O^{2}PD: OVERLAP-GATED ON-POLICY DISTILLATION ON RELIABLE PREFIX

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) supplies dense token-level supervision by aligning a student to the teacher's next-token distribution on its own generated rollouts. Existing works mostly implement top-k OPD, applying reverse Kullback-Leibler (RKL) divergence between the teacher's and the student's renormalized distributions over teacher top-k support. However, the teacher scores student-generated rollouts that are inherently off-policy for itself, leaving the reliability of its next-token guidance unclear. Restricting the divergence to the teacher top-k support also discards the remaining mass outside that support. These gaps leave two questions implicit: 1) which tokens to align, and 2) how to align those tokens. To answer these questions, we propose Overlap-gated On-Policy Distillation (O^{2}PD). As for 1), we quantitatively analyze the reliability of teacher supervision and find that alignment on the student's trajectory prefix is more reliable than alignment on the full context. Therefore, our O^{2}PD is applied only to the 20% prefix tokens of each student rollout. As for 2), we theoretically characterize two problems of top-k OPD: the *Teacher-support Uncovering* and the *Dominant-Mode Mismatch*. Based on these analyses, we propose two support strategies: the *Reset Support* strategy attaches a reset atom to every token, pooling mass outside the teacher top-k, and the *Support Convergence* strategy converges the support to the already aligned intersection. Our O^{2}PD employs the designed overlap-gated α-divergence to interpolate between mode-covering and mode-seeking for per-token alignment, where the overlap is used as a gate to switch the support and adaptively modulate α. Extensive experiments on seven common benchmarks demonstrate the superiority of O^{2}PD, improving average performance by up to 5.29 points compared to OPD. `https://anonymous.4open.science/r/Code-59A9`

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) improves reasoning by assigning advantages from outcome-level correctness Guo et al. (2025); Shao et al. (2024); Hu et al. (2025); Zeng et al. (2025). That signal is sparse along long trajectories: many tokens share one reward, and groups of equally correct or incorrect responses can yield vanishing advantages Yu et al. (2025); Liu et al. (2025); Cui et al. (2025). Knowledge distillation supplies denser targets by matching a student to a teacher's per-token distribution Hinton et al. (2015), but the usual recipe trains on teacher-generated sequences Kim & Rush (2016). The student is then updated off-policy relative to its own sampler Gu et al. (2024); Agarwal et al. (2024). This introduces a distribution mismatch between training and those generated by the student itself. On-policy distillation (OPD) reverses that sampling: the student generates the rollout trajectory, and the teacher scores the same rollout, so each token receives a dense, on-policy learning signal Agarwal et al. (2024); Gu et al. (2024); Lu & Thinking Machines Lab (2025); Yang et al. (2026). The alignment process is conducted by minimizing the divergence between the teacher's and student's per-token distributions. For example, reverse Kullback-Leibler (RKL) is a mode-seeking objective that can effectively capture the teacher's dominant modes. Existing OPD methods mainly adopt top-k OPD, minimizing the RKL on the teacher's top-k support atoms for tractable cost.

However, the student-generated rollouts are inherently off-policy for the teacher, so the reliability of teacher supervision remains unclear Fu et al. (2026); Zhou et al. (2026). Besides, the teacher's top-k

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

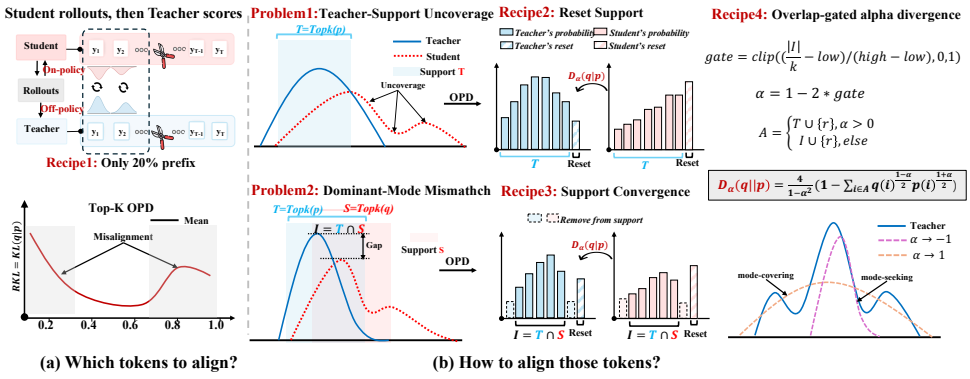

Figure 1: Framework of our O^{2}PD: (a) Restricting alignment to the prefix, where teacher supervision is reliable. (b) Overlap-gated α-divergence with Reset Support and Support Convergence strategies.

support atoms are just a subset of the full vocabulary Fu et al. (2026); Anshumann et al. (2025). Restricting the alignment to this subset discards the remaining mass outside that support. Empirically, as the teacher-student overlap decreases, the student more frequently samples tokens outside the teacher's top-k support, where the teacher's supervision is unreachable. These gaps leave two questions implicit: 1) *which tokens to align?*, and 2) *how to align those tokens?* Along these lines, we first identify a *Cross-prefix Asymmetry phenomenon*: as student prefixes accumulate, the teacher's accuracy falls, while as teacher prefixes accumulate, the student's accuracy rises (§3). Furthermore, we theoretically characterize two problems of top-k OPD (§4): 1) *Teacher-support Uncovering* happens because RKL on the teacher's top-k support fails to pull the student onto teacher-only tokens; thus high-confidence teacher modes stay uncovered when overlap is low; 2) *Dominant-Mode Mismatch* is that even the two top-k supports already overlap well, the same objective only matches the renormalized distribution and fails to pin the teacher's mass on the shared atoms.

To address these issues, we propose an overlap-gated α-divergence on refined support and reliable prefix for OPD, which we term O^{2}PD (shown in Figure 1). For *which*, inspired by the Cross-prefix Asymmetry phenomenon, we only optimize the 20% prefix tokens of student-generated rollouts to avoid the off-policy accumulation for the teacher. For *how*, we introduce two support strategies: the *Reset Support* strategy attaches a reset atom to every token, pooling mass outside the teacher top-k, and the *Support Convergence* strategy converges the support to the already aligned intersection. Theoretical analysis shows that Forward KL (FKL) with Reset Support supplies a first-order covering force on teacher-only atoms, and RKL with Support Convergence pins the teacher's mass on the shared atoms rather than leftover-renormalizing them. This motivates our design of overlap-gated α-divergence, where overlap is used as a gate to switch the support strategies and adaptively modulate α. For each token, α is mapped from the teacher-student overlap to control the interpolation between the mode-covering and the mode-seeking. Our contributions are:

- We identify a *Cross-prefix Asymmetry*: teacher supervision is more reliable on the prefix than on the suffix. We theoretically characterize two shortcomings of top-k OPD, *Teacher-support Uncovering* and *Dominant-Mode Mismatch* (§3, §4).
- We propose O^{2}PD, which applies overlap-gated α-divergence on the reliable prefix tokens. The per-token overlap between teacher and student adaptively modulates α and switches *Reset Support* and *Support Convergence* strategies (§3, §4).
- Extensive experiments on seven benchmarks show that O^{2}PD achieves state-of-the-art average performance among compared OPD methods, improving over OPD by up to $+5.29$ points across three teacher–student pairs.

2 Related Work

GKD, MiniLLM, and later OPD recipes minimize a token-level divergence loss on student-generated trajectories Gu et al. (2024); Yang et al. (2026); Ko et al. (2025). The standard implementation is

top-k OPD (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025; Anshumann et al., 2025), which minimizes RKL over the teacher's top-k atoms at every position. At each token, the teacher and student induce next-token distributions p and q on vocabulary V. Write $T = \mathrm{TopK}(p)$, $S = \mathrm{TopK}(q)$, $I = T \cap S$, and overlap $\rho = |I|/k$ (illustrated in Table 1). Top-k OPD renormalizes both distributions on T and minimizes

$$L_T = \mathrm{KL}(q^T \| p^T) = \sum_{a \in T} q_a^T \log \frac{q_a^T}{p_a^T}, \tag{1}$$

where $q_a^T = q(a)/q(T)$ and $p_a^T = p(a)/p(T)$. Mass outside T does not enter the loss. A concurrent line of work studies when this recipe succeeds or fails. Because the teacher scores student-sampled rollouts, its next-token guidance can be noisy, unteachable, or mismatched with the student Fu et al. (2026); Zhou et al. (2026); Wang et al. (2026b;a); Yu et al. (2026); Shao et al. (2026). Li et al. show that teacher–student top-k overlap tracks alignment-level and thinking-pattern consistency, and that high-overlap successful runs can be explained largely by overlap increasing Li et al. (2026). Other methods intervene on which tokens to optimize and how to lessen the overlap mismatch. For example, TIP selects tokens with high student entropy, and tokens with low student entropy plus high teacher–student divergence, where the student is overconfident and wrong Xu et al. (2026). ESR argues that early tokens do problem framing and strategic planning, and truncates the rollout to a fixed first N tokens Zhou et al. (2026). EOPD keeps the on-policy trajectory and mixes FKL into RKL on high teacher-entropy tokens to avoid diversity collapse Jin et al. (2026). However, these token filters are heuristic, ESR's cutoff is a fixed short horizon that does not scale with response length, and EOPD changes the alignment objective without using overlap to decide covering versus pinning. O^{2}PD takes both decisions together: a length-scaled prefix, and overlap-gated α-divergence with Reset Support and Support Convergence for better alignment.

