Rank What Matters: On-Policy Ranking Distillation for Decision Structure Transfer

Anonymous authors
Paper under double-blind review

Abstract

On-policy distillation (OPD) queries a teacher on student-generated trajectories, aligning teacher supervision with the prefixes encountered by the student. Standard OPD objectives express this supervision through numerical differences between token probabilities, while recent variants refine the distribution regions or token positions. At each prefix, however, these probabilities jointly define a local decision structure: which candidates compete and how strongly the teacher prefers one over another. To make this structure explicit, we introduce Rank-OPD, which selects decision-relevant relations, adapts their required separation to teacher preference strength, and complements them with candidate-set mass control. We further show that this ranking mechanism can serve as a plug-and-play component in distribution-matching OPD and propose Rank-OPD-T, which converts relation violations into bounded probability transport and combines the resulting ranking-refined dense target with candidate-set mass matching. Across eight mathematical reasoning benchmarks, Rank-OPD and Rank-OPD-T achieve average scores of 42.01% and 43.69%, respectively, compared with 41.19% for standard OPD. Rank-OPD-T obtains the best score on seven benchmarks. Mechanistic analyses show that ranking supervision corrects preference violations while preserving plausible alternatives, and ablations support both selective relational supervision and distribution target refinement. These results demonstrate that local ranking structure provides effective supervision for OPD.

1 Introduction

Transferring capabilities from large language models to smaller models is important for efficient deployment. Knowledge distillation provides richer supervision than hard labels by training a student on teacher-generated sequences or soft targets(Hinton et al., 2015; Kim & Rush, 2016; Sanh et al., 2020; Jiao et al., 2020). In autoregressive generation, however, a student eventually conditions on its own earlier decisions, which may lead it to prefixes not covered by fixed training trajectories.

On-policy distillation (OPD) instead samples trajectories from the student and queries the teacher on the resulting prefixes (Ross et al., 2011; Lin et al., 2020; Agarwal et al., 2024; Gu et al., 2024; Lu & Lab, 2025). By placing supervision on prefixes induced by the student itself, OPD covers states that fixed-trajectory distillation may miss and more closely matches the contexts encountered at inference. At each such prefix, the teacher provides a probability distribution over possible next tokens. Standard OPD objectives convert this distribution into numerical discrepancies through divergences or sampled estimators, while recent variants refine the distribution regions or token positions on which supervision is applied (Wu et al., 2025; Jung et al., 2025; Jin et al., 2026; Jia et al., 2026; Xing et al., 2026; Ziheng et al., 2026; Xie et al., 2026; Xu et al., 2026). Across these formulations, teacher guidance is represented through numerical differences between token probabilities.

However, at a given prefix, token probabilities do not define independent targets: their relative values jointly determine which candidates compete and which continuation the teacher prefers (Frydenlund et al., 2022; Ding et al., 2027). Existing probability objectives encode this local decision structure only implicitly through numerical values. Making these relations explicit reveals which teacher preferences the student already follows and which still require correction. A useful formulation must

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.

then decide which comparisons are worth transferring and how strongly each should be enforced. We therefore ask a simple question: how can this local decision structure guide OPD?

To instantiate this view, we first propose Rank-OPD. A cumulative-mass Top-p rule divides the teacher's Top-K candidates into head and tail sets and then selects the pairwise relations to transfer. Each retained pair receives a margin based on the teacher log-probability gap. Pairs whose student gap falls below this margin are treated as violated and contribute to the ranking loss. A set-mass constraint complements these relative relations by preventing probability from moving out of the Top-K candidate set. Together, these components supervise the teacher's decision-relevant structure without fixing the student's exact probability allocation.

Beyond direct pairwise supervision, the same ranking signal can also serve as a plug-and-play component within distribution-matching OPD. We therefore propose Rank-OPD-T, which converts each violation into a bounded probability transfer from a lower-ranked to a teacher-preferred candidate. Together, these transfers form a ranking-refined target over the Top-K set: larger violations produce larger adjustments, while satisfied relations leave the original target unchanged. A companion mass-matching term controls the total probability assigned to the Top-K set. Thus, Rank-OPD uses teacher ranking as direct constraints, whereas Rank-OPD-T turns the same relations into a dense target that provides ranking-guided supervision to every Top-K candidate.

