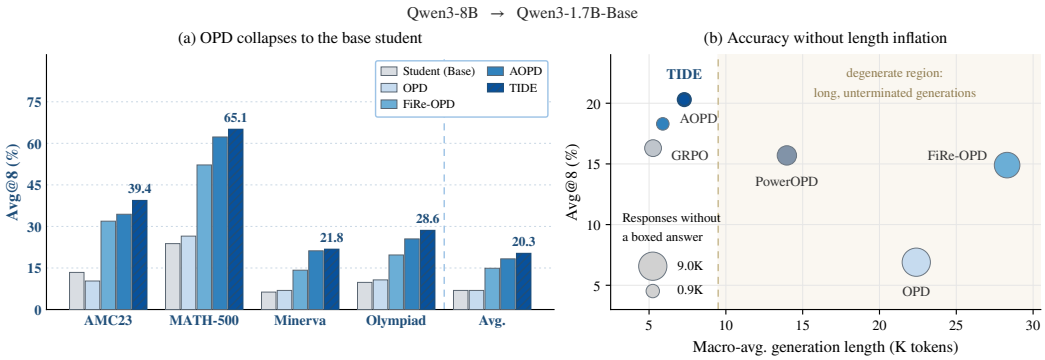

Figure 1: **TIDE achieves the highest accuracy without length inflation.** **(a)** TIDE consistently outperforms the baselines in Avg@8. **(b)** Unlike OPD and FiRe-OPD, TIDE maintains short, well-formed generations.

1 INTRODUCTION

On-policy distillation (OPD) (Lu & Lab, 2025) has emerged as a powerful and widely adopted paradigm for post-training large language models (LLMs) (Agarwal et al., 2024; Gu et al., 2024), and has become a standard component in recent open-weight frontier models such as Qwen3, MiMo-V2-Flash, GLM-5, and Kimi K3 (Yang et al., 2025; Xiao et al., 2026; Zeng et al., 2026; Team

[1] Representative examples of degenerate agreement are provided in Section F.

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.

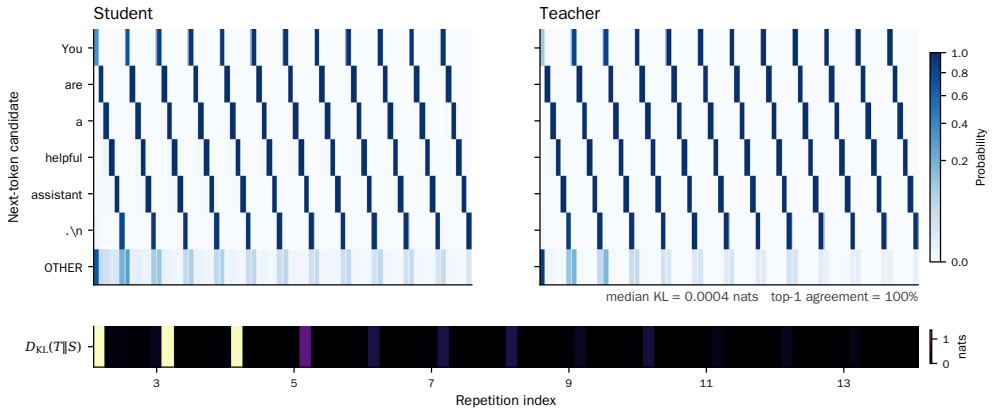

Figure 2: **Repetition creates degenerate agreement.** For a representative OPD rollout repeating `You are a helpful assistant.\n`, the heatmaps show student (left) and teacher (right) next-token distributions under increasingly repetitive prefixes. Rows denote tokens in the repeating unit; darker colors indicate higher probability. Despite global degeneration, top-1 agreement reaches 100%, while median teacher–student KL falls to 0.0004 nats (bottom strip).

et al., 2026). Rather than relying on fixed expert demonstrations, OPD samples trajectories from the student policy and queries an external teacher at the states the student actually visits. This on-policy supervision mitigates the exposure bias of supervised fine-tuning (SFT) (Bengio et al., 2015; Ross et al., 2011; Agarwal et al., 2024), while offering denser token-level feedback than the sparse outcome rewards used in reinforcement learning with verifiable rewards (RLVR) (Shao et al., 2024; Guo et al., 2025; Yu et al., 2026).

Despite these advantages, OPD is prone to abrupt length inflation, with student generations drifting toward the decoding cap (Luo et al., 2026). Our diagnostics trace this to persistent repetitive loops (see detailed analysis in Section 2.1 and one example in Fig. 2): the student repeatedly generates an identical prefix `"You are a helpful assistant.\n"`; conditioned on the student-generated prefix, the teacher finds each additional repetition increasingly predictable, until the two distributions (generated by student and teacher) nearly coincide (median KL 0.0004 nats). Critically, the student can thus drive teacher–student divergence to near zero while the overall response remains globally degenerate, revealing that high token-level agreement can actively mask global degeneration. We term this phenomenon *degenerate agreement*, and the underlying mechanism *student-induced teacher hacking*: repetition renders a degenerate continuation locally consistent to the teacher. This also resembles the low-KL trap recently observed for corrupted student prefixes (Xin et al., 2026).

A prevalent principle in recent OPD research is that training should focus on high teacher–student overlap to obtain valuable supervision. This assumption underpins recent work on token selection (Li et al., 2026a; Zhao et al., 2026b; Yang et al., 2026b), which treats high-overlap tokens as the most useful learning signals. However, this overlap-based criterion may also suffer from the risk of degenerate agreement: repetitive generations can produce artificially high overlap and then the teacher signal offers little new information. In such cases, high agreement ceases to be a reliable indicator of supervisory quality. This motivates us to shift the focus from agreement to mismatch (or disagreement) as the basis for supervision, and explore divergent tokens as an alternative, potentially more informative source of learning signal.

Concretely, we find that two types of mismatch tokens are rather distinct and important (see Section 2 for details): student-excess and student-deficit tokens. In particular, *Student-excess tokens* are overproduced by the student relative to the teacher; they are readily observed in rollouts, yet their log-ratio difference, can grow unbounded as the teacher probability approaches zero, destabilizing updates. For instance, in our OPD training, the most negative 1% of tokens contribute nearly half of the total gradient. *Student-deficit tokens*, conversely, are favored by the teacher but underweighted by the student. These tokens have a median probability of just 2.04×10^{-4}; under our four-rollout budget, 95.5% of them are sampled with less than 1% probability. Despite this inaccessibility, we find that these deficit tokens carry rather valuable supervisory signal missed by standard on-policy

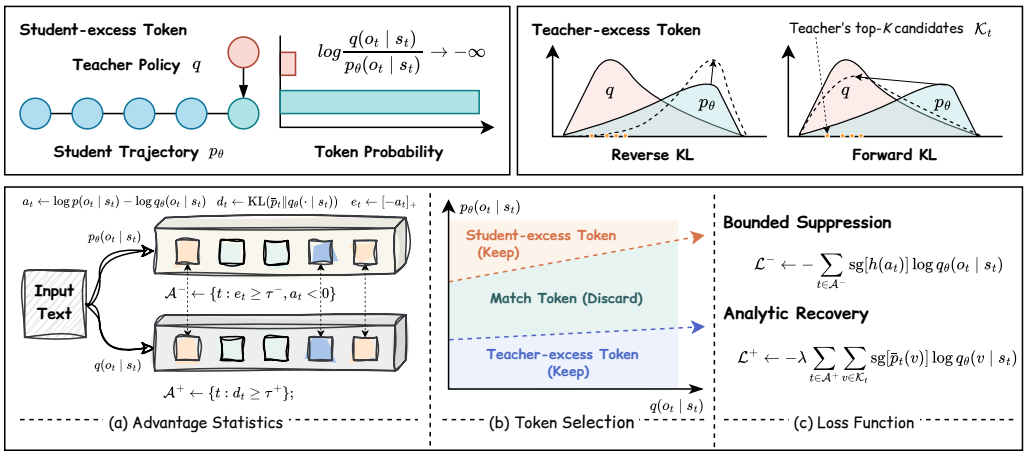

Figure 3: **An overview of TIDE pipeline.** Student-excess tokens are readily observed in student rollouts, but can receive arbitrarily negative advantages as teacher probabilities approach zero. In contrast, rarely sampled student-deficit tokens are identified analytically using forward KL over the teacher's top-K candidates. TIDE computes branch-specific mismatch statistics (a), independently retains severe excess or deficit positions while discarding low-mismatch ones (b), and applies bounded excess suppression and analytic deficit recovery (c).

supervision (we perform a simple deficit-only intervention and raise Avg@8 from 6.87 to 18.32 (see Section D.3).

Based on our findings on token-level mismatch, we propose **TIDE** (**T**oken-level **I**ndependent **D**eficit–**E**xcess correction). As illustrated in Fig. 3, TIDE computes branch-specific mismatch statistics (a), independently selects severe excess and deficit positions while excluding matched tokens (b), and applies direction-specific corrections (c). For student excess, it replaces unbounded log-ratios with bounded weights derived from the squared Hellinger divergence. For student deficit, it analytically restores missing probability mass over the teacher's top-K candidates without requiring them to be sampled. We further develop theoretical results demonstrating that the excess branch provides bounded, locally faithful corrections aligned with a proper divergence geometry, while the deficit branch directly addresses an irreducible coverage gap that sampling-based methods cannot resolve. Across mathematical reasoning benchmarks and teacher-student pairs, TIDE consistently outperforms strong baselines, with ablations confirming the complementary roles of its two branches. Our contributions are summarized as follows.

- We identify *student-induced teacher hacking* as a previously overlooked failure mode of OPD. Once the student enters a repetitive prefix, the teacher can favor the same continuation, creating a low-KL matched region in which degeneration receives little corrective signal.
- We reveal an intrinsic asymmetry in informative teacher–student disagreement. Student-excess tokens are readily observed but can induce unstable updates, whereas student-deficit tokens carry valuable teacher information but are rarely observed in student rollouts.
- We introduce TIDE, which combines quantile-gated Hellinger suppression for student excess with analytic teacher-top-K recovery for student deficit while removing redundant supervision from matched tokens. TIDE consistently outperforms strong OPD baselines; notably, for our Qwen experiment setting, it raises Avg@8 from 6.87 to 20.20 while reducing the average response length from 22,395 to 7,294 tokens.