Table 1: Supports at each token.

Set	Role
$T = \mathrm{TopK}(p)$	teacher support
$S = \mathrm{TopK}(q)$	student support
$I = T \cap S$	already aligned atoms
$T \setminus S$	teacher modes still to cover
$S \setminus T$	student-only modes to suppress

3 Where Should Distillation Concentrate?

3.1 Cross-prefix Asymmetry phenomenon

Student prefixes are off-policy for the teacher. To test whether teacher supervision remains reliable on student-generated prefixes, we run a cross-prefix completion protocol on MATH500 with a Qwen3-1.7B student and a Qwen3-4B teacher. When each model generates the entire response, the student and teacher obtain 73.8% and 85.8% pass@1. We then vary a prefix ratio ℓ: the student continues from a teacher-generated prefix of relative length ℓ, and the teacher continues from a student-generated prefix of the same length. Fig. 2 (a) shows an opposite dependence on ℓ. Student pass@1 increases as the teacher prefix lengthens; teacher pass@1 decreases as the student prefix lengthens, and the decline grows with ℓ.

Aligning prefix raises suffix overlap. To test whether aligning only those early tokens can raise teacher–student overlap on the rest of the trajectory, we use the same Qwen3-1.7B student and Qwen3-4B teacher and train top-k OPD on DAPO-MATH-17K with the loss applied only to the first 20% of each rollout. As shown in Fig. 2 (b), we compute mean top-k overlap between the student and the teacher, grouped by relative-position bin. Overlap on the untrained suffix bins rises together with the trained prefix, so the prefix-only objective improves alignment for suffix tokens. The 0–20% bin remains the lowest throughout, so the prefix is still the hardest region to match. §5.3 further verifies that aligning on a prefix fraction exceeds full-sequence alignment, and concentrating on the prefix outperforms aligning the suffix.

Misalignment concentrates on the prefix. As shown in Fig. 2 (c), we plot mean L_T against relative position at initialization and after full-trajectory top-k OPD. Mean RKL is highest near the start of the response and declines with relative position at both checkpoints. Training lowers the curve along the trajectory, so teacher-student mismatch is alleviated, yet the prefix remains the region of largest mismatch.

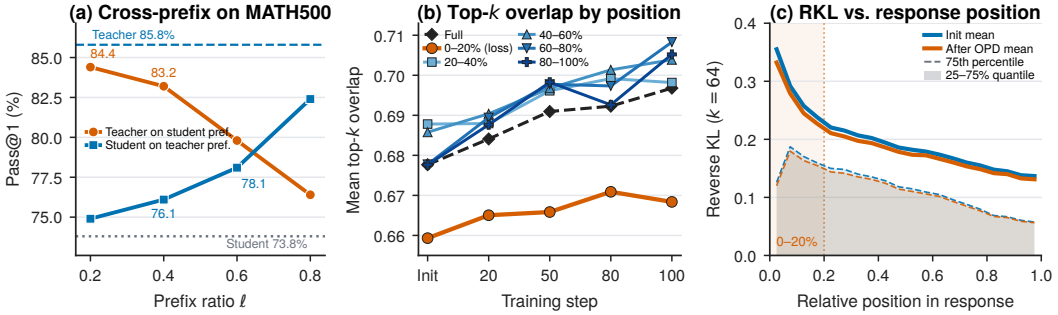

Figure 2: (a) Cross-prefix pass@1 versus prefix ratio ℓ on MATH500. (b) Mean top-k overlap by relative-position bin when L_T is applied only to the first 20% of each student trajectory. (c) Top-k OPD loss L_T versus relative position at initialization and after full-trajectory training.

3.2 RECIPE1: PREFIX-ONLY ALIGNMENT

Based on this analysis, the prefix is both the most reliable source of teacher supervision and the region of largest remaining mismatch, so these are the tokens most worth aligning. Prefix-only training still raises overlap on the untrained suffix, so withholding suffix tokens from the loss does not leave the suffix unaligned. We therefore replace full-trajectory L_T by the Where objective $\mathcal{L}^{(\ell)}_{\text{where}}$:

$$\mathcal{L}^{(\ell)}_{\text{where}} = \mathbb{E}_{y\sim\pi_s(\cdot|x)} \sum_{i=1}^{|y|} \mathbf{1}\left[i \le \ell\,|y|\right] L_T^{(i)}. \tag{2}$$

Suffix tokens are sampled with the student but do not enter the optimization. Because the window is a fraction of $|y|$, the number of aligned tokens scales with response length.

4 HOW TO ALIGN PREFIX TOKENS

4.1 TEACHER-SUPPORT UNCOVERING AND DOMINANT-MODE MISMATCH

We theoretically characterize two shortcomings of top-k OPD. The first failure is Teacher-support Uncovering: minimizing RKL on T involves gradient coordinates only on T, so it therefore assigns no gradient to student-only atoms $S \setminus T$, and its first-order force on teacher-only atoms in $T \setminus S$ is multiplied by student mass.

Lemma 1 (Support restriction). *$L_T = \mathrm{KL}(q^T \| p^T)$ depends on the student logits only through $\{z_a : a \in T\}$. Consequently $\partial L_T/\partial z_a = 0$ for every $a \notin T$. In particular, RKL on T has no coordinate on $S \setminus T$.*

On the simplex over T, the coordinate gradient of L_T is proportional to student mass q^T_a, so the first-order force on $T \setminus S$ is weak whenever that mass is small (Appendix B.1).

Proposition 1 (Covering channel). *Assume $I \neq \emptyset$, $q(I) > 0$, and $q(T \setminus S) = 0$. Then $q(T) = q(I)$, q^T vanishes on $T \setminus S$, and*

$$\frac{\partial L_T}{\partial \zeta_a} = 0 \qquad \text{for all } a \in T \setminus S.$$

In this case L_T is determined by q^I only, so the first-order update lives only on already aligned atoms. The covering channel on $T \setminus S$ is off, and Lemma 1 additionally gives $\partial L_T/\partial z_a = 0$ for every $a \in S \setminus T$. When $q(T \setminus S)$ is small but nonzero, the covering gradient remains proportional to q^T_a, so the first-order force on $T \setminus S$ is weak rather than identically zero.

Proposition 1 is why L_T does not pull $T \setminus S$ into S. Fig. 3 checks this on prefix tokens with initial overlap $\rho < 0.3$: mean overlap moves only from 0.265 to 0.282 after top-k OPD training, 87.3% of the cohort remain below 0.3, and none reach ≥ 0.6.

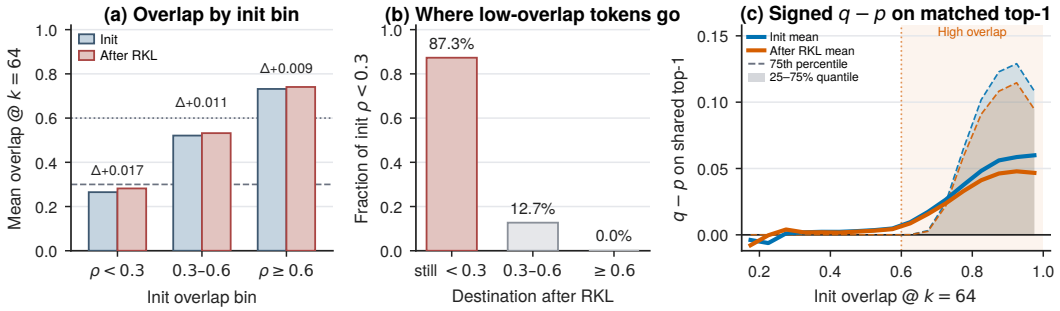

Figure 3: (a) Mean overlap by initial-overlap bin before and after training. (b) Destination of tokens with initial $\rho < 0.3$. (c) Signed gap $q - p$ on the shared top-1.

The second failure is Dominant-Mode Mismatch: minimizing RKL on the full-vocabulary is associated with $q = p$ as a mode-seeking target. Top-k OPD is not full-vocabulary OPD: its loss L_T lives on the renormalized simplex of T.