Across eight mathematical reasoning benchmarks, Rank-OPD achieves an average score of 42.01%, outperforming all five baselines. Rank-OPD-T further improves the average to 43.69%, exceeding OPD by 2.50 percentage points. Mechanistic analyses explain these gains by showing that ranking supervision corrects unresolved preference violations while preserving plausible alternatives. Complementary ablations validate the joint use of ranking relations and candidate-set mass control, while showing that target refinement works best with a modest transport budget.

2 Preliminaries

2.1 On-Policy Distillation

Let p_θ be a student policy and q a frozen teacher sharing vocabulary $\mathcal{V}$. On-policy distillation (OPD) periodically collects training responses from the student itself: at round r, a snapshot $\bar{\theta}_r$ samples $x \sim \mathcal{D}$ and $y \sim p_{\bar{\theta}_r}(\cdot \mid x)$, inducing student-visited prefixes $s_t = (x, y_{<t})$ (Agarwal et al., 2024; Lu & Lab, 2025). The teacher and student are evaluated at these same prefixes, giving $q_t(v) = q(v \mid s_t)$ and $p_t(v) = p_\theta(v \mid s_t)$. Standard OPD transfers the teacher distribution through a token-level discrepancy D (Agarwal et al., 2024; Gu et al., 2024). For the valid response positions $\mathcal{R}$ in a sampled batch, its empirical objective is

$$\mathcal{L}_{\text{div}}^{(r)}(\theta) = \frac{1}{|\mathcal{R}|} \sum_{t \in \mathcal{R}} D(q_t, p_t). \tag{1}$$

The two commonly used KL-based divergences are forward KL (FKL) and reverse KL (RKL):

$$D_{\text{FKL}}(q_t \| p_t) = \sum_{v \in \mathcal{V}} q_t(v) \log \frac{q_t(v)}{p_t(v)}, \qquad D_{\text{RKL}}(p_t \| q_t) = \sum_{v \in \mathcal{V}} p_t(v) \log \frac{p_t(v)}{q_t(v)}. \tag{2}$$

2.2 Fixed-Margin Pairwise Ranking

Pairwise ranking represents supervision through the relative importance of candidates rather than their exact probabilities (Burges et al., 2005; Reddi et al., 2021; Frydenlund et al., 2022). For a candidate set at position t, a directed preference $i \succ j$ states that candidate i is preferred to candidate j. We collect these ordered pairs in $\mathcal{E}_t = \{(i, j) : i \succ j\}$. Let $z_t(v)$ be the model score for candidate v, and let $p_t(v)$ be the corresponding normalized probability. The score gap for a preferred pair is

$$g_{t,ij} = z_t(i) - z_t(j) = \log p_t(i) - \log p_t(j) = \log \frac{p_t(i)}{p_t(j)}, \tag{3}$$

where the shared softmax normalizer cancels. A fixed-margin objective requires every preferred candidate to exceed its paired alternative by at least the same margin $m_0 > 0$:

$$\ell_t^{\text{fixed}} = \frac{1}{|\mathcal{E}_t|} \sum_{(i,j) \in \mathcal{E}_t} [m_0 - g_{t,ij}]_+, \qquad [u]_+ = \max(u, 0). \tag{4}$$