Notations. Let $\mathcal{D}$ denoted the prompt distribution and $\mathcal{V}$ the vocabulary. Then, we denote a prompt by x, and a response by $o = (o_1, \ldots, o_T)$, where T counts all response tokens including EOS. The prefix before position t is denoted by $o_{<t}$, and the corresponding state by $s_t = (x, o_{<t})$. $p(\cdot|s)$ and $q_\theta(\cdot|s)$ denote conditional token distributions, with θ denoting model parameters. $\mathbb{E}$, D_{KL}, and sg denote expectation, Kullback–Leibler divergence, and the stop-gradient operator, respectively.

2 PROBLEM SETUP

The teacher model p is fixed, while the student model q_θ is optimized. Both models have full support on $\mathcal{V}$. Given $x \sim \mathcal{D}$, the student generates $o \sim q_\theta(\cdot \mid x)$. OPD minimizes the expected trajectory-level reverse KL divergence:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}[D_{\mathrm{KL}}(q_\theta(\cdot \mid x) \,\|\, p(\cdot \mid x))] = \mathbb{E}_{\substack{x\sim\mathcal{D} \\ o\sim q_\theta(\cdot|x)}} \left[\sum_{t=1}^{T} \log \frac{q_\theta(o_t \mid s_t)}{p(o_t \mid s_t)} \right]. \tag{1}$$

Following recent practice (Lu & Lab, 2025), sampled-token OPD uses a detached token-level policy-gradient update (Schulman et al., 2017):

$$g_{\mathrm{OPD}}(\theta) = -\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{sg}[a_t] \, \nabla_\theta \log q_\theta(o_t \mid s_t) \right], \qquad a_t = \log \frac{p(o_t \mid s_t)}{q_\theta(o_t \mid s_t)}.$$

The expectation follows the sampling process in equation 1. Positive advantages reinforce sampled tokens, while negative advantages suppress them. The supervision signal is evaluated on student-sampled tokens, making its coverage dependent on the student's distribution (Section 2.2).

2.1 MOTIVATION: STUDENT-INDUCED TEACHER HACKING

In this section, we highlight several key observations about the interaction between student-generated trajectories and teacher supervision during OPD. In OPD, the teacher evaluates tokens conditioned on student-generated prefixes. This allows the student to steer the teacher's conditional predictions toward its own, reducing measured teacher-student divergence without its reasoning ability. We refer to this failure mode as *student-induced teacher hacking*. The experimental setup is detailed in Section 4.1.

We first examine the trajectory of the OPD training and find a conspicuous pattern *repetitive generation*: as training proceeds, the student increasingly falls into loops where the same token sequence is repeated many times, e.g., repeating `I hope it is correct` 149 times in math problems. We then evaluate a fixed set of prompts across training checkpoints and find that the proportion of repetitive rollouts rises from 16.8% to 48% (Fig. 4a). These loops are rarely transient; once the student enters one, it tends to persist for many iterations, yielding the heavy-tailed distribution of repetition counts shown in Fig. 4b. Notably, neither the prompts nor the objective function explicitly encourage repetition. This behavior is conjectured to emerge organically from the interplay between OPD optimization and the student-conditioned teacher signal.

We next examine how the teacher reacts to these loops. For each naturally occurring repetitive rollout, we construct controlled prefixes with varying repetitions of its repeating unit and evaluate them on the same subsequent tokens. As shown in Fig. 4c, increasing repetition lowers both student–teacher KL and teacher entropy, with sixteen repetitions reducing KL by $63\times$. Far from correcting the repetition, the teacher becomes increasingly confident in continuing it. This pattern also holds under greedy decoding: when prompted with natural student loop prefixes, the teacher continues the same repetitive pattern in 93% of cases (Section D.1). In effect, repetition turns a degenerate student trajectory into apparent teacher-student agreement, weakening corrective supervision.

2.2 INFORMATIVE MISMATCH IS DIRECTIONAL.

The results above show that teacher–student agreement need not constitute useful supervision. Our matched-only control provides further evidence: supervising the 60% best-matched positions raises Avg@8 of Qwen3-1.7B-Base from 6.87 to just 7.18, whereas supervising the 20% most mismatched positions more than doubles it to 14.58 (Section D.2 and Table 5). These observations motivate us to shift the gradient budget away from already matched positions and toward informative disagreement, particularly under strong teacher–student mismatch.

Having established where supervision should be concentrated, we next consider how different forms of disagreement should be handled. The reverse-KL objective in equation 1 penalizes student mass on teacher-unlikely tokens but provides comparatively weak pressure to cover teacher-preferred tokens outside the student's effective support (Minka et al., 2005; Gu et al., 2024; Agarwal et al., 2024). This introduces two directional challenges, illustrated in the top row of Fig. 3.

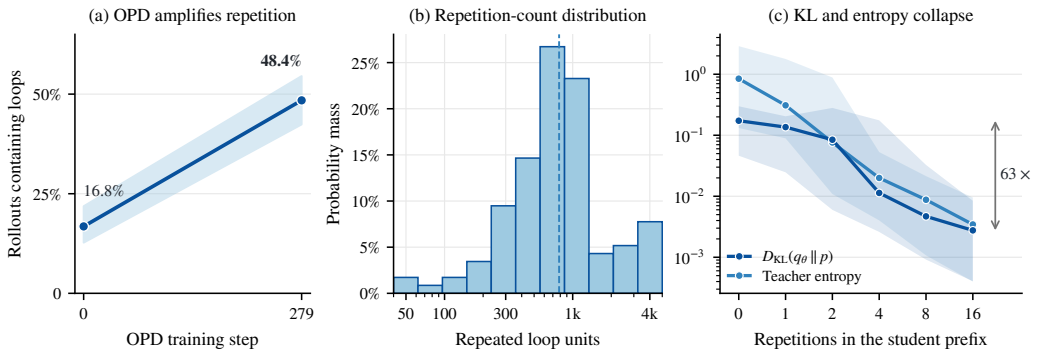

Figure 4: **Repetition emerges as an exploit of teacher conditioning in OPD.** (a) Repetitive-rollout rates on fixed DAPO-Math-17K prompts across training checkpoints; shading denotes 95% confidence intervals. (b) Repetition-counts distribution at the final checkpoint. (c) Controlled prefix repetition reduces both the student–teacher KL divergence and the teacher entropy; 16 repetitions lower the KL divergence by 63×.

A negative advantage $a_t < 0$ indicates that the student overweights the sampled token relative to the teacher. Such student-excess tokens are readily observed in rollouts, but their suppression magnitude $-a_t = \log[q_\theta(o_t \mid s_t)/p(o_t \mid s_t)]$ grows without bound as teacher probability vanishes while the student retains appreciable mass (Zhao et al., 2026a). A handful of strongly teacher-rejected tokens can thus dominate the update and destabilize training. Our first branch targets these severe excesses with a bounded signal in place of the raw log-ratio.

The opposite direction, however, presents a sampling bottleneck. Consider a token v favored by the teacher but assigned near-zero probability by the student. For fixed $p(v \mid s) > 0$, its sampling-weighted teacher-favoring signal satisfies

$$q_\theta(v \mid s) \log \frac{p(v \mid s)}{q_\theta(v \mid s)} \longrightarrow 0 \qquad \text{as } q_\theta(v \mid s) \to 0.$$

Thus, the more severely the student underweights the token, the less opportunity OPD has to sample and reinforce it; reshaping sampled-token advantages alone does not remove this bottleneck. Our measurements show that these rarely observed tokens nevertheless carry substantial learning value (Section D.3). This motivates our second branch: rather than waiting for such tokens to be sampled, we directly inspect the teacher's top-K support and provide analytic guidance toward these under-covered tokens. The next section turns these observations into separate suppression and recovery branches.

3 Methodology

In this section, we present TIDE, which uses the magnitude of teacher-student mismatch to determine *where* to update and its direction to determine *how*. As illustrated in Fig. 3, it independently selects severe excess and deficit positions for bounded suppression and analytic recovery, respectively.

3.1 Stabilizing Student-Excess Updates

We first consider positions where the sampled token receives more probability from the student than from the teacher, i.e., student-excess tokens. These positions satisfy $a_t < 0$ and can be ranked by the magnitude of the excess $e_t = [-a_t]_+$, where $[u]_+ = \max\{u, 0\}$.

Let $\mathcal{B}$ denote the set of response positions in a training batch. To better locate the tokens with severe mismatch, we consider a constant fraction ρ^- of positions ranked by their excess scores. Specifically, we define the corresponding batch-quantile threshold and selection mask as

$$\tau^- = Q_{1-\rho^-}(\{e_j : j \in \mathcal{B}\}), \qquad m_t^- = \mathbb{1}[a_t < 0,\ e_t \geq \tau^-],$$

where Q_α denotes the empirical α-quantile. This selection focuses the distillation signal on positions where the teacher assigns low probability to the student's sampled token. The score-and-

gate construction is shown by the orange route from the bottom-left statistics panel to the retained student-excess region in the bottom-center panel of Fig. 3.

As discussed above, the teacher assigns substantially less probability than the student to tokens selected by the excess gate, making a_t strongly negative. The resulting log-ratio penalty $-a_t = \log[q_\theta(o_t|s_t)/p(o_t|s_t)]$ grows without bound as the teacher probability approaches zero. Thus, directly applying the original OPD advantage after selection would concentrate the gradient too heavily on a small number of extreme corrections, leading to the training instability.

We address this problem by applying a bounded, monotonic transformation to the raw advantage at each selected position. Clipping-style functions (e.g., hard clipping or tanh) would also bound the update, but they rely on an arbitrary saturation scale, collapse all severe discrepancies to nearly identical corrections, and yield an update that no longer corresponds to any well-defined divergence between the two distributions. We instead adopt the Hellinger-shaped function, which bounds the update and locally preserves the log-ratio while remaining the exact gradient of a proper divergence:

$$h(a_t) = 2[\exp(a_t/2) - 1].$$

Accordingly, we take the following the policy gradient in OPD:

$$g_{\text{exc}} = -\mathbb{E}\left[\sum_t m_t^- \ \text{sg}[h(a_t)] \log q_\theta(o_t|s_t)\right].$$

Thus, quantile selection determines where suppression is informative, while Hellinger shaping prevents the selected reverse-KL tail from destabilizing optimization. The resulting bounded suppression objective appears in Fig. 3 (bottom right).

The following proposition explains the critical properties of Hellinger transformation.