Proposition 2 (Renormalized match on T)**.** *On the relative interior $\{q^T : q(T) > 0\}$, every minimizer of L_T satisfies $q^T = p^T$. This matches the teacher only as a distribution on T. It does not pin $q(T) = p(T)$ and does not constrain $V \setminus T$, so it is not $q = p$ on V.*

Remaining mode-mass gaps are therefore a shape mismatch on the support T, not evidence that the student is insufficiently mode-seeking. Fig. 3 (c) checks tokens that already share the teacher's top-1: the signed gap $q - p$ on that mode stays positive after top-k OPD and grows with initial overlap. High overlap does not mean the student has matched the teacher's dominant-mode mass. Proofs of Lemma 1 and Propositions 1–2 are in Appendix B.

4.2 Recipe2: A reset atom on every token (Reset Support)

Teacher-support Uncovering can be handled by our Reset Support strategy, which attaches a reset atom to every token. The reset atom pools the mass outside T: teacher reset mass is $1 - p(T)$ and student reset mass is $1 - q(T)$. On the simplex over $T \cup \{r\}$,

$$\pi_p(a) = p(a)\ (a \in T), \pi_p(r) = 1 - p(T), \pi_q(a) = q(a)\ (a \in T), \pi_q(r) = 1 - q(T). \tag{3}$$

FKL on this simplex is $L_F = \mathrm{KL}(\pi_p \| \pi_q)$. Without a reset atom, FKL opens only a covering channel on $T \setminus S$. The reset coordinate additionally supplies a pooled suppression channel for $S \setminus T \subset V \setminus T$.

Proposition 3 (FKL covering)**.** *On simplex coordinates $\pi_q = \mathrm{softmax}(\zeta)$,*

$$\frac{\partial L_F}{\partial \zeta_a} = \pi_q(a) - \pi_p(a).$$

If $a \in T \setminus S$, $\pi_p(a) > 0$ and $\pi_q(a) = 0$, then $\partial L_F / \partial \zeta_a = -\pi_p(a) < 0$, so a descent step raises ζ_a. Teacher mass on $T \setminus S$ supplies covering force without requiring the student to already place mass there. The reset coordinate yields $\partial L_F / \partial \zeta_r = q_r - p_r$ with $q_r = 1 - q(T)$ and $p_r = 1 - p(T)$. When $T \setminus S$ is large the student typically places substantial mass outside T, so $q_r > p_r$ and a descent step lowers the reset coordinate, a pooled suppression of $S \setminus T$. If $p(T \setminus S) > 0$, then $L_F \to +\infty$ as $q^T|_{T \setminus S} \to 0$, whereas L_T remains finite.

The reason to use FKL when $T \setminus S$ is large: the covering channel stays open when student mass on $T \setminus S$ is near zero. RKL on the same support still multiplies those coordinates by $q(a)$.

4.3 Recipe3: shrink the support, then RKL (Support Convergence)

Dominant-Mode Mismatch means that minimizing RKL on T matches only $q^T = p^T$. Attaching a reset atom to T makes the unconstrained RKL minimizer $\pi_q = \pi_p$ on $T \cup \{r\}$. On $T \cup \{r\}$ this is $q|_T = p|_T$, which pins $c = q(T)/p(T) = 1$ and closes the scale gap in Proposition 2. Reset Support

is therefore already enough for the unconstrained problem. As overlap rises, $T \setminus S$ shrinks and the job becomes mode-seeking on the already shared atoms I. The interiors of $T \cup \{r\}$ and $I \cup \{r\}$ coincide only when $T = I$. Leftover atoms $T \setminus I$ still carry teacher mass but little student mass. Proposition 1 keeps the RKL factor on those atoms proportional to $q(a)$, so leftover student mass $\varepsilon = q(T \setminus I)$ typically remains below $p(T \setminus I)$ rather than rising to the unconstrained leftover target. On this slice RKL on $T \cup \{r\}$ still renormalizes I by leftover teacher mass. Support Convergence strategy is therefore proposed to restrict the support to I. Leftover teacher mass is folded into the reset atom, $\pi_q(a) = q(a)$ $(a \in I), \pi_q(r) = 1 - q(I), \pi_p(a) = p(a)$ $(a \in I), \pi_p(r) = 1 - p(I)$, and $L_{I,r} = \mathrm{KL}(\pi_q \| \pi_p)$.

Proposition 4 (Leftover renormalization)**.** *Assume $I \neq \emptyset$, $T \setminus I \neq \emptyset$, $q(I) > 0$, and $p(T) < 1$. Write $\varepsilon = q(T \setminus I) \in [0, 1)$ and freeze the leftover coordinates. Then* RKL *on $T \cup \{r\}$, as a function of the remaining mass $1 - \varepsilon$ on $I \cup \{r\}$, is uniquely minimized at*

$$q(a) = \frac{(1-\varepsilon)\, p(a)}{1 - p(T \setminus I)} \quad (a \in I), \qquad 1 - q(I) - \varepsilon = \frac{(1-\varepsilon)\,(1 - p(T))}{1 - p(T \setminus I)}.$$

Hence $q(a) > p(a)$ on I if and only if $\varepsilon < p(T \setminus I)$. The uncovered face $\varepsilon = 0$ is the special case $q(a) = p(a)/(1 - p(T \setminus I))$. Minimizing RKL *on $I \cup \{r\}$ does not include $T \setminus I$; its interior minimizer is $q|_I = p|_I$ and $1 - q(I) = 1 - p(I)$. Leftover teacher mass therefore renormalizes the aligned atoms under $T \cup \{r\}$ whenever student leftover mass is smaller than teacher leftover mass, and is folded into the reset atom under $I \cup \{r\}$.*

Proposition 4 is why mode-seeking RKL uses support $I \cup \{r\}$: when leftover mass remains below leftover teacher mass, $T \cup \{r\}$ still overshoots p on I, not only on the exact-zero face. Proofs of Propositions 3 and 4 are in Appendix B.

4.4 RECIPE4: OVERLAP-GATED α-DIVERGENCE

Low overlap needs the covering force of FKL on $T \cup \{r\}$ (Proposition 3); high overlap needs RKL on $I \cup \{r\}$, where $T \cup \{r\}$ would still leftover-renormalize I (Proposition 4). Overlap ρ selects which of those two jobs a prefix token presents. We therefore propose overlap-gated α-divergence: ρ interpolates those two regimes and switches the support strategies. Let π_q and π_p be the student and teacher distributions on the simplex over the current support A, either $T \cup \{r\}$ or $I \cup \{r\}$. For $\alpha \in (-1, 1)$, the α-divergence is

$$D_\alpha(\pi_q \| \pi_p) = \frac{4}{1 - \alpha^2}\left(1 - \sum_{i \in A} \pi_q(i)^{\frac{1-\alpha}{2}}\, \pi_p(i)^{\frac{1+\alpha}{2}}\right). \tag{4}$$

The endpoints recover the two KL directions used above: $D_\alpha \to \mathrm{KL}(\pi_p \| \pi_q)$ as $\alpha \to 1$, and $D_\alpha \to \mathrm{KL}(\pi_q \| \pi_p)$ as $\alpha \to -1$. Appendix C verifies the two limits for (4). Intermediate α interpolates those two vector fields rather than adding a second loss. On each update, let $\mathcal{P}$ be the prefix tokens that enter the loss. Set $\mathrm{low} = \min_{i \in \mathcal{P}} \rho_i, \mathrm{high} = \max_{i \in \mathcal{P}} \rho_i$, and let $\rho_{0.5}$ be the median percentile of $\{\rho_i : i \in \mathcal{P}\}$. Define

$$\mathrm{gate} = \mathrm{clip}\big((\rho - \mathrm{low})/(\mathrm{high} - \mathrm{low}), 0, 1\big), \qquad \alpha = 1 - 2 \cdot \mathrm{gate}. \tag{5}$$

The support switches at the median: $T \cup \{r\}$ when $\rho < \rho_{0.5}$, and $I \cup \{r\}$ when $\rho \geq \rho_{0.5}$. We therefore propose the overlap-gated α-divergence for adaptively aligning each token via ρ:

$$\mathcal{L}_{\mathrm{how}} = \mathbb{E}_{y \sim \pi_s(\cdot|x)} \sum_{i=1}^{|y|} D_{\alpha_i}\big(\pi_q^{(i)} \| \pi_p^{(i)}\big), \tag{6}$$

where ρ_i determines both α_i and the support of D_{α_i} as above. The full objective puts the Where prefix window on this How divergence. With prefix fraction $\ell = 0.2$ from §3,

$$\mathcal{L} = \mathbb{E}_{y \sim \pi_s(\cdot|x)} \sum_{i=1}^{|y|} \mathbf{1}\big[i \leq \ell\,|y|\big]\, D_{\alpha_i}\big(\pi_q^{(i)} \| \pi_p^{(i)}\big). \tag{7}$$

Suffix tokens are sampled with the student but do not enter the loss.