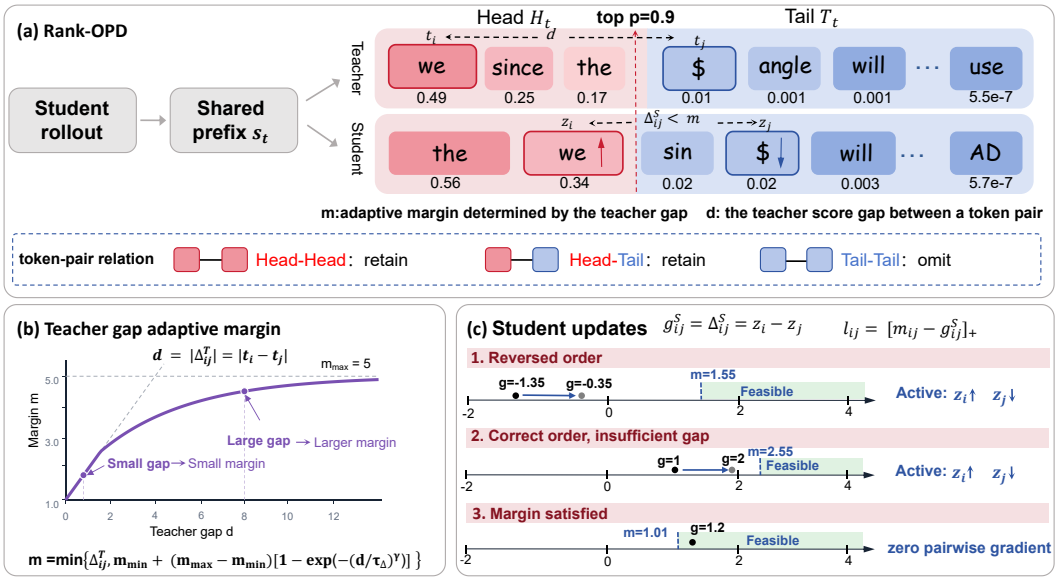

Figure 1: Overview of Rank-OPD. **(a)** Teacher probabilities define a cumulative-mass head. Head–head and head–tail preferences are retained, while tail–tail comparisons are omitted. **(b)** Larger teacher gaps induce larger, saturating margins. **(c)** Reversed or insufficiently separated pairs receive gradients; margin-satisfied pairs do not.

The loss is zero when $\mathcal{E}_t$ is empty. A pair with $g_{t,ij} < m_0$ contributes a positive loss, pushing $z_t(i)$ above $z_t(j)$. Once $g_{t,ij} \geq m_0$, it produces no further update. Thus, satisfying the margin requires $p_t(i)/p_t(j) \geq e^{m_0}$ while leaving the individual probabilities unspecified. However, a shared margin imposes the same separation on near-tied and strongly ordered candidates. In the next section, we refine both relation selection and margin design and incorporate them into OPD.

3 Rank-OPD

We propose Rank-OPD, which only transfers the teacher's local ranking preferences along trajectories sampled from the student policy. As illustrated in Figure 1, Rank-OPD uses cumulative probability mass to partition the teacher's Top-K candidates into head and tail sets, and uses adaptive margins to identify the pairs that require correction.

3.1 Head Selection and Pair Construction

At prefix s_t, let $\mathcal{C}_t = \{v_{t,1}, \ldots, v_{t,k_t}\}$ be the teacher's Top-K candidates, ordered by decreasing $q_t(v)$. We use the original full-vocabulary probabilities without renormalizing them within $\mathcal{C}_t$ and denote their total mass by $Q_t = \sum_{v \in \mathcal{C}_t} q_t(v)$. The head $\mathcal{H}_t$ is the shortest prefix of this ranking whose cumulative mass reaches ρ_H:

$$b_t = \min \left\{ r : \sum_{u=1}^{r} q_t(v_{t,u}) \geq \rho_H \right\}, \qquad \mathcal{H}_t = \{v_{t,1}, \ldots, v_{t,b_t}\}. \tag{5}$$

If the returned Top-K mass $Q_t < \rho_H$, all returned candidates form the head. We retain head–head (HH) relations to preserve the ordering among the principal candidates and head–tail (HT) relations to keep each head candidate above lower-ranked alternatives. Tail–tail (TT) relations are omitted. Appendix A.3 gives the exact pairs and sampling rule.

3.2 Adaptive Pairwise Margins

As reviewed in Section 2, a pairwise margin specifies the minimum student log-odds separation between a preferred candidate and a lower-ranked one. For every retained pair (i, j), the teacher

satisfies $q_t(v_{t,i}) > q_t(v_{t,j})$, and the desired student relation is $z_t(v_{t,i}) - z_t(v_{t,j}) \geq m_{t,ij}$. We derive this margin from the teacher preference gap $\Delta^T_{t,ij} = \log[q_t(v_{t,i})/q_t(v_{t,j})] > 0$:

$$m_{t,ij} = \phi(\Delta^T_{t,ij}) := \min\Big\{\Delta^T_{t,ij},\ m_{\min} + (m_{\max} - m_{\min})\Big[1 - e^{-(\Delta^T_{t,ij}/\tau_\Delta)^\gamma}\Big]\Big\}. \tag{6}$$

The margin increases with the teacher's preference gap and saturates at $m_{\max}$. The cap $m_{t,ij} \leq \Delta^T_{t,ij}$ ensures that a near-tied pair is not required to be more widely separated than it is under the teacher. The constraint is one-sided: once the student reaches the required separation, the pair incurs no loss.