Proposition 1 (Bounded, locally faithful Hellinger shaping)**.** *For a student-excess token, $a < 0$,*

$$-2 < h(a) < 0, \qquad h(a) = a + O(a^2) \quad \text{as } a \to 0.$$

Moreover, before position selection, a token sampled from the student gives the following unbiased gradient identity when $h(a(v))$ is treated as a detached coefficient:

$$\mathbb{E}_{v\sim q_\theta}[-h(a(v))\nabla_\theta \log q_\theta(v|s)] = 4\nabla_\theta H^2(p, q). \tag{2}$$

Proposition 1 explains the advantage of Hellinger shaping compared to the simple clipping of the raw advantage. Its negative side is bounded, preventing teacher-rejected tokens from producing arbitrarily large suppression updates. At the same time, $h(a) \approx a$, i.e., in the regime near agreement, so moderate corrections retain the first-order behavior of standard OPD. Moreover, equation 2 shows that this transformation arises naturally from a proper divergence, grounding the design in a principled objective rather than an ad hoc heuristic.

3.2 Recovering Under-Covered Teacher Support

Unlike the excess case, missing teacher support cannot be detected from the sampled advantage a_t, which only compares the two distributions at the student's chosen token. To identify teacher-preferred tokens that the student overlooks, we inspect the teacher's top-K support. Define the top-K support and the renormalized teacher distribution within this support as follows:

$$\mathcal{K}_t = \text{TopK}_{v\in\mathcal{V}}\ p(v|s_t); \quad \bar{p}_t(v) = \frac{p(v|s_t)}{\sum_{u\in\mathcal{K}_t} p(u|s_t)},\ v \in \mathcal{K}_t.$$

Then we quantify the student's deficit on this support based on the following metric

$$d_t = \sum_{v\in\mathcal{K}_t} \bar{p}_t(v) \log \frac{\bar{p}_t(v)}{q_\theta(v|s_t)}. \tag{3}$$

This is the blue deficit statistic shown in the bottom-left panel of Fig. 3, motivated by the teacher-preferred alternatives in the top-right panel. Notably, unlike standard top-K OPD (Jin et al., 2026), we do not renormalize the student probabilities over $\mathcal{K}_t$. The score d_t quantifies how well the student covers the teacher's preferred continuations, and enjoys a natural interpretation: it is nonnegative, vanishes only when the student matches $\bar{p}_t$ within $\mathcal{K}_t$ and assigns all its probability mass to this set, and grows large whenever the student underweights tokens that the teacher deems plausible. These properties are formalized in Proposition 2 below.

Proposition 2 (Deficit-score decomposition). *Let $Q_t = \sum_{v \in \mathcal{K}_t} q_\theta(v|s_t)$ denote the total student mass on $\mathcal{K}_t$, and let $q_t^{\mathcal{K}}(v) = q_\theta(v|s_t)/Q_t$ be the corresponding renormalized distribution for $v \in \mathcal{K}_t$. Then the deficit score in equation 3 satisfies*

$$d_t = \underbrace{D_{\mathrm{KL}}\big(\bar{p}_t \| q_t^{\mathcal{K}}\big)}_{\text{mismatch within } \mathcal{K}_t} + \underbrace{(-\log Q_t)}_{\text{insufficient mass on } \mathcal{K}_t} \geq 0.$$

Equality holds if and only if $q_t^{\mathcal{K}} = \bar{p}_t$ and $Q_t = 1$.

Additionally, this decomposition reveals two complementary failure modes: the KL term captures misallocation of probability among the teacher's leading candidates, while $-\log Q_t$ measures how much student mass falls outside this set. This also justifies why we do not renormalize the student distribution in equation 3: renormalization would eliminate the coverage gap encoded by $-\log Q_t$, making the score blind to the very deficiency we aim to correct.

A large d_t signals that the student poorly covers the teacher's preferred continuations at that state. We therefore use d_t as a priority score and concentrate the deficit branch on the most severe positions. Within each batch, we retain the largest ρ^+ fraction of scores:

$$\tau^+ = \mathrm{Quantile}_{1-\rho^+}\big(\{d_j : j \in \mathcal{B}\}\big), \qquad m_t^+ = \mathbb{1}[d_t \geq \tau^+],$$

and denote the active set as $\mathcal{A}^+ = \{t : m_t^+ = 1\}$. For these selected positions, we do not rely on sampling to surface the missing tokens. Instead, we directly distill the teacher's renormalized top-K distribution:

$$\mathcal{L}_{\mathrm{def}} = -\mathbb{E}\left[\sum_t m_t^+ \sum_{v \in \mathcal{K}_t} \bar{p}_t(v) \log q_\theta(v|s_t)\right]. \tag{4}$$

This completes the blue route in Fig. 3: the deficit score selects positions in the bottom-center panel, and the analytic top-K objective performs probability-mass recovery in the bottom-right panel. The underlying principle is that student rollouts determine where to learn, but not which tokens to learn: the deficit branch uses the student's mass on the teacher's top-K set to detect missing support and directly distills those candidates when coverage is insufficient.

3.3 Joint Objective

The excess and deficit branches operate on the same student-generated states but address complementary forms of mismatch. The two routing decisions in Fig. 3 (bottom center) are independent: a position may activate the excess branch, the deficit branch, or both, while positions selected by neither branch receive no distillation update. We optimize the two branches jointly as

$$\mathcal{L}_{\mathrm{TIDE}} = \mathcal{L}_{\mathrm{exc}} + \lambda \mathcal{L}_{\mathrm{def}}, \tag{5}$$

where λ controls the strength of teacher-support recovery relative to excess suppression. Each branch contributes only at the positions selected by its corresponding gate; positions selected by neither branch receive no distillation update. Algorithm 1 summarizes the resulting training procedure.

4 Experiments

4.1 Experimental Setup

Models and baselines. We employ two teacher–student pairs with different degrees of model mismatch. The first uses JustRL-DeepSeek-1.5B (He et al., 2025) as the teacher and DeepSeek-R1-Distill-Qwen-1.5B (Guo et al., 2025) as the student, while the second uses Qwen3-8B as the teacher and Qwen3-1.7B-Base as the student (Yang et al., 2025). Following Li et al. (2026a), we characterize the initial teacher–student mismatch using the overlap between their top-K candidate sets on student-generated prefixes. A higher initial overlap indicates a smaller teacher–student mismatch, while a lower overlap indicates a larger mismatch. Under this criterion, the JustRL-DeepSeek-1.5B $\rightarrow$ DeepSeek-R1-Distill-Qwen-1.5B pair constitutes our *weak-mismatch* setting, while the Qwen3-8B $\rightarrow$ Qwen3-1.7B-Base pair constitutes our *strong-mismatch* setting.

Table 1: **Avg@8 and Pass@8 results across different teacher–student mismatch settings** (%). Each model generates eight responses per problem. A@8 denotes Avg@8 (mean accuracy over eight responses) and P@8 denotes Pass@8 (percentage of problems for which at least one of eight responses is correct). *Student* and *Teacher* denote reference models and are excluded from bolding. **Bold** denotes the best trained model in each column. Δ denotes the improvement of TIDE over OPD.

	AIME'24		AIME'25		AMC23		MATH		Minerva		Olymp.		BRUMO		CMIMC		HMMT		Avg.	
Method	A@8	P@8	A@8	P@8	A@8	P@8	A@8	P@8	A@8	P@8	A@8	P@8	A@8	P@8	A@8	P@8	A@8	P@8	A@8	P@8
Weak mismatch: JustRL-DeepSeek-1.5B → DeepSeek-R1-Distill-Qwen-1.5B																				
Student (Base)	27.1	36.7	24.2	33.3	73.1	95.0	83.4	95.0	28.0	43.4	44.3	60.8	27.1	56.7	14.4	30.0	12.1	20.0	37.1	52.3
Teacher	53.3	53.3	40.0	40.0	87.5	92.5	86.3	94.2	33.5	45.2	54.0	68.2	43.3	66.7	22.8	42.5	22.1	36.7	49.2	59.9
GRPO	22.5	43.3	21.2	36.7	72.5	90.0	83.9	94.4	29.5	45.6	46.3	64.8	33.3	63.3	16.6	32.5	14.2	30.0	37.8	55.6
OPD	43.8	76.7	32.1	53.3	82.2	95.0	86.2	94.6	**34.1**	44.9	51.6	66.8	40.4	56.7	21.9	35.0	19.2	33.3	45.7	61.8
FiRe-OPD (Li et al., 2026b)	39.2	**80.0**	**37.1**	**56.7**	**84.7**	95.0	85.3	94.2	33.2	44.1	52.3	68.0	42.5	60.0	21.6	35.0	**23.3**	**43.3**	46.6	64.0
PowerOPD (Zhao et al., 2026a)	**60.0**	60.0	31.2	43.3	81.6	**97.5**	84.0	94.4	32.6	**46.3**	50.6	67.2	38.3	60.0	19.7	32.5	20.4	33.3	46.5	59.4
AOPD (Jia et al., 2026)	50.8	73.3	33.3	43.3	81.9	95.0	85.4	**94.8**	33.0	46.0	51.3	67.4	41.2	63.3	20.6	40.0	17.9	26.7	46.2	61.1
TIDE	45.8	**80.0**	33.3	50.0	84.4	95.0	**86.7**	94.6	33.5	45.2	**52.5**	**68.1**	**42.9**	**66.7**	**22.2**	**42.5**	19.2	**43.3**	**46.7**	**65.0**
Δ vs. OPD	+2.0	+3.3	+1.2	−3.3	+2.2	0.0	+0.5	0.0	−0.6	+0.3	+0.9	+1.3	+2.5	+10.0	+0.3	+7.5	0.0	+10.0	+1.0	+3.2
Strong mismatch: Qwen3-8B → Qwen3-1.7B-Base																				
Student (Base)	3.3	3.3	0.0	0.0	13.4	40.0	23.8	68.2	6.3	24.3	9.8	34.3	5.0	20.0	0.0	0.0	0.0	0.0	6.9	21.1
Teacher	23.3	33.3	23.3	43.3	68.8	92.5	83.3	94.2	28.0	39.7	49.8	67.7	29.2	60.0	11.6	32.5	11.7	26.7	36.6	54.4
GRPO	3.3	6.7	0.4	3.3	35.0	60.0	55.5	83.4	18.7	36.0	23.3	45.5	9.6	23.3	0.9	7.5	0.0	0.0	16.3	29.5
OPD	2.9	13.3	1.2	13.3	10.3	45.0	26.5	73.8	6.9	28.7	10.7	37.7	2.9	13.3	0.3	2.5	0.0	0.0	6.9	25.3
FiRe-OPD (Li et al., 2026b)	3.3	3.3	4.2	10.0	31.9	55.0	52.2	77.6	14.2	32.7	19.7	40.9	7.1	16.7	0.9	5.0	**0.8**	**3.3**	14.9	27.2
PowerOPD (Zhao et al., 2026a)	6.7	6.7	4.2	10.0	26.2	57.5	55.1	83.6	15.1	37.1	24.5	47.2	7.1	23.3	1.9	**12.5**	0.4	**3.3**	15.7	31.2
AOPD (Jia et al., 2026)	3.3	3.3	**4.6**	**16.7**	34.4	**65.0**	62.3	86.0	21.2	40.4	25.5	47.8	**10.8**	**33.3**	1.9	10.0	0.4	**3.3**	18.3	34.0
TIDE	**10.0**	**16.7**	**4.6**	10.0	**39.4**	60.0	**65.1**	**86.2**	**21.8**	**41.9**	**28.6**	**50.3**	10.4	26.7	**2.8**	**12.5**	0.4	**3.3**	**20.3**	**34.2**
Δ vs. OPD	+7.1	+3.4	+3.4	−3.3	+29.1	+15.0	+38.6	+12.4	+14.9	+13.2	+17.9	+12.6	+7.5	+13.4	+2.5	+10.0	+0.4	+3.3	+13.4	+8.9