5 EXPERIMENTS

5.1 SETUP

Models and training dataset. We use three student–teacher pairs (Table 2). Qwen3-1.7B distilled from Qwen3-4B Yang et al. (2025) is a same-family distillation with a larger teacher. DSR1-1.5B distilled from JustRL-1.5B Guo et al. (2025); He et al. (2025) is a same-size distillation on long trajectories. Qwen3-1.7B distilled from Qwen3-4B-Non-Thinking-RL (NTRL) Yang et al. (2025) keeps the first student and changes only the teacher, isolating teacher strength. The training set is DAPO-MATH-17K with no additional filtering Yu et al. (2025). The compared baselines are: OPD, ESR Zhou et al. (2026), EOPD Jin et al. (2026), and TIP Xu et al. (2026). OPD is the default top-k recipe. ESR, EOPD, and TIP are trained with the configurations from their papers.

Evaluation. For fair comparison, we follow previous evaluation pipeline Fu et al. (2026); Lu & Thinking Machines Lab (2025) and evaluate on seven common benchmarks: AIME24, AIME25, AIME26, HMMT25, Minerva Lewkowycz et al. (2022), Olympiad He et al. (2024), and MATH500 Hendrycks et al. (2021); Lightman et al. (2024). AIME24/25/26 and HMMT25 have few problems and report Avg@8; Minerva, Olympiad, and MATH500 have many more and report pass@1 from a single sample. Reported scores use the best checkpoint of each method.

Implementation. All methods are trained in VERL Sheng et al. (2025). OPD and O^{2}PD use distillation top-k size $k = 16$. O^{2}PD applies the loss to a prefix fraction $\ell = 0.2$, attaches a reset atom, maps α from the prefix overlap min and max, and switches named support at the median of prefix overlap: $T \cup \{r\}$ below the median and $I \cup \{r\}$ at or above it. Further training hyperparameters are in Appendix E, and evaluation details in Appendix F.

Table 2: Performance (%) for three OPD student–teacher pairs. Len is the mean generated length in tokens. Bold marks the best distilled method in each column within a pair; underline marks the second best. Δ vs. OPD is O^{2}PD minus OPD.

Method	AIME24	AIME25	AIME26	HMMT25	Minerva	Olympiad	MATH500	Len	Avg
Qwen3-1.7B (student) ← Qwen3-4B (teacher)									
Qwen3-1.7B	8.33	12.08	5.83	6.25	20.59	36.74	73.8	3049	23.37
Qwen3-4B	23.33	31.67	21.67	15.00	31.99	45.48	85.8	4127	36.42
OPD	13.75	14.17	<u>7.08</u>	5.83	<u>23.53</u>	36.00	74.0	5033	24.91
ESR	13.75	14.17	<u>7.08</u>	**8.33**	21.32	38.37	74.2	4480	25.32
EOPD	<u>14.17</u>	12.50	**11.67**	5.00	**23.90**	38.37	74.6	5110	<u>25.74</u>
TIP	12.92	<u>14.58</u>	6.67	<u>6.67</u>	21.32	**39.70**	**76.8**	5377	25.52
O^{2}PD	**15.42**	**15.00**	**11.67**	<u>6.67</u>	**23.90**	<u>38.96</u>	<u>76.0</u>	4684	**26.80**
Δ vs. OPD	+1.67	+0.83	+4.59	+0.84	+0.37	+2.96	+2.0	−−	+1.89
DSR1-1.5B (student) ← JustRL-1.5B (teacher)									
DSR1-1.5B	28.75	23.75	21.25	12.08	18.38	28.00	84.4	15627	30.94
JustRL-1.5B	50.83	40.00	35.83	22.08	27.94	42.96	91.4	9308	44.43
OPD	42.92	32.08	31.67	16.67	16.91	18.67	83.8	12241	34.67
ESR	36.67	30.00	27.50	15.83	<u>20.22</u>	25.04	83.4	12376	34.09
EOPD	42.50	**34.17**	23.75	17.50	18.38	<u>29.48</u>	86.0	11677	35.97
TIP	<u>45.83</u>	31.67	**31.67**	<u>18.75</u>	17.65	25.63	<u>86.6</u>	11552	<u>36.83</u>
O^{2}PD	**46.67**	<u>33.33</u>	<u>30.42</u>	**20.00**	**23.16**	**33.78**	**89.0**	10875	**39.48**
Δ vs. OPD	+3.75	+1.25	−1.25	+3.33	+6.25	+15.11	+5.2	−−	+4.81
Qwen3-1.7B (student) ← Qwen3-4B-NTRL (teacher)									
Qwen3-1.7B	8.33	12.08	5.83	6.25	20.59	36.74	73.8	3049	23.37
Qwen3-4B-NTRL	70.00	63.33	65.00	38.75	39.34	64.44	94.8	11751	62.24
OPD	45.42	40.42	37.08	21.25	31.62	51.85	89.4	19129	45.29
ESR	46.67	37.92	42.50	22.92	33.09	<u>54.22</u>	90.2	13893	46.79
EOPD	<u>52.08</u>	39.17	40.83	<u>23.33</u>	<u>34.56</u>	53.19	<u>90.2</u>	17696	47.62
TIP	50.42	<u>44.58</u>	<u>43.33</u>	22.92	32.72	52.74	89.8	19036	<u>48.07</u>
O^{2}PD	**54.17**	**45.83**	**44.58**	**25.83**	**36.76**	**55.70**	**91.2**	19631	**50.58**
Δ vs. OPD	+8.75	+5.41	+7.50	+4.58	+5.14	+3.85	+1.8	−−	+5.29

5.2 Main distillation

Table 2 reports seven-suite Avg for the three student–teacher pairs. O^{2}PD attains 26.80 with the Qwen3-4B teacher, 39.48 with JustRL-1.5B, and 50.58 with NTRL. On every pair it exceeds ESR, EOPD, and TIP on Avg, and it improves over OPD by $+1.89$, $+4.81$, and $+5.29$. TIP and ESR raise Avg performance over OPD on all three teacher-student pairs, by up to $+3.63$ and $+2.36$. This reveals that optimizing a subset of tokens is more effective than optimizing all tokens. However, ESR's Avg gains remain small, and on some benchmarks ESR falls behind OPD, which demonstrates that dynamically adjusting the number of selected tokens may be necessary on harder problems. EOPD also raises Avg over OPD, by up to $+3.18$, by mixing FKL into RKL on high-entropy teacher tokens, showing that RKL alone is insufficient and benefits from an FKL covering channel. Instead of applying complex token-selective strategies, we optimize a fixed fraction of each trajectory's prefix, so the number of aligned tokens scales with the response length. We then use overlap-gated α-divergence to combine FKL covering with RKL mode-seeking and switch the support with overlap. This yields a further Avg gain over ESR, EOPD, and TIP on all three teacher-student pairs.

5.3 Discussion

Method components. We ablate five variants (Fig. 4). *w/o How* applies the prefix restriction to OPD. *w/o Where* applies overlap-gated α-divergence with Reset Support and Support Convergence to the full trajectory. FKL/RKL uses FKL on $T \cup \{r\}$ below the prefix overlap median and RKL on $I \cup \{r\}$ at or above it. *w/o I* keeps the prefix and runs overlap-gated α-divergence on $T \cup \{r\}$ only. *w/o r* keeps the prefix and runs overlap-gated α-divergence on the teacher top-k T, without the reset atom.

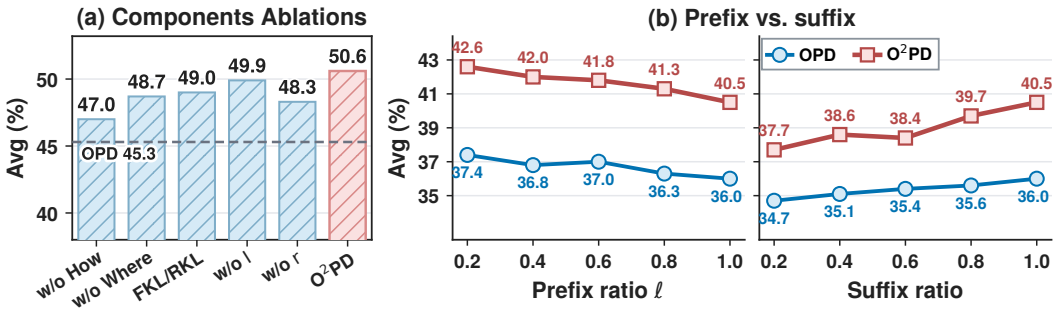

Figure 4: (a) Method-component ablation studies (%). Each bar removes one component of O^{2}PD. (b) Prefix versus suffix performance. Avg is the mean of Avg@8 on AIME24, AIME25, AIME26, and HMMT25.