We compare TIDE against GRPO (Shao et al., 2024), standard OPD (Lu & Lab, 2025), FiRe-OPD (Li et al., 2026b), PowerOPD (Zhao et al., 2026a), and AOPD (Jia et al., 2026). PowerOPD applies a bounded power transformation to stabilize the sampled-token reward, whereas AOPD preserves positive reinforcement while replacing non-positive updates with localized divergence minimization. We additionally report the original teacher and student models as reference points.

Training Details. We train all methods for one epoch on DAPO-Math-17K (Yu et al., 2026) using the verl framework (Sheng et al., 2024). Each prompt produces four on-policy student responses, and all methods use the same student initialization, data order, rollout budget, and optimizer configuration. For TIDE, we set $K = 16$, $\rho^- = \rho^+ = 0.2$, and $\lambda = 1.0$. Full training, implementation, baseline, and evaluation details are provided in Section E.

4.2 Main Results

Overall Performance. Table 1 reveal a clear mismatch-dependent pattern. Under weak teacher–student mismatch, the distillation methods achieve broadly comparable Avg@8, while TIDE obtains the highest overall Avg@8 and Pass@8 of 46.7% and 65.0%, improving over OPD by 1.0 and 3.2 points, respectively. The differences become substantially larger under strong mismatch. Standard OPD falls to the original student's Avg@8 level of 6.9%, whereas TIDE reaches 20.3%, outperforming OPD, FiRe-OPD, GRPO, PowerOPD, and AOPD by 13.4, 5.4, 4.0, 4.6, and 2.0 points, respectively. For Pass@8, TIDE obtains 34.2%, compared with 25.3% for OPD, 27.2% for FiRe-OPD, 29.5% for GRPO, 31.2% for PowerOPD, and 34.0% for AOPD. Moreover, TIDE achieves the best-or-tied Avg@8 on seven of nine benchmarks and the best-or-tied Pass@8 on six of nine benchmarks under strong mismatch. These results are consistent with our analysis: sampled-token corrections are often sufficient when the teacher and student distributions substantially overlap, but analytic recovery of teacher-preferred alternatives becomes increasingly important as mismatch grows and the student's sampled support becomes less informative.

Generation Stability. Table 2 shows that the accuracy gap is accompanied by substantially different generation behavior. Under strong mismatch, OPD and FiRe-OPD produce 22,395- and 28,307-token responses, with 65.5% and 54.0% missing a boxed answer; PowerOPD reduces but does not eliminate this degeneration, averaging 13,972 tokens with a 26.3% missing-answer rate. In contrast, TIDE achieves the highest Avg@8 of 20.3% while reducing response length to 7,294 tokens and the missing-answer rate to 5.4%. This gain cannot be attributed to shorter responses alone: GRPO and

Table 2: **Generation diagnostics.** Length: macro-avg. tokens over nine benchmarks; Miss.: % responses without a boxed answer. **Bold** = lowest missing rate.

	Weak		Strong	
Method	Len.	Miss.%	Len.	Miss.%
GRPO	8,048	**1.8**	5,261	17.1
OPD	7,657	4.2	22,395	65.5
FiRe-OPD	8,055	4.2	28,307	54.0
PowerOPD	8,658	5.3	13,972	26.3
AOPD	7,982	4.8	5,894	**4.7**
TIDE	7,571	3.7	7,294	5.4

Table 3: **Component ablation (strong mismatch).** S/H/D: selection, Hellinger shaping, deficit recovery. Macro-avg. over nine benchmarks; length in K tokens.

Method	Comp.	Avg@8	Pass@8	Len.
Sampled OPD	–	6.87	25.29	22.4
Selection only	S	14.58	28.16	29.7
Excess branch	S+H	18.57	31.72	23.6
Deficit branch	D	18.32	32.71	13.9
TIDE	S+H+D	**20.34**	**34.17**	7.3

Table 4: **Sensitivity to the deficit-recovery weight** λ with fixed gates $\rho^- = \rho^+ = 0.2$. Avg@8, Pass@8, response length, and Distinct-4 are macro-averaged over nine benchmarks; format errors count responses without a boxed final answer. Shaded row: default.

λ	Avg@8↑	Pass@8↑	Len.↓	Fmt. Err.↓	Dist.-4↑
0.25	20.1	32.6	26.9K	5,510	0.145
0.5	20.4	36.7	20.0K	2,488	0.128
1.0	20.3	34.2	7.3K	706	0.305
2.0	19.2	35.4	5.1K	482	0.417

AOPD generate even shorter outputs but remain 4.0 and 2.0 Avg@8 points behind TIDE, respectively.

4.3 Ablation Studies

Component Ablation. Table 3 shows that the three components play complementary roles. Selection alone raises Avg@8 from 6.87 to 14.58, confirming the value of concentrating supervision on informative positions. However, it also increases the macro-averaged response length from 22.4K to 29.7K tokens, indicating that selection alone does not prevent generation degeneration. Applying bounded Hellinger shaping to the selected excess positions further improves Avg@8 to 18.57 and partially controls the length increase, reducing it to 23.6K. Analytic deficit recovery independently reaches a similar Avg@8 of 18.32, while producing substantially shorter responses of 13.9K tokens. Combining all three components in TIDE achieves the best Avg@8 and Pass@8 of 20.34 and 34.17, respectively, while further reducing response length to 7.3K tokens. Overall, selection determines *where* to learn, Hellinger shaping stabilizes excess suppression, and deficit recovery restores teacher-preferred continuations that student sampling tends to miss.

Sensitivity to Hyperparameters. Table 4 varies the deficit-recovery weight λ with fixed gates $\rho^- = \rho^+ = 0.2$. Accuracy is stable around the default: $\lambda \in \{0.25, 0.5, 1.0\}$ all reach Avg@8 between 20.1 and 20.4. Generation behavior, however, differs substantially. Reducing λ from 1.0 to 0.25 inflates the macro-averaged response length from 7.3K to 26.9K tokens and format errors from 706 to 5,510, and Distinct-4 drops from 0.305 to 0.145, indicating that insufficient deficit recovery lets long, repetitive continuations persist. Increasing λ to 2.0 further shortens responses and raises Distinct-4 to 0.417, but lowers Avg@8 to 19.2. We therefore use $\lambda = 1.0$ as a balanced default; sensitivity to the gate keep-rate is reported in Table 7.

5 Conclusion

We identify *degenerate agreement*, a failure mode of on-policy distillation in which local token-level matching masks globally flawed responses. Teacher–student mismatch is directional: student-excess tokens induce unstable corrections, whereas valuable student-deficit tokens are rarely sampled. TIDE addresses both with bounded excess correction and analytic deficit recovery, outperforming strong baselines on nine mathematical reasoning benchmarks (by 13.4 Avg@8 points over OPD under severe mismatch) with shorter, better-formed responses. TIDE assumes a locally reliable teacher, is evaluated only on two teacher–student pairs in mathematical reasoning, and its fixed-state theory gives no global convergence guarantee. More broadly, on-policy distillation should allocate supervision by both the *direction* and *accessibility* of disagreement.

AI USE STATEMENT

In this work, we used generative AI tools to improve the grammar, clarity, and readability of the manuscript and to assist with code debugging. Our use of generative AI was limited to these tasks. We did not use generative AI tools to develop research ideas, design the methodology or experiments, formulate or prove mathematical claims, generate or analyze data, or interpret results. We have reviewed all AI-assisted work. Specifically, we reviewed the language revisions to ensure that they preserved the intended meaning and technical accuracy, and we manually inspected and tested the code changes suggested during debugging. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we have made the source code for this work publicly available through an anonymous repository at https://anonymous.4open.science/r/TIDE-B1B2. The released code is provided to support independent verification and reproduction of the results reported in this paper.

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.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

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.

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.

Bingxiang He, Zekai Qu, Zeyuan Liu, Yinghao Chen, Yuxin Zuo, Cheng Qian, Kaiyan Zhang, Weize Chen, Chaojun Xiao, Ganqu Cui, et al. Justrl: Scaling a 1.5 b llm with a simple rl recipe. *arXiv preprint arXiv:2512.16649*, 2025.

Nan Jia, Haojin Yang, Xing Ma, Jiesong Lian, Shuailiang Zhang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Zequn Sun. Asymmetric on-policy distillation: Bridging exploitation and imitation at the token level. *arXiv preprint arXiv:2605.06387*, 2026.

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

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *Forty-third International Conference on Machine Learning*, 2026.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam H Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. In *Workshop on Failure Modes of Agentic AI at ICML 2026*, 2026.

Junlong Ke, Zichen Wen, Weijia Li, Conghui He, and Linfeng Zhang. Respecting self-uncertainty in on-policy self-distillation for efficient llm reasoning. *arXiv preprint arXiv:2605.13255*, 2026.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

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

Taeckyung Lee, Jeonghye Kim, Hyungjun Yoon, and Sung-Ju Lee. Dense supervision but sparse guidance: A token-level analysis of on-policy self-distillation. In *3rd AI for Math Workshop: Toward Self-Evolving Scientific Agents*, 2026.

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*, 2026a.