On the Qwen3-1.7B student with the NTRL teacher, all five variants lie between OPD and full O^{2}PD on seven-suite Avg (Fig. 4 (a)). *w/o How* exceeds full-trajectory OPD, so the prefix restriction alone beats applying the loss to every token. *w/o Where* is higher, so How on the full trajectory contributes more than the prefix restriction alone. Dropping either block leaves a gap to full O^{2}PD. Within How, *w/o r* is the largest single-component drop; the reset atom carries the covering channel. With the support already switching with overlap, FKL/RKL stays below both *w/o I* and full O^{2}PD: continuous α interpolates the two vector fields rather than adding a second loss. Converging the support to $I \cup \{r\}$ as overlap rises is the smallest increment.

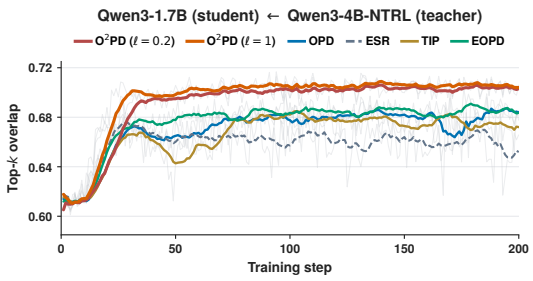

Figure 5: Mean top-k overlap trend.

Prefix versus suffix. To test whether the teacher provides effective supervision on prefix tokens relative to suffix tokens and to the full rollout, we vary where the loss sits on the full trajectory for both OPD and O^{2}PD. The prefix sweep applies the loss to the first fraction ℓ of each

Table 3: Support-switch percentile on the Qwen3-1.7B (student) ← Qwen3-4B-NTRL (teacher).

	AIME24		AIME25		HMMT25		
Percentile	Avg@8	pass@8	Avg@8	pass@8	Avg@8	pass@8	Avg
0.25	51.67	66.67	42.92	63.33	23.75	36.67	39.45
0.5	**54.17**	**73.33**	**45.83**	**70.00**	**25.83**	**43.33**	**41.94**
0.75	52.50	70.00	45.42	66.67	25.42	40.00	41.11
1.0	53.75	**73.33**	43.33	66.67	24.58	40.00	40.55

student response; the suffix sweep applies it to the last fraction. The two methods share the full-trajectory endpoint $\ell = 1$. Both methods are higher at $\ell = 0.2$ than at full-trajectory training (Fig. 4 (b)). Optimizing the suffix pulls down that distillation improvement. At a matched budget of 0.2, the prefix outperforms the suffix for both methods, and suffix-only training falls below the full trajectory. The teacher therefore provides effective supervision on prefix tokens. O^{2}PD remains above OPD at every measured endpoint, and the best point is prefix $\ell = 0.2$ with the How recipes.

Overlap dynamics. To test whether O^{2}PD raises teacher–student alignment level, measured by mean top-k overlap, and whether the prefix fraction $\ell = 0.2$ still attains the alignment of full-trajectory distillation, we track that overlap over training on the Qwen3-1.7B student with the NTRL teacher for O^{2}PD at $\ell = 0.2$ and $\ell = 1$, together with OPD, ESR, TIP, and EOPD (Fig. 5). All methods start near 0.61, rise in the first few steps, and then plateau through step 200. Distillation under O^{2}PD at $\ell = 1$ rises faster than at $\ell = 0.2$; the prefix run then catches up, and both finish near 0.70. That level is above OPD, EOPD, and TIP near 0.68, and above ESR. Prefix-only distillation therefore delays alignment and does not lower the level O^{2}PD attains.

5.4 Sensitivity to the support-switch percentile

In §4.4, the support switches when overlap ρ crosses a percentile of the prefix overlaps, by default the median $\rho_{0.5}$. This subsection measures how sensitive O^{2}PD is to that percentile. We keep $\ell = 0.2$, a reset atom, and overlap-gated α fixed, and replace $\rho_{0.5}$ by $\rho_{0.25}$, $\rho_{0.5}$, $\rho_{0.75}$, or $\rho_{1.0}$ (Tab. 3). Tokens below the chosen percentile stay on $T \cup \{r\}$; tokens at or above it use $I \cup \{r\}$. We report Avg@8 and pass@8 on AIME24, AIME25, and HMMT25. The median is the best of the four percentiles. Percentile 0.5 has the highest Avg@8 on every benchmark. Switching to $I \cup \{r\}$ at 0.25 hurts more than leaving tokens on $T \cup \{r\}$ through 0.75 or 1.0. Actor entropy declines over training at every percentile, and the decline is larger as the percentile approaches 1.0 (Fig. 6), where almost all prefix tokens remain on $T \cup \{r\}$. That trend is consistent with leftover renormalization on $T \cup \{r\}$ (Proposition 4). We therefore use percentile 0.5: it yields the best performance of the four percentiles, and its entropy is not the lowest, so the student can keep exploration.

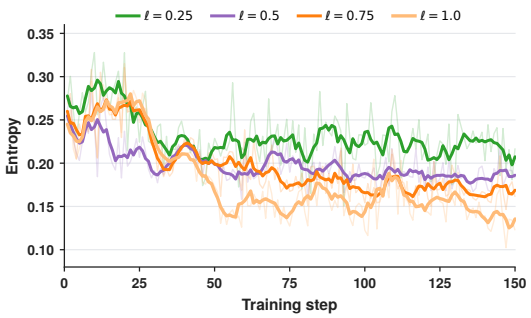

Figure 6: Actor entropy during training at the four support-switch percentiles on the Qwen3-1.7B student with the NTRL teacher.

6 Conclusion

In this paper, we study two questions in OPD: *which tokens to align*, and *how to align those tokens*. For *which*, we find that teacher supervision is more reliable on the student prefix than on the suffix. For *how*, we prove that minimizing RKL on the teacher top-k neither covers teacher-only atoms at low overlap nor pins teacher mass on shared atoms at high overlap. Therefore, we propose O^{2}PD, which concentrates on the alignment of the 20% prefix of each rollout via a designed overlap-gated α-divergence. The per-token overlap between teacher and student is used to adaptively modulate α and switch between Reset Support and Support Convergence. On seven benchmarks O^{2}PD attains the best average performance among compared methods, improving over OPD by $+1.89$, $+4.81$, and $+5.29$ on the three different student–teacher pairs.

AI USE STATEMENT

In this work we used generative AI tools (LLM writing and coding assistants) for the three uses reported on the ICLR 2027 submission form.

Writing. We used them to outline the paper, draft sections, edit for clarity, format LaTeX tables and captions, and suggest titles and keywords. The authors reviewed and revised all submitted text.

Retrieval and discovery. We used them to search and summarize retrieved papers. Every citation was checked against a public source; unverified references were discarded.

Research ideation and execution. We used them to refine the research questions and conceptual framework, to formulate and check mathematical claims and proof sketches. They were not used to generate synthetic data.

ETHICS STATEMENT

The study distills language models on public math benchmarks and does not involve human subjects. Dataset licenses and intended-use constraints will be listed here before submission.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Shun-ichi Amari. α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, 2009. doi: 10.1109/TIT.2009.2030485.

Anshumann, Mohd Abbas Zaidi, Akhil Kedia, Jinwoo Ahn, Taehwak Kwon, Kangwook Lee, Haejun Lee, and Joohyung Lee. Sparse logit sampling: Accelerating knowledge distillation in LLMs, 2025. URL https://arxiv.org/abs/2503.16870.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models, 2025. URL https://arxiv.org/abs/2505.22617.

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, 2026. URL https://arxiv.org/abs/2603.25562.

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. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Bingxiang He, Zekai Qu, Zeyuan Liu, Yinghao Chen, Yuxin Zuo, Cheng Qian, Kaiyan Zhang, Weize Chen, Chaojun Xiao, Ganqu Cui, Ning Ding, and Zhiyuan Liu. JustRL: Scaling a 1.5B LLM with a simple RL recipe, 2025. URL https://arxiv.org/abs/2512.16649.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 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, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

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. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2103.03874.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model, 2025. URL https://arxiv.org/abs/2503.24290.

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, 2026. URL https://arxiv.org/abs/2603.07079.

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, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A contrastive approach boosts the distillation of LLMs, 2025. URL https://arxiv.org/abs/2503.07067.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2206.14858.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe, 2026. URL https://arxiv.org/abs/2604.13016.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-zero-like training: A critical perspective, 2025. URL https://arxiv.org/abs/2503.20783.

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

Bing Shao, Jiazheng Zhang, Long Ma, Yujiong Shen, Senjie Jin, Xin Guo, Yuming Yang, Mingxu Chai, Zhiheng Xi, Boyang Liu, Junlin Shang, Tao Gui, Qi Zhang, and Xuanjing Huang. A token-level analysis of sampled-token reverse-KL on-policy distillation, 2026. URL https://arxiv.org/abs/2608.25643.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

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. doi: 10.1145/3689031.3696075.

Rui Wang, Hongru Wang, Yi Chen, Boyang Xue, Tianqing Fang, Wenhao Yu, and Kam-Fai Wong. Demystifying on-policy distillation: Roles, pathologies, and regulations, 2026a. URL https://arxiv.org/abs/2607.13399.

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, 2026b. URL https://arxiv.org/abs/2605.26844.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token importance in on-policy distillation, 2026. URL https://arxiv.org/abs/2604.14084.

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, 2025. URL https://arxiv.org/abs/2505.09388.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation, 2026. URL https://arxiv.org/abs/2602.12125.

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

Zichao Yu, Chengzhi Yu, Shengze Xu, Yujin Han, Bingqing Jiang, Xu Wang, and Difan Zou. Mismatch matters: On-policy distillation beyond token agreement, 2026. URL https://arxiv.org/abs/2608.09836.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. SimpleRL-Zoo: Investigating and taming zero reinforcement learning for open base models in the wild. In *Second Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2503.18892.

Ziheng Zhou, Jiaqi Li, Huacong Tang, Ying Nian Wu, and Demetri Terzopoulos. Less is more: Early stopping rollout for on-policy distillation, 2026. URL https://arxiv.org/abs/2605.27028.

CONTENTS

A O^{2}PD training step **14**

B Gradient proofs **14**

B.1 RKL simplex gradient 14

B.2 Support restriction 15

B.3 Covering channel 15

B.4 Renormalized match on T 15

B.5 FKL covering 16

B.6 Leftover renormalization 16

C Endpoints of α-divergence **17**

D The two failures after O^{2}PD **18**

E Training configuration **18**

F Evaluation **18**

A O^{2}PD TRAINING STEP

Algorithm 1 implements (7). Only prefix tokens enter the loss. Overlap ρ maps α from the prefix min and max, and switches Reset Support $T \cup \{r\}$ with Support Convergence $I \cup \{r\}$ at the median of prefix overlaps. Reported runs use $\ell = 0.2$ and $k = 16$ (§5.1).

Algorithm 1 O^{2}PD update on one student rollout.

Require: student π_s, teacher π_t, prompt x, prefix fraction ℓ, top-k size k
sample $y \sim \pi_s(\cdot \mid x)$
$\mathcal{P} \leftarrow \{i : 1 \le i \le \ell\,|y|\}$
for $i \in \mathcal{P}$ **do**
  $p^{(i)} \leftarrow \pi_t(\cdot \mid x, y_{<i})$, $q^{(i)} \leftarrow \pi_s(\cdot \mid x, y_{<i})$
  $T_i \leftarrow \mathrm{TopK}(p^{(i)})$, $S_i \leftarrow \mathrm{TopK}(q^{(i)})$, $I_i \leftarrow T_i \cap S_i$
  $\rho_i \leftarrow |I_i|/k$
end for
low $\leftarrow \min_{i\in\mathcal{P}} \rho_i$, high $\leftarrow \max_{i\in\mathcal{P}} \rho_i$
$\rho_{0.5} \leftarrow$ median of $\{\rho_i : i \in \mathcal{P}\}$
$\mathcal{L} \leftarrow 0$
for $i \in \mathcal{P}$ **do**
  gate $\leftarrow \mathrm{clip}\big((\rho_i - \text{low})/(\text{high} - \text{low}), 0, 1\big)$
  $\alpha_i \leftarrow 1 - 2 \cdot \text{gate}$
  if $\rho_i < \rho_{0.5}$ **then**
    $A \leftarrow T_i \cup \{r\}$ ▷ Reset Support
    $\pi_p(a) \leftarrow p^{(i)}(a)$ for $a \in T_i$, $\pi_p(r) \leftarrow 1 - p^{(i)}(T_i)$; likewise π_q from $q^{(i)}$
  else
    $A \leftarrow I_i \cup \{r\}$ ▷ Support Convergence
    $\pi_p(a) \leftarrow p^{(i)}(a)$ for $a \in I_i$, $\pi_p(r) \leftarrow 1 - p^{(i)}(I_i)$; likewise π_q from $q^{(i)}$
  end if
  $\mathcal{L} \leftarrow D_{\alpha_i}(\pi_q \| \pi_p)$
end for
update π_s on $\mathcal{L}$

B GRADIENT PROOFS

This section proves Lemma 1 and Propositions 1–4. Throughout, $q = \mathrm{softmax}(z)$ on V, p does not depend on z, $T = \mathrm{TopK}(p)$, $S = \mathrm{TopK}(q)$, and $I = T \cap S$. We always assume the support in a display is nonempty and that the corresponding student mass is positive, so renormalized restrictions are well-defined. No further assumption is placed on the shapes of p and q. The reset atom is introduced only in §B.5–§B.6; the failure statements in §B.2–§B.4 concern L_T without the reset atom.

B.1 RKL SIMPLEX GRADIENT

Let $\pi = \mathrm{softmax}(\zeta)$ on a finite simplex with strictly positive target π_p, and let $L = \mathrm{KL}(\pi\|\pi_p) = \sum_i \pi_i \log(\pi_i/\pi_{p,i})$. The softmax Jacobian is $\partial\pi_i/\partial\zeta_a = \pi_i(\delta_{ia} - \pi_a)$. Differentiating,

$$\begin{aligned}\frac{\partial L}{\partial \zeta_a} &= \sum_i \big(\log \tfrac{\pi_i}{\pi_{p,i}} + 1\big)\,\pi_i(\delta_{ia} - \pi_a) \\ &= \pi_a\big(\log \tfrac{\pi_a}{\pi_{p,a}} + 1\big) - \pi_a \sum_i \pi_i\big(\log \tfrac{\pi_i}{\pi_{p,i}} + 1\big) \\ &= \pi_a\big(\log \tfrac{\pi_a}{\pi_{p,a}} + 1\big) - \pi_a(L+1) = \pi_a\big(\log \tfrac{\pi_a}{\pi_{p,a}} - L\big). \end{aligned} \tag{8}$$

The convention $0 \log 0 = 0$ is used throughout. If $\pi_a \to 0$ while L remains bounded, then $\pi_a \log \pi_a \to 0$, hence $\partial L/\partial\zeta_a \to 0$. Applying (8) on the simplex of T with $\pi = q^T$ and $\pi_p = p^T$ yields

$$\frac{\partial L_T}{\partial \zeta_a} = q_a^T\Big(\log \frac{q_a^T}{p_a^T} - L_T\Big), \qquad a \in T. \tag{9}$$

The right-hand side vanishes whenever $q_a^T = 0$. Membership $a \in T \setminus S$ does not force $q(a) = 0$, but it does place a outside S, so student mass on those atoms is typically far below the mass on S. Softmax never attains exact zeros; $q(T \setminus S) = 0$ is a slice used in Proposition 1, not a training constraint.

B.2 Support restriction

Proof of Lemma 1. For $c \in T$,

$$q_c^T = \frac{q(c)}{q(T)} = \frac{e^{z_c}}{\sum_{b \in T} e^{z_b}}.$$

The partition function of q over V cancels, so q^T depends on z only through $\{z_b : b \in T\}$. The teacher restriction p^T does not depend on z. Therefore $L_T = \sum_{a \in T} q_a^T \log(q_a^T / p_a^T)$ is a function of those logits alone, which gives $\partial L_T / \partial z_a = 0$ for every $a \notin T$. Every $a \in S \setminus T$ and every $a \in V \setminus (T \cup S)$ falls in this set, so RKL on T has no coordinate there. □

B.3 Covering channel

Proof of Proposition 1. The assumption $q(T \setminus S) = 0$ with $q(I) > 0$ gives $q(T) = q(I)$ and $q_a^T = 0$ for every $a \in T \setminus S$. Substituting into (8) yields $\partial L_T / \partial \zeta_a = 0$ on $T \setminus S$. The remaining coordinates live on I, and

$$L_T = \sum_{a \in I} q_a^I \log \frac{q_a^I}{p(a)/p(T)},$$

which is RKL between q^I and the restriction of p^T to I. Lemma 1 gives $\partial L_T / \partial z_a = 0$ for $a \in S \setminus T$.