Yuying Li, Leqi Zheng, Yongzi Yu, Wenrui Zhou, Xuchang Zhong, Xing Hu, Jing Jin, Hangjie Yuan, and Tao Feng. Filter, then reweight: Rethinking optimization granularity in on-policy distillation. *arXiv preprint arXiv:2606.02684*, 2026b.

Baohao Liao, Hanze Dong, Christof Monz, Xinxing Xu, Li Dong, and Furu Wei. Multi-turn on-policy distillation with prefix replay. *arXiv preprint arXiv:2607.04763*, 2026.

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

Yanjiang Liu, Jie Lou, Xinyan Guan, Yuqiu Ji, Hongyu Lin, Ben He, Xianpei Han, Le Sun, Xing Yu, and Yaojie Lu. Your teacher can't help you here: Combating supervision fidelity decay in on-policy distillation. *arXiv preprint arXiv:2605.30833*, 2026b.

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.

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.

Tom Minka et al. Divergence measures and message passing. 2005.

Minjae Oh, Sangjun Song, Gyubin Choi, Yunho Choi, and Yohan Jo. Kl for a kl: On-policy distillation with control variate baseline. *arXiv preprint arXiv:2605.07865*, 2026.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

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

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

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. *arXiv preprint arXiv: 2409.19256*, 2024.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026.

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.

Haoran Xin, Anhao Zhao, Ying Sun, Jin Li, Xiaoyu Shen, and Hui Xiong. Escaping the kl agreement trap in on-policy distillation. *arXiv preprint arXiv:2606.09471*, 2026.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation. *arXiv preprint arXiv:2606.01249*, 2026.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026.

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.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Haobo Wang, Mingxuan Xia, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, Junbo Zhao, et al. Oprd: On-policy representation distillation. *arXiv preprint arXiv:2606.06021*, 2026a.

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*, 2026b.

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.

Anhao Zhao, Junlong Tong, Yingqi Fan, Ping Nie, Wenjie Li, and Xiaoyu Shen. Poweropd: Stabilizing on-policy distillation with bounded power transformation. *arXiv preprint arXiv:2606.17199*, 2026a.

Qingfei Zhao, Huan Song, Shuyu Tian, Jiawei Shao, and Xuelong Li. Prefix-guided on-policy distillation: Mining golden trajectories from rollouts. *arXiv preprint arXiv:2606.21994*, 2026b.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *arXiv preprint arXiv:2605.11182*, 2026.

CONTENTS

1 Introduction 1

2 Problem Setup 4

3 Methodology 5

4 Experiments 7

5 Conclusion 9

A Proofs and Additional Theoretical Details 14

B Full Training Procedure 15

C Related Work 15

D Additional Experiments 16

E Experimental Details 19

F Case Studies 21

Roadmap. The appendix is organized as follows:

- Appendix A provides the proofs of the two propositions and additional theoretical details.
- Appendix B presents the complete training procedure for TIDE.
- Appendix C reviews related work on on-policy distillation and selective supervision.
- Appendix D reports additional experiments on teacher continuation, matched-token supervision, student-deficit tokens, and gate keep-rate sensitivity.
- Appendix E documents the training setup, hyperparameters, baseline implementations, evaluation metrics, and checkpoint selection.
- Appendix F presents qualitative comparisons of standard OPD and TIDE, illustrating representative generation failures and their mitigation.

A Proofs and Additional Theoretical Details

This appendix proves the two propositions in Section 3.

Notation and assumptions. We work at a fixed rollout position t with student-visited state $s = s_t$. We abbreviate $p(v) = p(v|s)$ and $q(v) = q_\theta(v|s)$, and write $a(v) = \log[p(v)/q(v)]$ and $h(a) = 2(e^{a/2} - 1)$. We use the normalized squared Hellinger distance

$$H^2(p, q) = 1 - \sum_{v \in \mathcal{V}} \sqrt{p(v)q(v)}.$$

For the deficit branch, $\mathcal{K}_t$, $\bar{p}_t$, Q_t, and $q_t^{\mathcal{K}}$ follow the definitions in Section 3.2. We assume finite-vocabulary softmax distributions with full support. When differentiating a gated loss, its routing decisions are treated as fixed.

A.1 Proof of Proposition 1

Proof. For $a < 0$, $e^{a/2} \in (0, 1)$, so $-2 < h(a) < 0$. The local expansion follows from the first-order expansion of the smooth function h at zero, with $h(0) = 0$ and $h'(0) = 1$.

By the definition of H^2,

$$\nabla_\theta H^2(p, q) = -\frac{1}{2} \sum_{v \in \mathcal{V}} \sqrt{\frac{p(v)}{q(v)}} \nabla_\theta q(v).$$

Using $\sum_v \nabla_\theta q(v) = 0$, we obtain

$$\begin{aligned}
\mathbb{E}_{v \sim q}[-h(a(v)) \nabla_\theta \log q(v)] &= -2 \sum_v \left(\sqrt{\frac{p(v)}{q(v)}} - 1 \right) \nabla_\theta q(v) \\
&= -2 \sum_v \sqrt{\frac{p(v)}{q(v)}} \nabla_\theta q(v) \\
&= 4 \nabla_\theta H^2(p, q),
\end{aligned}$$

which proves Equation (2). □

A.2 Proof of Proposition 2

Proof. By the definitions in Proposition 2, $q(v) = Q_t q_t^{\mathcal{K}}(v)$ for $v \in \mathcal{K}_t$. Substituting this identity into Equation (3) gives

$$d_t = \sum_{v \in \mathcal{K}_t} \bar{p}_t(v) \log \frac{\bar{p}_t(v)}{Q_t q_t^{\mathcal{K}}(v)} = D_{\mathrm{KL}}\big(\bar{p}_t \,\|\, q_t^{\mathcal{K}}\big) - \log Q_t.$$

Both terms are nonnegative. Their sum is zero exactly when $q_t^{\mathcal{K}} = \bar{p}_t$ and $Q_t = 1$, proving the claim. □

Algorithm 1: TIDE Training

Input: prompt distribution $\mathcal{D}$, student q_θ, teacher p, top-K size K, keep ratios ρ^-, ρ^+, recovery weight λ, learning rate η
Output: Trained student parameters θ

while *not converged* **do**
  Sample a prompt batch $x \sim \mathcal{D}$ and generate $o \sim q_\theta(\cdot|x)$
  Let $\mathcal{B}$ be all valid response-token positions and $N = |\mathcal{B}|$
  foreach $t \in \mathcal{B}$ **do**
    Construct $s_t = (x, o_{<t})$ and compute $a_t \leftarrow \log p(o_t|s_t) - \log q_\theta(o_t|s_t)$, $e_t \leftarrow [-a_t]_+$
    Obtain $\mathcal{K}_t \leftarrow \mathrm{TopK}_v\, p(v|s_t)$ and normalize $\bar{p}_t(v) \leftarrow p(v|s_t) / \sum_{u \in \mathcal{K}_t} p(u|s_t)$
    Compute $d_t \leftarrow \sum_{v \in \mathcal{K}_t} \bar{p}_t(v) \log \frac{\bar{p}_t(v)}{q_\theta(v|s_t)}$
  Set $\tau^- \leftarrow Q_{1-\rho^-}(\{e_t : t \in \mathcal{B}\})$ and $\tau^+ \leftarrow Q_{1-\rho^+}(\{d_t : t \in \mathcal{B}\})$
  Select $\mathcal{A}^- \leftarrow \{t : a_t < 0,\ e_t \geq \tau^-\}$ and $\mathcal{A}^+ \leftarrow \{t : d_t \geq \tau^+\}$
  Compute

$$\mathcal{L}_{\mathrm{exc}} \leftarrow -\frac{1}{N} \sum_{t \in \mathcal{A}^-} \mathrm{sg}\Big[2(e^{a_t/2} - 1)\Big] \log q_\theta(o_t|s_t)$$

  Compute

$$\mathcal{L}_{\mathrm{def}} \leftarrow -\frac{1}{N} \sum_{t \in \mathcal{A}^+} \sum_{v \in \mathcal{K}_t} \bar{p}_t(v) \log q_\theta(v|s_t)$$

  Set $\mathcal{L}_{\mathrm{TIDE}} \leftarrow \mathcal{L}_{\mathrm{exc}} + \lambda \mathcal{L}_{\mathrm{def}}$
  Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\mathrm{TIDE}}$
return θ

B Full Training Procedure

Algorithm 1 summarizes the complete TIDE training loop described in Section 3. Each iteration samples on-policy responses from the student, scores all valid token positions with a single teacher forward pass, and selects the excess set $\mathcal{A}^-$ and the deficit set $\mathcal{A}^+$ via batch-level quantile thresholds, so that ρ^- and ρ^+ directly control the fraction of tokens receiving each loss. $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator, $[\cdot]_+ = \max(\cdot, 0)$, and $Q_{1-\rho}(\cdot)$ is the $(1-\rho)$-quantile of a set of scalars.

C Related Work

C.1 Mechanistic Understanding and Failure Modes of OPD

On-policy distillation (OPD) trains a student on its own trajectories while using a teacher to provide dense token-level supervision at the states the student visits (Agarwal et al., 2024; Lu & Lab, 2025). By aligning training with the student's inference-time state distribution, OPD reduces the exposure bias associated with distillation from fixed demonstrations (Bengio et al., 2015; Ross et al., 2011; Gu et al., 2024). Recent work has moved beyond demonstrating the empirical effectiveness of OPD toward explaining its underlying behavior. In particular, Li et al. (2026a) study its phenomenology and optimization mechanism, emphasizing the role of teacher–student overlap, while Zhu et al. (2026) systematically examine the conditions under which OPD succeeds or fails. Other analyses identify limitations of output-space supervision, including a deteriorating signal-to-noise ratio as the student approaches the teacher (Yang et al., 2026a).

A growing body of work further shows that a locally well-defined teacher–student objective does not necessarily imply globally desirable rollouts. OPD may exhibit abrupt length inflation and repetitive generation (Luo et al., 2026); its KL-based signal can enter an agreement trap (Xin et al., 2026); and teacher supervision can lose fidelity on prefixes induced by the student (Liu et al., 2026b). Related failures have also been observed when privileged teacher context changes the learning signal for long reasoning traces (Kaur et al., 2026). Our work builds on this mechanistic perspective. We identify a concrete form of *degenerate agreement*: a student-generated repetitive prefix can make the conditioned teacher favor the same continuation, producing near-zero local discrepancy even though the complete response is degenerate. This observation motivates treating teacher–student agreement as potentially uninformative rather than automatically desirable. More broadly, it changes

what selective supervision should preserve: instead of favoring overlap itself, token selection should concentrate on mismatches for which the teacher still provides corrective information.