If instead $q(T \setminus S)$ is small but positive, (8) remains proportional to q_a^T on $T \setminus S$. Emptiness of $T \setminus S$ as a set is not what shuts the covering channel. For $a \in T \setminus S$,

$$\frac{\partial L_T}{\partial \zeta_a} = 0 \quad \Longleftrightarrow \quad q_a^T = 0 \quad \text{or} \quad \log \frac{q_a^T}{p_a^T} = L_T.$$

Thus $a \notin S$ does not imply a vanishing gradient. The first-order covering share

$$\rho(L_T) := \frac{\sum_{a \in T \setminus I} \left| \partial L_T / \partial \zeta_a \right|}{\sum_{a \in T} \left| \partial L_T / \partial \zeta_a \right|}$$

is zero whenever every $a \in T \setminus I$ satisfies the display above (and is undefined if the denominator is zero). □

B.4 Renormalized match on T

Proof of Proposition 2. $L_T = \mathrm{KL}(q^T \| p^T)$ is RKL on the finite simplex of T. Gibbs' inequality gives $\mathrm{KL}(\nu \| \mu) \geq 0$ with equality if and only if $\nu = \mu$, provided μ is strictly positive on that simplex. Softmax teachers assign $p(a) > 0$ on T, so p^T is strictly positive. The unique minimizer on the relative interior is therefore $q^T = p^T$. This identity is equivalent to $q(a) = c\, p(a)$ for $a \in T$ with the same $c = q(T)/p(T)$. It does not force $c = 1$, and L_T does not depend on logits outside T (Lemma 1), so the minimizer is not $q = p$ on V. □

For completeness we record the face without a reset atom that is *not* used as a training constraint. If $q(T \setminus S) = 0$ and $q(I) = 1$, then $q^T = q^I$ and

$$L_T = \sum_{a \in I} q(a) \log \frac{q(a)}{p(a)/p(T)} = -H(q) - \sum_{a \in I} q(a) \log p(a) + \log p(T).$$

The last term does not depend on q. Minimizing over the simplex of I gives $q(a) \propto p(a)$ on I, hence $q(a) = p(a)/p(I)$ for $a \in I$. Then $q(a) \geq p(a)$, with equality for a positive $p(a)$ if and only if $p(I) = 1$. This characterizes the slice; it does not prove that SGD remains on the slice.

B.5 FKL COVERING

On a simplex with $\pi_q = \mathrm{softmax}(\zeta)$ and $L_F = \mathrm{KL}(\pi_p \| \pi_q) = \sum_i \pi_{p,i} \log(\pi_{p,i}/\pi_{q,i})$, the same Jacobian yields

$$\frac{\partial L_F}{\partial \zeta_a} = \sum_i \pi_{p,i} \Big(-\frac{1}{\pi_{q,i}}\Big) \partial \pi_{q,i}/\partial \zeta_a = -\sum_i \pi_{p,i}(\delta_{ia} - \pi_q(a)) = \pi_q(a) - \pi_p(a). \tag{10}$$

Proof of Proposition 3. The cancellation in Lemma 1 shows that $\mathrm{KL}(p^T \| q^T)$ depends on z only through $\{z_a : a \in T\}$. Identity (10) on that simplex, or on $T \cup \{r\}$, is $\partial L_F/\partial \zeta_a = \pi_q(a) - \pi_p(a)$. If $a \in T \setminus S$, $\pi_q(a) = 0$ and $\pi_p(a) > 0$, then $\partial L_F/\partial \zeta_a = -\pi_p(a) < 0$, so gradient descent raises ζ_a.

If $p(T \setminus S) > 0$ and $q^T|_{T \setminus S} \to 0$, the sum $\sum_{a \in T \setminus S} p_a^T \log(p_a^T/q_a^T)$ diverges to $+\infty$. For RKL the same limit is finite: q^T remains a distribution on I and L_T reduces as in Proposition 1.

On $T \cup \{r\}$, teacher reset mass is $p_r = 1 - p(T)$ and student reset mass is $q_r = 1 - q(T)$. Identity (10) on the reset coordinate is $\partial L_F/\partial \zeta_r = q_r - p_r$. Low overlap typically places substantial student mass outside T, so $q_r > p_r$ and descent lowers the reset coordinate, which is a pooled suppression of $S \setminus T$. □

B.6 LEFTOVER RENORMALIZATION

Proposition 1 keeps leftover RKL factors proportional to student mass, so $\varepsilon = q(T \setminus I)$ stays small rather than rising to $p(T \setminus I)$. A reset atom does not reopen that channel. The next statement compares RKL on $T \cup \{r\}$ and on $I \cup \{r\}$ with leftover coordinates frozen at a possibly positive ε; the uncovered face is $\varepsilon = 0$.

Proof of Proposition 4. The assumption $T \setminus I \neq \emptyset$ with $I = T \cap S$ gives $T \setminus I = T \setminus S$. Softmax teachers assign $p(a) > 0$ on T, so $p(T \setminus I) > 0$. Combined with $p(T) < 1$,

$$c := p(I) + 1 - p(T) = 1 - p(T \setminus I)$$

lies in $(0, 1)$, and $p(I) = p(T) - p(T \setminus I) < 1$.

On $T \cup \{r\}$, $\pi_p(a) = p(a)$ for $a \in T$ and $\pi_p(r) = 1 - p(T)$. Freeze q on $T \setminus I$ with total mass $\varepsilon \in [0, 1)$. The leftover RKL terms are then constant in the $I \cup \{r\}$ coordinates. Write w for the remaining masses on $I \cup \{r\}$, so $w(a) = q(a)$ for $a \in I$, $w(r) = 1 - q(I) - \varepsilon$, and $\sum w = 1 - \varepsilon$. Let $m = (p|_I,\, 1 - p(T))$, hence $\sum m = c$, and

$$L_{T,r} = \sum_i w_i \log \frac{w_i}{m_i} + C_\varepsilon = (1 - \varepsilon)\, \mathrm{KL}\Big(\frac{w}{1-\varepsilon} \Big\| \frac{m}{c}\Big) - (1 - \varepsilon) \log \frac{c}{1 - \varepsilon} + C_\varepsilon,$$

where $C_\varepsilon = \sum_{a \in T \setminus I} q(a) \log(q(a)/p(a))$ does not depend on w. Gibbs' inequality yields a unique minimizer $w = (1 - \varepsilon)\, m/c$, i.e.

$$q(a) = \frac{(1 - \varepsilon)\, p(a)}{1 - p(T \setminus I)}\ (a \in I), \qquad 1 - q(I) - \varepsilon = \frac{(1 - \varepsilon)\,(1 - p(T))}{1 - p(T \setminus I)}.$$

Thus $q(a)/p(a) = (1 - \varepsilon)/(1 - p(T \setminus I))$ on I, so $q(a) > p(a)$ if and only if $\varepsilon < p(T \setminus I)$. Setting $\varepsilon = 0$ recovers the uncovered-face formulae. As $\varepsilon \to 0$, the leftover contribution $C_\varepsilon \to 0$ because $x \log(x/p(a)) \to 0$ for each $a \in T \setminus I$.

On $I \cup \{r\}$, leftover mass is part of the reset atom: $\pi_p(a) = p(a)$ on I and $\pi_p(r) = 1 - p(I)$. This pair is a strictly positive probability. Gibbs' inequality gives the unique interior minimizer $q|_I = p|_I$ and $1 - q(I) = 1 - p(I)$. The coordinate $T \setminus I$ is not in the simplex over $I \cup \{r\}$, so the I-target does not depend on ε. If leftover mass is additionally frozen at $\varepsilon \leq 1 - p(I)$, that interior point remains feasible by placing the remaining reset mass on $V \setminus T$. When $T = I$ the two simplices coincide, and the unconstrained interior of $T \cup \{r\}$ is $q|_T = p|_T$, which requires $\varepsilon = p(T \setminus I)$.

The simplex gradient on $I \cup \{r\}$ has only two kinds of coordinates:

$$\frac{\partial L_{I,r}}{\partial \zeta_a} = q(a)\Big(\log \frac{q(a)}{p(a)} - L_{I,r}\Big)\ (a \in I), \qquad \frac{\partial L_{I,r}}{\partial \zeta_r} = (1 - q(I))\Big(\log \frac{1 - q(I)}{1 - p(I)} - L_{I,r}\Big). \tag{11}$$

Identity (11) is (8) on $I \cup \{r\}$. For vocabulary logits z, the simplex logits may be realized as $\zeta_a = z_a$ for $a \in I$ and $\zeta_r = \mathrm{logsumexp}(z|_{V\setminus I})$. Then $\partial L_{I,r}/\partial z_c = q(c)(\log(q(c)/p(c)) - L_{I,r})$ for $c \in I$, and $\partial L_{I,r}/\partial z_c = q(c)(\log((1-q(I))/(1-p(I))) - L_{I,r})$ for $c \notin I$. Tokenwise $\log(q/p)$ therefore appears only on I. The analogous identities on $T \cup \{r\}$ place tokenwise $\log(q/p)$ on every $a \in T$, including $T \setminus S$. □

Lemma 2 (Gradient coincidence)**.** *If $q(T \setminus I) = 0$ and $q(I) > 0$, then $L_T = L_I + \log\big(p(T)/p(I)\big)$. The second term does not depend on z, so L_T and L_I induce the same student-logit gradient. In particular both first-order covering shares are zero whenever they are defined.*

Proof. The assumption gives $q(T) = q(I)$ and $q_j^T = q_j^I$ for $j \in I$, together with $q^T|_{T\setminus I} = 0$. Then

$$L_T = \sum_{j\in I} q_j^I \log \frac{q_j^I}{p_j/p(T)} = \sum_{j\in I} q_j^I \log \frac{q_j^I}{p_j} + \log p(T),$$

$$L_I = \sum_{j\in I} q_j^I \log \frac{q_j^I}{p_j/p(I)} = \sum_{j\in I} q_j^I \log \frac{q_j^I}{p_j} + \log p(I).$$

Subtracting yields $L_T = L_I + \log(p(T)/p(I))$. For $a \in T \setminus I$ the RKL identity (8) is zero because $q_a^T = 0$. Gradients on I therefore coincide, and the covering share of L_T is zero. □

Lemma 2 is the boundary case of Proposition 1 for RKL on T. Leftover renormalization (Proposition 4) does not degenerate: the $T \cup \{r\}$ I-target equals $p|_I$ only when $\varepsilon = p(T \setminus I)$, which is the unconstrained interior rather than the small-ε regime.