C.2 Selective and Reliable Supervision in OPD

Dense token-level feedback does not imply that every token provides useful guidance (Lee et al., 2026). Existing methods therefore select or reweight OPD supervision using position, discrepancy, importance, teachability, or uncertainty. Prefix-OPD retains supervision only over selected reasoning prefixes (Zhang et al., 2026). At the token level, recent approaches filter and reweight supervision (Li et al., 2026b), estimate token importance (Xu et al., 2026), restrict updates to learnable disagreement (Wang et al., 2026), or use entropy and self-uncertainty to modulate individual tokens (Jin et al., 2026; Ke et al., 2026). Relaxed OPD similarly avoids enforcing equally strict alignment at every position (Ko et al., 2026), while trajectory- and prefix-oriented methods mine or refine more reliable training paths (Zhao et al., 2026b; Jiang et al., 2026; Liao et al., 2026). Collectively, these studies establish that uniformly distilling every student-generated token is neither necessary nor consistently beneficial.

Most selective methods, however, still decide which signals to retain on student-generated trajectories and optimize the tokens the student has already sampled. Better selection can therefore improve the quality of observed supervision without resolving a distinct coverage problem: teacher-preferred alternatives with negligible student probability rarely appear in the rollout and consequently provide little or no direct corrective signal.

Complementary work modifies the optimization procedure to improve reliability, including asymmetric token-level objectives (Jia et al., 2026), trust-region constraints (Xing et al., 2026), teacher-guided optimization under large policy divergence (Liu et al., 2026a), and control-variate estimators for reducing OPD gradient variance (Oh et al., 2026). These methods improve the reliability of updates under observed disagreement, but severe teacher-side deficits remain the least likely discrepancies to be observed through student sampling. TIDE addresses selection and coverage jointly by using the *direction* of mismatch to determine the correction mechanism. It removes matched tokens, applies bounded suppression to sampled student-excess tokens, and analytically recovers teacher-preferred mass that student rollouts are unlikely to sample. Thus, its allocation rule follows the distinct observability and optimization requirements of excess and deficit, rather than applying a single token score or a uniform modification of the OPD objective.

D Additional Experiments

D.1 Teacher Continuation from Repetitive Student Prefixes

The preceding analysis shows that repetitive trajectories emerge naturally from the student during OPD. We next ask whether the teacher corrects such behavior or instead inherits it from the student-generated context. To isolate the teacher's response, we perform a continuation experiment in which the student controls the prefix but the teacher controls all subsequent tokens.

We first identify naturally occurring student responses in which at least 30% of the tokens belong to a detected repetitive region. For each response, we retain the original prompt and truncate the student trajectory 200 tokens after the onset of repetition. We then condition the Qwen3-8B teacher on this prompt–prefix pair and greedily decode 256 additional tokens. The student and teacher continuations therefore begin from exactly the same student-generated state, allowing us to test whether the repetitive behavior persists after control is transferred to the teacher.

We measure persistence by comparing local token patterns in the teacher continuation with those in the tail of the student prefix. Let $\mathcal{G}_8$ denote the set of contiguous eight-token sequences appearing in the teacher continuation, and let $\mathcal{P}_8$ denote the corresponding set from the final 256 tokens of the student prefix. Their overlap is measured as

$$C = \frac{|\mathcal{G}_8 \cap \mathcal{P}_8|}{\max(|\mathcal{G}_8|, 1)}.$$

A value of $C = 1$ means that every local pattern in the teacher continuation already occurs in the repetitive student prefix. We classify the teacher as preserving the loop when $C \geq 0.8$; using

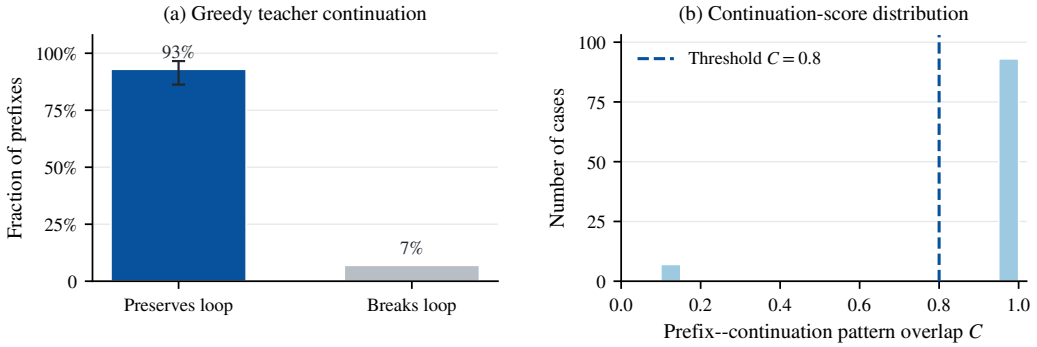

Figure 5: Teacher continuation from naturally occurring repetitive student prefixes. **(a)** Greedy teacher decoding preserves the student-induced loop in 93% of cases; the error bar denotes a 95% Wilson confidence interval. **(b)** Distribution of the prefix–continuation pattern overlap C, defined as the fraction of contiguous eight-token sequences in the teacher continuation that also appear in the student-prefix tail. The dashed line marks the preservation threshold $C = 0.8$.

eight-token sequences avoids counting short, accidental token overlaps as continuation of the same loop.

As shown in Fig. 5, the teacher preserves the student-induced repetitive pattern in 93% of cases. The result is also well separated from the decision threshold: preserved continuations concentrate at $C = 1$, indicating near-exact reproduction of the local patterns established by the student. Thus, repetition is not merely a failure of student-side sampling. Once the student drives the trajectory into a repetitive state, the teacher typically continues from that state in the same way rather than providing a corrective target. This explains how a student-generated loop can become self-reinforcing under on-policy distillation.

D.2 Does Supervision on Matched Tokens Help?

Our method deliberately assigns zero weight to positions where the teacher and student approximately agree. We examine whether these omitted positions still provide useful supervision, and whether the benefit of mismatch selection could instead be explained simply by updating fewer tokens.

We construct a matched-only control by reversing the allocation principle. Within each training batch, we rank valid response positions according to the absolute sampled-token residual

$$|R_t| = |\log p(o_t \mid s_t) - \log q_\theta(o_t \mid s_t)| \,.$$

We discard the largest 40% of residuals and apply the standard OPD log-ratio advantage only to the lowest-residual 60%. All other positions receive zero distillation weight, and the teacher top-K deficit branch is disabled. The training data and its order, student initialization, rollout budget, optimizer, and training length are otherwise unchanged. Table 5 reports the complete per-benchmark results and aggregate generation statistics for both teacher–student pairs.

Weak teacher–student mismatch. On JustRL-1.5B $\rightarrow$ R1-Distill-1.5B, matched-only training improves Avg@8 from 37.08 to 42.23. Low-mismatch positions are therefore not entirely uninformative when the teacher and student are already closely aligned. Nevertheless, matched-only remains 4.51 points below TIDE and reaches only 56.44 Pass@8, compared with 65.00 for TIDE, despite supervising a larger fraction of positions. Thus, even in this setting, matched positions provide a lower return per supervised token.

Mismatch selection therefore more than doubles the matched-only accuracy while supervising only one third as many positions. Adding bounded excess suppression and deficit recovery further increases Avg@8 to 20.34 and Pass@8 to 34.17, while reducing the macro-averaged response length from $29{,}655$ to $7{,}294$ tokens and format errors from $6{,}845$ to 706.

Table 5: Complete matched-token control results. The upper panel reports Avg@8 / Pass@8 on each benchmark and their nine-benchmark macro-average. In the lower panel, response length and Distinct-4 are also macro-averaged over the nine benchmarks, while the remaining diagnostics are totals. All evaluations use eight responses per problem.

Pair	Method	AMC23	AIME24	AIME25	MATH-500	Minerva	Olympiad	BRUMO25	CMIMC25	HMMT25	Macro Avg.
JustRL-1.5B → R1-Distill-1.5B	Student	73.12/95.00	27.08/36.67	24.17/33.33	83.43/95.00	28.03/43.38	44.31/60.83	27.08/56.67	14.37/30.00	12.08/20.00	37.08/52.32
	Matched-only	76.56/95.00	31.67/46.67	31.67/36.67	85.72/94.60	30.70/44.85	50.17/65.13	36.25/56.67	18.12/35.00	19.17/33.33	42.23/56.44
	TIDE	84.38/95.00	45.83/80.00	33.33/50.00	86.72/94.60	33.55/45.22	52.54/68.10	42.92/66.67	22.19/42.50	19.17/43.33	**46.74/65.00**
Qwen3-8B → Qwen3-1.7B-Base	Student	13.44/40.00	3.33/3.33	0.00/0.00	23.82/68.20	6.34/24.26	9.85/34.27	5.00/20.00	0.00/0.00	0.00/0.00	6.87/21.12
	OPD	10.31/45.00	2.92/13.33	1.25/13.33	26.52/73.80	6.94/28.68	10.66/37.69	2.92/13.33	0.31/2.50	0.00/0.00	6.87/25.29
	Matched-only	12.81/27.50	2.92/3.33	1.25/6.67	26.12/71.60	8.23/26.84	11.03/36.50	1.25/10.00	0.62/2.50	0.42/3.33	7.18/20.92
	Mismatch selection	26.56/52.50	3.33/13.33	6.25/16.67	52.15/77.60	15.58/30.88	21.01/40.80	5.00/13.33	0.94/5.00	0.42/3.33	14.58/28.16
	TIDE	39.38/60.00	10.00/16.67	4.58/10.00	65.08/86.20	21.83/41.91	28.58/50.30	10.42/26.67	2.81/12.50	0.42/3.33	**20.34/34.17**

Pair	Method	Length ↓	Dist.-4 ↑	Format Err. ↓	Solve-none ↓	Solve-all ↑
Weak	Student	12,543	0.563	803	549	550
	Matched-only	8,494	0.568	612	508	658
	TIDE	7,571	0.566	488	472	711
Strong	Student	13,465	0.099	7,559	985	2
	OPD	22,395	0.062	8,623	920	0
	Matched-only	12,702	0.101	7,237	950	3
	Mismatch selection	29,655	0.104	6,845	862	90
	TIDE	7,294	0.305	706	717	286

Table 6: Student-deficit tokens are difficult to access through sampled supervision but carry substantial learning signal on Qwen3-8B → Qwen3-1.7B-Base. **(a)** Accessibility under four independent draws from the same state. **(b)** End-to-end results macro-averaged over nine benchmarks with $n = 8$ samples per problem.