C Endpoints of α-divergence

This appendix verifies the two KL limits of (4) used in §4.4. The family is due to Amari Amari (2009); the calculation is for the displayed coefficients and argument order. Throughout, A is finite and $\pi_q(i) > 0$, $\pi_p(i) > 0$ for every $i \in A$.

Write

$$D_\alpha(\pi_q\|\pi_p) = \frac{4}{1-\alpha^2}\left(1 - \sum_{i\in A} \pi_q(i)^{\frac{1-\alpha}{2}} \pi_p(i)^{\frac{1+\alpha}{2}}\right), \qquad \alpha \in (-1, 1).$$

$\alpha \to 1$. Set $\beta = (1-\alpha)/2$, so $\beta \to 0^+$ and $\alpha = 1 - 2\beta$. Then $1 - \alpha^2 = 4\beta(1-\beta)$ and $4/(1-\alpha^2) = 1/(\beta(1-\beta))$. The inner sum is

$$\sum_{i\in A} \pi_p(i)\Big(\frac{\pi_q(i)}{\pi_p(i)}\Big)^\beta = \sum_{i\in A} \pi_p(i)\, e^{\beta \log(\pi_q(i)/\pi_p(i))} = 1 + \beta \sum_{i\in A} \pi_p(i) \log \frac{\pi_q(i)}{\pi_p(i)} + O(\beta^2).$$

Hence

$$1 - \sum_{i\in A} \pi_q(i)^\beta \pi_p(i)^{1-\beta} = \beta\, \mathrm{KL}(\pi_p\|\pi_q) + O(\beta^2),$$

and $D_\alpha = (\beta\, \mathrm{KL}(\pi_p\|\pi_q) + O(\beta^2))/(\beta(1-\beta)) \to \mathrm{KL}(\pi_p\|\pi_q)$.

$\alpha \to -1$. Set $\gamma = (1+\alpha)/2$, so $\gamma \to 0^+$ and $\alpha = 2\gamma - 1$. Then $1 - \alpha^2 = 4\gamma(1-\gamma)$ and $4/(1-\alpha^2) = 1/(\gamma(1-\gamma))$. The inner sum is

$$\sum_{i\in A} \pi_q(i)\Big(\frac{\pi_p(i)}{\pi_q(i)}\Big)^\gamma = 1 + \gamma \sum_{i\in A} \pi_q(i) \log \frac{\pi_p(i)}{\pi_q(i)} + O(\gamma^2).$$

Hence

$$1 - \sum_{i\in A} \pi_q(i)^{1-\gamma} \pi_p(i)^\gamma = \gamma\, \mathrm{KL}(\pi_q\|\pi_p) + O(\gamma^2),$$

and $D_\alpha \to \mathrm{KL}(\pi_q\|\pi_p)$.

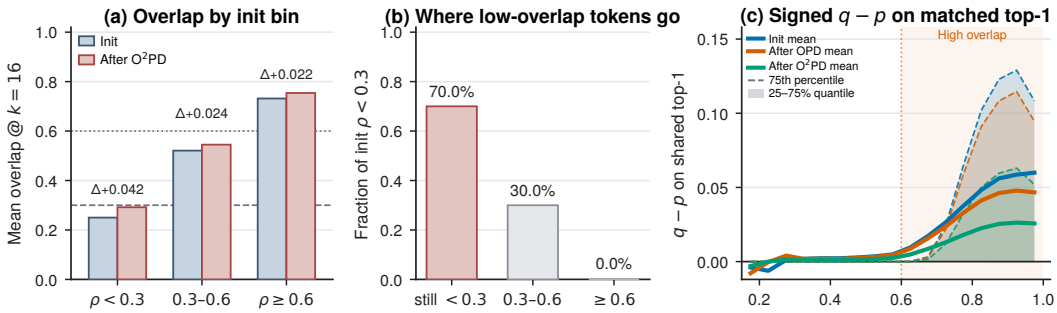

Figure 7: Prefix tokens at $k = 16$ before and after O^{2}PD. (a) Mean overlap by initial-overlap bin. (b) Destination of tokens with initial $\rho < 0.3$. (c) Signed $q - p$ on the shared top-1.

Table 4: Student–teacher pairs used for O^{2}PD.

Student	Teacher
Qwen3-1.7B	Qwen3-4B
DSR1-1.5B	JustRL-1.5B
Qwen3-1.7B	Qwen3-4B-NTRL

D THE TWO FAILURES AFTER O^{2}PD

To test whether O^{2}PD relieves the two failures of top-k OPD in §4.1, we follow prefix tokens at $k = 16$ (Fig. 7). Teacher-support Uncovering is read from mean overlap by initial-overlap bin and from the destination of tokens with initial $\rho < 0.3$. Dominant-Mode Mismatch is read from the signed gap $q - p$ on the shared top-1.

Mean overlap rises in every bin, by 0.042 for $\rho < 0.3$, 0.024 for 0.3–0.6, and 0.022 for $\rho \geq 0.6$. Of the tokens that start below 0.3, 30% move into 0.3–0.6 and none reach ≥ 0.6; 70% remain below 0.3. Under top-k OPD, 87.3% of that cohort remain below 0.3 and none reach ≥ 0.6. O^{2}PD moves more of the low-overlap cohort out of the lowest bin, and most of it is still there. On tokens that share the teacher's top-1, the signed gap $q - p$ after O^{2}PD lies below the gap after top-k OPD and remains positive at high overlap. O^{2}PD therefore relieves Teacher-support Uncovering and Dominant-Mode Mismatch problems relative to top-k OPD.

E TRAINING CONFIGURATION

We train O^{2}PD in VERL Sheng et al. (2025) on 16-NPU Ascend 910C machines. Chat templates disable thinking. The update is the distillation loss alone. Teacher log-probabilities are computed on the student prompt. Table 4 lists the three student–teacher pairs. DSR1-1.5B is DeepSeek-R1-Distill-Qwen-1.5B, and the NTRL teacher is Qwen3-4B-Non-Thinking-RL. All runs use DAPO-MATH-17K with no additional filtering, right truncation, and data-loader seed 42. Table 5 records the optimizer and rollout recipe. Table 6 records the O^{2}PD loss settings used in the reported runs.

O^{2}PD on the NTRL and DSR1 pairs was scheduled for 300 steps, and O^{2}PD with the Qwen3-4B teacher for 200 steps. Downstream scores use intermediate checkpoints at step 60 or 100, not necessarily the last scheduled step.

F EVALUATION

Downstream evaluation uses the same decoding and grading pipeline for distilled checkpoints and reference models. The cross-prefix protocol in §3 reports pass@1 on MATH500 and is separate from the settings below.

Every model is decoded with vLLM. Thinking mode is disabled. All problems use the model's chat template with the user prompt:

Table 5: Shared training recipe for O^{2}PD.

Setting	Value
Hardware	16× Ascend 910C
Data	DAPO-MATH-17K, unfiltered
Optimizer	AdamW, $\beta = (0.9, 0.999)$, weight decay 0.01
Learning rate	2×10^{-6}, cosine decay, min ratio 0.1
Warmup / grad clip	10 steps / 1.0
Train / mini-batch	64 / 16
Micro-batch per NPU	4
Max prompt / response	1536 / 4096
Rollout n / temperature	1 / 1.0
Precision	bfloat16 rollout, FSDP param and optimizer offload
KL loss	off
Loss clamps	max 10.0, $\log q$ min -10.0
Thinking	off

Table 6: O^{2}PD loss settings.

Setting	Value
Top-k size k	16
Prefix fraction ℓ	0.2
Reset atom	on
α	mapped from prefix overlap min and max
Support switch	median of prefix overlap: $T \cup \{r\}$ below, $I \cup \{r\}$ at or above
Scheduled steps	200 (Qwen3-4B teacher); 300 (NTRL and DSR1 pairs)

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

Decoding uses temperature 0.6, top-$p = 0.95$, sampling top-$k = 20$, and a maximum of $31,744$ tokens.

A response is graded by extracting the final `\boxed{}` expression and checking mathematical equivalence against the reference answer with a rule-based verifier. The verifier applies symbolic-equivalence normalization for fractions, radicals, and other equivalent forms. A response without a parseable boxed answer is counted as incorrect.