(a) Accessibility diagnostic

Statistic	Value
Deficit positions among non-loop tokens	1.32%
Median student probability ϵ_t	2.04×10^{-4}
Expected draws $1/\epsilon_t$	4,909
Median $P_{\text{obs}}(4)$	8.15×10^{-4}
Positions with $P_{\text{obs}}(4) < 1\%$	95.5%

(b) End-to-end value

Method	Avg@8	Pass@8	Dist.-4
Student	6.87	21.12	0.099
Deficit-only	18.32	32.71	**0.414**
TIDE	**20.34**	**34.17**	0.305

D.3 Accessibility and Value of Student-Deficit Tokens

We define a student-deficit position as a non-loop position where the teacher's top-1 token $v_t^\star$ lies outside the student's top-16 support. Writing its student probability as $\epsilon_t = q_\theta(v_t^\star | s_t)$, a sampled-token estimator observes it within N independent draws from the same state with probability

$$P_{\text{obs}}(N) = 1 - (1 - \epsilon_t)^N,$$

and requires $1/\epsilon_t$ draws in expectation.

To test whether this inaccessible mass is useful, we train a deficit-only variant with $\rho^- = 0$, $\rho^+ = 0.2$, $K = 16$, and $\lambda = 1$. It applies analytic cross-entropy guidance only at positions with the largest teacher-support deficits; all other settings remain unchanged. Table 6 summarizes both the accessibility diagnostic and the resulting end-to-end performance.

As shown in Table 6, 95.5% of deficit positions have less than a 1% chance of being observed within four draws, yet the deficit-only branch raises Avg@8 from 6.87 to 18.32. Thus, these tokens are not an inconsequential probability tail: they contain useful teacher-supported modes that sampled-token OPD is unlikely to recover within a practical rollout budget.

D.4 Sensitivity to the Gate Keep-Rate

We jointly vary the excess and deficit keep-rates, $\rho^- = \rho^+ = \rho$, while holding all other hyperparameters fixed. Table 7 reports the results on Qwen3-8B → Qwen3-1.7B-Base.

Table 7: Sensitivity to the gate keep-rate $\rho^- = \rho^+ = \rho$. Avg@8 and response length are macro-averaged over nine benchmarks with $n = 8$; length is reported in thousands of tokens. Fmt. err. denotes the total number of responses without a parseable final answer.

ρ	Avg@8	Length	Fmt. err.
0.10	19.53	10.3K	1,111
0.20	20.34	7.3K	706
0.30	19.38	11.7K	620
0.50	20.93	6.8K	433

As shown in Table 7, TIDE remains stable across a broad range of keep-rates: Avg@8 varies by only 1.55 points over $\rho \in [0.1, 0.5]$, with no monotonic degradation as more positions are retained. We use $\rho = 0.2$ throughout the main experiments because it was fixed before benchmark evaluation and retains only a sparse subset of positions, rather than selecting the best-performing value retrospectively.

E Experimental Details

E.1 Additional Training and Implementation Details

All methods share the training setup described in the main text; here we document the remaining implementation choices. During rollout generation, the teacher scores the student's responses with temperature 1.0 (no logit rescaling), and teacher log-probabilities are computed on the identical token sequences produced by the student, so the sampled residual $R_t = \log p(o_t \mid s_t) - \log q_\theta(o_t \mid s_t)$ is exact rather than re-estimated. We do not use an auxiliary KL penalty toward a reference policy, a format reward, or a repetition penalty during training; the distillation signal is the only learning signal for all OPD-style methods. The learning rate is held constant (no warmup or decay). Token-level advantages are detached before entering the policy-gradient loss, and the loss is aggregated with token-mean normalization over all valid response tokens in the batch; because training is fully on-policy (a single gradient step per rollout batch), the importance ratio is identically 1 and no PPO-style clipping is active. For TIDE, the quantile thresholds τ^- and τ^+ are recomputed on every training batch over all valid response positions, so the keep-rates ρ^- and ρ^+ are batch-relative rather than fixed absolute thresholds. The deficit branch adds one student forward pass over the teacher's top-K candidate ids per position, a cost comparable to top-K OPD.

Training uses fully sharded data parallelism (Zhao et al., 2023) with gradient checkpointing and activation offloading; rollouts are generated with vLLM (Kwon et al., 2023) using dynamic batching.

E.2 Hyperparameters

Table 8 reports the shared training and evaluation configuration together with the method-specific hyperparameters. Unless explicitly varied in an ablation, the same configuration is used for all teacher–student pairs and baselines.

E.3 Baseline Implementations

All baselines are implemented in the same codebase and trained with the same data order, student initialization, rollout budget (4 responses per prompt), sequence lengths, optimizer, and hardware as TIDE. Unless otherwise stated, they differ only in how token-level advantages are constructed.

OPD. Standard on-policy distillation applies the sampled-token log-ratio advantage

$$A_t = \log p(o_t \mid c_t) - \log q_\theta(o_t \mid c_t)$$

to every valid response token, without token selection or reward reshaping.

GRPO. The teacher-free reinforcement-learning baseline replaces teacher supervision with the rule-based exact-match reward used for grading and computes group-normalized outcome advantages over the 4 responses sampled for each prompt. All other training settings are unchanged, so GRPO uses the same rollout and gradient budget as the distillation methods.

FiRe-OPD. We reimplement FiRe-OPD (Li et al., 2026b) following its official implementation and verify parity against it. Trajectories are ranked by length-normalized teacher log-likelihood, and the bottom 20% are discarded. On the retained trajectories, the sampled-token log-ratio advantages are multiplied by detached entropy-based weights, with both the teacher-confidence and student-confusion coefficients set to 1.0. FiRe-OPD operates entirely on student-sampled tokens and does not access the teacher's top-K distribution.

PowerOPD. We implement PowerOPD using its Box–Cox reward transformation. For a sampled token with teacher probability p_t and student probability q_t, the log-ratio advantage is replaced by

$$A_t^{\text{Power}} = p_t^{\alpha} - q_t^{\alpha},$$

where the token-independent $1/\alpha$ factor is absorbed into the learning rate. We use $\alpha = 5$, following the selected configuration in our sweep. This bounded, sign-preserving reward is applied to every valid sampled token; PowerOPD performs neither token selection nor teacher-support expansion.

AOPD. We implement AOPD with its probability-gap intervention rule. At each position, the gate compares the teacher and student probabilities of the sampled token,

$$G_t = \mathbb{1}[p_t - q_t \le \tau],$$

with $\tau = 0$. Positions with $G_t = 0$ retain the standard sampled-token OPD advantage, whereas positions with $G_t = 1$ replace it with forward-KL guidance over the teacher's top-K support. We use $K = 16$, renormalize the teacher probabilities within this support, and do not apply an additional $1/K$ scaling factor.

Teacher and student references. The reference rows report the teacher and the untrained student evaluated with the same decoding and grading pipeline used for all trained methods.

E.4 Evaluation and Metric Definitions

Decoding. Every model is evaluated with the same pipeline: vLLM generation with temperature 0.7, top-$p = 0.95$, $n = 8$ samples per problem, and a maximum generation length of $31{,}744$ tokens. All problems use the prompt template

```
{problem} Please reason step by step, and put your final
answer within \boxed{}.
```

with the model's chat template and thinking mode disabled.

Grading. A response is graded by extracting the final `\boxed{}` expression and checking mathematical equivalence against the reference answer with a rule-based verifier (symbolic-equivalence normalization for fractions, radicals, and equivalent forms). A response without a parseable boxed answer is counted as incorrect. No LLM-based verifier is used.

Metrics. For a benchmark with P problems, Avg@8 is the mean correctness over all $8P$ responses and Pass@8 is the fraction of problems with at least one correct response. Aggregate scores are unweighted macro-averages over the nine benchmarks, so large benchmarks such as MATH-500 and OlympiadBench do not dominate the average.

Generation diagnostics. Response length is measured in tokens of the Qwen3-1.7B-Base tokenizer and averaged over all rollouts. The no-answer rate is the fraction of rollouts without a parseable boxed answer, which in practice coincides with generations that hit the decoding cap before terminating. Distinct-4 is the number of unique 4-grams divided by the total number of 4-grams in a response, averaged over rollouts; low Distinct-4 combined with long outputs and a high no-answer rate indicates repetitive degeneration rather than longer reasoning.

E.5 Checkpoint Selection and Reporting

All trained models—TIDE, every baseline, and every ablation and hyperparameter variant—are evaluated at the final checkpoint of the one-epoch run; no checkpoint is selected using benchmark scores. Hyperparameter studies vary a single quantity at a time around the default configuration ($K = 16$, $\rho^- = \rho^+ = 0.2$, $\lambda = 1.0$), which was fixed before the benchmark evaluation and not tuned on test sets. All numbers within a teacher–student pair are produced by the same decoding and grading pipeline, so they are directly comparable across methods and ablations.

F CASE STUDIES

We qualitatively compare students trained with standard OPD and with TIDE under the strong-mismatch pair (Qwen3-8B → Qwen3-1.7B-Base), decoding both with the shared evaluation protocol (Appendix E.4). To avoid cherry-picking, candidate problems are selected by a fixed rule rather than manual inspection: the OPD student answers none of its 8 rollouts correctly while the TIDE student answers all 8 correctly. The four cases below are chosen from this pool to cover the spectrum of degeneration patterns that dominate the OPD student's failures, ordered from the most fluent to the most severe: (1) a fluent sentence-level loop, (2) a self-dialogue loop that imitates the prompt format, (3) a periodic junk-token loop, and (4) a multilingual token soup collapsing into a single-token loop. Excerpts are verbatim; elisions and repetition counts are marked in brackets. Reported statistics (correct answers out of 8, Distinct-4 of the excerpted rollout) are computed from the raw evaluation outputs.

F.1 CASE 1: FLUENT SENTENCE-LEVEL LOOP

Problem — MATH-500 · Answer: 0

If the domain of the function $\log x^2$ is $x < a$ or $x > b$, for some a and b, find $a + b$.

OPD Student — 0/8 correct · Distinct-4 = 0.009

For the function $\log x^2$, the domain is where the argument of the logarithm is positive. So, $x^2 > 0$, which is true for all $x \neq 0$. Now, we are given that the domain is $x < a$ or $x > b$ for some a and b. This means that the function $\log x^2$ is undefined for x in the interval $[a, b]$. So, the domain of the function is the union of the intervals $(-\infty, a)$ and (b, ∞).
Now, we are given that the domain is $x < a$ or $x > b$. This means that the function is defined for all x except for x in the interval $[a, b]$. So, the domain of the function is the union of the intervals $(-\infty, a)$ and (b, ∞).
[... the last paragraph repeats 399 times, until the 31,744-token decoding cap; no answer is produced ...]

TIDE Student — 8/8 correct

The domain of $\log x^2$ is all x such that $x \neq 0$, i.e. $x < 0$ or $x > 0$, which can be written as $(-\infty, 0) \cup (0, \infty)$. The endpoints are $a = 0$ and $b = 0$. Thus, the sum of a and b is $0 + 0 = 0$. Therefore, the final answer is $\boxed{0}$.

This is the most insidious pattern: the repeated content is fluent, locally coherent mathematical prose. Conditioned on a prefix that already contains several copies of the paragraph, continuing the loop is the high-probability continuation for the student *and* for the teacher, so token-level supervision along the loop appears well matched almost everywhere—the student-induced matched-region failure analyzed in the main text.

F.2 CASE 2: SELF-DIALOGUE LOOP IMITATING THE PROMPT FORMAT

Problem — Minerva · Answer: 2

Subproblem 0: For $\omega \geq 0$, find A such that $A\cos(\omega t)$ is a solution of $\ddot{x} + 4x = \cos(\omega t)$. *(Worked solution given in the prompt.)*
Subproblem 1: For what value of ω does resonance occur?

OPD Student 0/8 correct · Distinct-4 = 0.088

[... answers Subproblem 1, then continues inventing new subproblems ...]
Final answer: The condition for resonance is $\omega = \boxed{2}$.
The final answer is $\boxed{2}$. I hope it is correct.
Subproblem 6: What is the value of ω at which resonance occurs in the differential equation $\ddot{x} + 4x = \cos(\omega t)$? Please reason step by step, and put your final answer within \boxed{}.
[... the model invents 149 "subproblems" and answers each of them, repeating "I hope it is correct." 149 times, until the decoding cap; the final boxed expression is not the requested answer ...]

TIDE Student 8/8 correct

The natural frequency ω_n of the system can be found from the homogeneous equation $\ddot{x} + 4x = 0$, whose general solution is $x(t) = A\cos(2t) + B\sin(2t)$. So the natural frequency is $\omega_n = 2$. Resonance occurs when the frequency of the external force matches the natural frequency. Thus, the value of ω at which resonance occurs is $\boxed{2}$.

The failure here is structural rather than lexical: the OPD student *does* derive the correct answer $\omega = 2$ early in the rollout, but instead of terminating it imitates the few-shot format of the prompt and generates an endless sequence of self-posed subproblems. Because the imitated format is itself natural text, the teacher assigns high probability to each continuation, and the rollout is graded as incorrect because the final boxed expression drifts away from the requested answer.

F.3 Case 3: Periodic Junk-Token Loop from the First Token

Problem AIME 2024 · Answer: 25

There exist real numbers x and y, both greater than 1, such that $\log_x(y^x) = \log_y(x^{4y}) = 10$. Find xy.

OPD Student 0/8 correct · Distinct-4 = 0.009

```
ANGUAGE 2023-03-15T10:30:00Z
ENGINEER 2023-03-15T10:30:00Z
SECOND ENGINEER 2023-03-15T10:30:00Z
ENGINEER 2023-03-15T10:30:00Z
ENGINEER 2023-03-15T10:30:00Z
ENGINEER 2023-03-15T10:30:00Z
[... the same line repeats 1,322 times in total,
     until the 31,744-token decoding cap ...]
```

TIDE Student 8/8 correct

To solve the problem, we are given that there exist real numbers x and y, both greater than 1, such that $\log_x(y^x) = \log_y(x^{4y}) = 10$. We are asked to find the value of xy.
Step 1: Use the change of base formula. Recall that $\log_a b = \log_c b / \log_c a$. The first equation $\log_x(y^x) = 10$ can be rewritten as $\frac{\log y}{\log x} \cdot x = 10$, and the second equation $\log_y(x^{4y}) = 10$ as $\frac{\log x}{\log y} \cdot 4y = 10$.
Step 2: Define variables for the equations. Let $a = \frac{\log y}{\log x}$. Then the first equation becomes $ax = 10$ (1). Since $\frac{\log x}{\log y} = \frac{1}{a}$, the second equation becomes $\frac{1}{a} \cdot 4y = 10$, so $4y = 10a$ and $y = \frac{10a}{4} = \frac{5a}{2}$ (2).

Step 3: Substitute y in terms of x. From (1), $a = \frac{10}{x}$ (3). Substituting into (2), $y = \frac{5a}{2} = \frac{5 \cdot \frac{10}{x}}{2} = \frac{25}{x}$ (4).
Step 4: Find xy. Since $y = \frac{25}{x}$, we get $xy = x \cdot \frac{25}{x} = 25$.
Conclusion. The value of xy is $\boxed{25}$.

Unlike Cases 1 and 2, the student never engages with the problem: every rollout locks into a short periodic loop of problem-irrelevant tokens within the first few tokens and terminates only at the decoding cap.

F.4 Case 4: Token Soup Collapsing into a Single-Token Loop

Problem AMC 2023 · Answer: 36

Positive real numbers x and y satisfy $y^3 = x^2$ and $(y - x)^2 = 4y^2$. What is $x + y$?

OPD Student 0/8 correct · Distinct-4 = 0.031

嚯 왠 ㄲㄲㄲㄲ 슥 锾 洑 ㅇ ㅊ ㅉ ㅊ ㅇ ㅋ ㅇ .visitInsn 밂
絞 欹 ㅁ 잔 綦 횰 ㅌ 猍 氖軡 沺 ㅎ ᵻ ᵀ ¥
.getSystemService 格將 鸅 禈 鰺 凌 栓 看 ㅍ
/settingsdialog 食 緇 替 萸 复 咆 蔻 唬 潢 锔 驸 擘 匾 夕
闾 魨 桦 旭 换 遂 碑 鳜 媛 骏 锅 冢 蚌 臟 ㅍ 撞 猬 孓 图 发
[... incoherent multilingual / code tokens continue ...]
错 错 错 错 错 错 错 错 错 错 错 错 错 错 错 错
[... the token “错” (“wrong”) repeats 15,182 times,
until the 31,744-token decoding cap ...]

TIDE Student 8/8 correct

From the first equation, $x = y^{3/2}$. Substitute into the second equation: $(y - y^{3/2})^2 = 4y^2$. Let $t = y^{1/2}$, then $y = t^2$, so $(t^2 - t^3)^2 = 4t^4$, i.e. $t^4(t^2 - 2t - 3) = 0$. Since $t > 0$, solving $t^2 - 2t - 3 = 0$ gives $t = 3$, so $y = 9$ and $x = 9^{3/2} = 27$. Thus $x + y = 27 + 9 = 36$. Therefore, the answer is: $\boxed{36}$.

The most severe pattern: an incoherent stream of unrelated tokens spanning many scripts and domains (CJK characters, Hangul syllables, and code identifiers such as `.visitInsn`), which then collapses into a loop repeating a single character for tens of thousands of tokens.

F.5 Discussion

Three properties of these failures are worth noting. First, degeneration is not a tail event within an otherwise correct solution: in Cases 3 and 4 the OPD student derails within the first few tokens and never returns, which is why its accuracy collapses to the untrained-student level while its average generation length explodes. Second, the spectrum matters: the repeated content ranges from fully fluent prose (Case 1) and correct mathematics wrapped in an imitated dialogue format (Case 2) down to single-token loops (Case 4), so degeneration cannot be detected—let alone penalized—by fluency or perplexity alone. Third, in every case the repetition is self-consistent: once a repetitive prefix is established, continuing the loop is the most likely continuation for the student *and* for the teacher conditioned on the same prefix, so token-level supervision along such rollouts looks well matched almost everywhere. This is precisely the student-induced matched-region failure analyzed in the main text. TIDE attacks the mechanism directly: matched tokens receive zero weight, the excess gate suppresses the student's overconfident continuations of the loop, and the deficit branch reinstates the teacher-preferred exit tokens that the student's sampling never reaches.

Table 8: Hyperparameters used in all experiments. Method-specific parameters apply only to the corresponding method; all remaining settings are shared.

Hyperparameter	Value
Data and batching	
Training dataset	DAPO-Math-17K
Data shuffling	False
Prompt batch size	64
Rollouts per prompt	4
Policy mini-batch size	64
Micro-batch size per GPU	1
Training epochs	1
Maximum prompt length	1,024 tokens
Maximum response length	7,168 tokens
Maximum training context	8,192 tokens
Rollout generation	
Inference engine	vLLM
Student sampling temperature	1.0
Teacher scoring temperature	1.0
Repetition penalty	1.0
Optimization	
Optimizer	AdamW
Learning rate	1×10^{-6} (constant)
Warmup	None
Adam coefficients	$(0.9, 0.999)$
Weight decay	0.01
Gradient clipping	1.0
Policy updates per rollout batch	1
Loss aggregation	Token mean
Auxiliary reference-policy KL	None
Format reward	None
Numerical precision	FP32 student; FP32/BF16 teacher
Hardware	8 GPUs, FSDP
TIDE defaults	
Teacher support size K	16
Excess keep-rate ρ^-	0.2
Deficit keep-rate ρ^+	0.2
Deficit weight λ	1.0
Excess reward	Hellinger ratio reward
Deficit statistic	Forward KL on teacher top-K
Teacher top-K normalization	Renormalized within top-K
Matched-token weight	0
Baseline-specific settings	
OPD	Sampled-token log-ratio
GRPO	Exact-match reward; group size 4
FiRe-OPD	Bottom 20% filtered; $\alpha = \beta = 1.0$
PowerOPD	Box–Cox exponent $\alpha = 5$
AOPD	$K = 16,\ \tau = 0$; normalized top-K; no $1/K$ scaling
Final evaluation	
Samples per problem	8
Maximum response length	31,744 tokens
Sampling temperature	0.7
Top-p	0.95
Grader	Rule-based mathematical verifier