3.3 Rank-OPD Objective

Let $\widehat{\mathcal{P}}_t$ be the sampled preference pairs, $g^S_{t,ij} = z_t(v_{t,i}) - z_t(v_{t,j})$ the student gap, and $P_t = \sum_{v\in\mathcal{C}_t} p_t(v)$ the student's total probability on the candidate set. The objective combines pairwise ranking with candidate-set mass control:

$$\mathcal{L}_{\text{Rank-OPD}} = \frac{1}{|\mathcal{R}|}\sum_{t\in\mathcal{R}}\left[\frac{1}{|\widehat{\mathcal{P}}_t|}\sum_{(i,j)\in\widehat{\mathcal{P}}_t}[m_{t,ij} - g^S_{t,ij}]_+ + \lambda_C[\text{logit}(\beta_C Q_t) - \text{logit}(P_t)]_+\right]. \tag{7}$$

The pairwise term can order candidates within $\mathcal{C}_t$ even when the student assigns little total probability to the set. The second term prevents this failure mode by requiring $P_t \geq \beta_C Q_t$, ensuring that the ranked candidates retain sufficient probability relative to the rest of the vocabulary.

Because Equation (6) gives $m_{t,ij} \leq \Delta^T_{t,ij}$, exact teacher matching satisfies the objective. The objective does not require exact matching: any student distribution with sufficient candidate-set mass and the required HH/HT gaps is feasible, leaving TT order and excess separation unconstrained.

For clarity, let ℓ^{rank}_t denote the pairwise term in Equation (7), and let $\mathcal{A}_t = \{(i,j) \in \widehat{\mathcal{P}}_t : g^S_{t,ij} < m_{t,ij}\}$ be its active pairs. Its gradient is

$$\nabla_{z_t}\ell^{\text{rank}}_t = \frac{1}{|\widehat{\mathcal{P}}_t|}\sum_{(i,j)\in\mathcal{A}_t}(\mathbf{e}_j - \mathbf{e}_i), \tag{8}$$

where $\mathbf{e}_i$ selects candidate $v_{t,i}$. Under gradient descent, each active pair raises the preferred logit and lowers the other; a satisfied pair contributes no pairwise gradient. Proofs and complete gradients are provided in Appendix D.

4 Rank-OPD-T

Rank-OPD can be viewed as a new ranking-based supervision paradigm for OPD. Beyond its standalone use, the same ranking mechanism can also serve as a plug-and-play component within distribution-matching OPD methods. We introduce Rank-OPD-T as such an extension. As shown in Figure 2, it consists of violation-guided target redistribution and candidate-set mass matching. The former adjusts the conditional target within $\mathcal{C}_t$ according to ranking violations, while the latter aligns the total probability assigned to $\mathcal{C}_t$. In this section, we reuse Q_t and P_t from Section 3. Within $\mathcal{C}_t$, let $r^T_t(i) = q_t(v_{t,i})/Q_t$ and $r^S_t(i) = p_t(v_{t,i})/P_t$ be the teacher and student conditional distributions.

4.1 Violation-Guided Target Redistribution

A pairwise hinge loss supervises only violated relations. To retain this selective ranking signal while providing dense supervision over $\mathcal{C}_t$, Rank-OPD-T converts each violation into a modification of the teacher target. It reuses the head, retained pairs, and adaptive margins of Rank-OPD.

For each retained pair $i \succ j$, define its margin deficit as $d_{t,ij} = [m_{t,ij} - g^S_{t,ij}]_+$ and map it to the bounded severity $\nu_{t,ij} = 1 - e^{(-d_{t,ij}/\tau_R)}$. A satisfied pair has zero severity. As the deficit grows, $\nu_{t,ij}$ increases toward one with diminishing increments, so a very large deficit cannot produce an unbounded correction. Each violated pair then transfers probability from the lower-ranked donor j to the teacher-preferred candidate i:

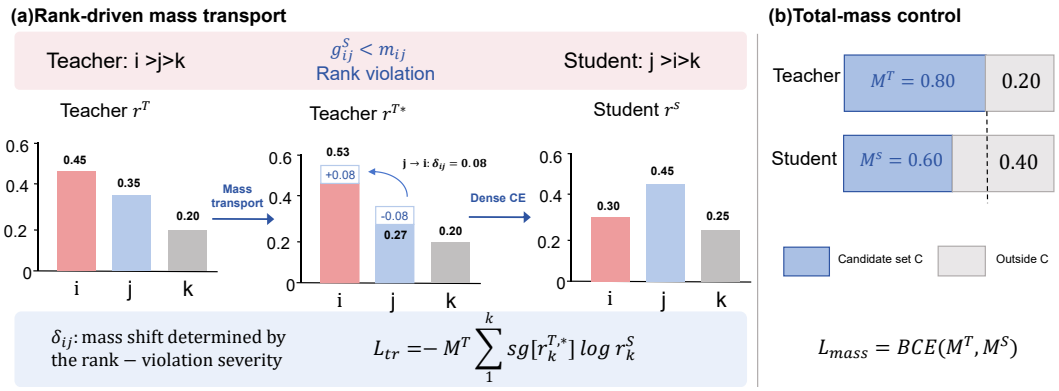

Figure 2: Overview of Rank-OPD-T. **(a)** Ranking violations trigger transfers from lower-ranked candidates to teacher-preferred ones, producing a new target for conditional cross-entropy. **(b)** A binary cross-entropy term matches the student's probability on the Top-K candidate set to the teacher's.

$$\delta_{t,ij} = \rho r_t^T(j) \frac{\nu_{t,ij}}{\max\left(1, \sum_{u:\, u \succ j} \nu_{t,uj}\right)}. \tag{9}$$

Here, $\rho \in [0,1]$ is the maximum fraction of a donor's probability that can be moved. When several violated pairs share donor j, its budget is divided in proportion to their severities, ensuring $\sum_i \delta_{t,ij} \leq \rho r_t^T(j)$. The implementation details are provided in Appendix A.3.

We apply all pairwise transfers simultaneously to construct the redistributed target. For each candidate k, we add transfers for which k is the preferred recipient and subtract transfers for which k is the donor:

$$r_t^{T,*}(k) = r_t^T(k) + \sum_j \delta_{t,kj} - \sum_i \delta_{t,ik}. \tag{10}$$

Thus, every transfer $j \to i$ subtracts $\delta_{t,ij}$ from $r_t^T(j)$ and adds the same amount to $r_t^T(i)$. This shifts the target toward teacher-preferred candidates precisely along relations that the student violates. Total probability is preserved, while the donor budget guarantees nonnegativity; therefore, $r_t^{T,*}$ is a valid conditional distribution over $\mathcal{C}_t$, with $\sum_k r_t^{T,*}(k) = 1$ and $\|r_t^{T,*} - r_t^T\|_1 \leq 2\rho$. If all retained margins are satisfied, no transfer occurs and $r_t^{T,*} = r_t^T$. We detach this target before computing the loss.

4.2 Candidate-Set Mass Matching

Because $r_t^{T,*}$ is normalized within $\mathcal{C}_t$, target redistribution does not constrain the total candidate-set mass P_t. We therefore match it to the teacher mass Q_t using binary cross-entropy:

$$\mathcal{L}_{\text{mass},t} = -Q_t \log P_t - (1 - Q_t) \log(1 - P_t). \tag{11}$$

This term is minimized at $P_t = Q_t$, so it matches the candidate-set mass without imposing any additional ordering within the set.

4.3 Rank-OPD-T Objective

We combine conditional target matching with the candidate-set mass loss in Equation (11):

$$\mathcal{L}_{\text{tr},t} = -Q_t \sum_i \text{sg}[r_t^{T,*}(i)] \log r_t^S(i), \quad \mathcal{L}_{\text{Rank-OPD-T}} = \frac{1}{|\mathcal{R}|} \sum_{t \in \mathcal{R}} \left(\lambda_M \mathcal{L}_{\text{mass},t} + \lambda_T \mathcal{L}_{\text{tr},t}\right). \tag{12}$$

The first loss is the conditional cross-entropy from the detached redistributed target $r_t^{T,*}$ to the student distribution r_t^S over $\mathcal{C}_t$. Its factor Q_t downweights positions where the returned candidates cover little teacher probability. λ_M and λ_T weight mass matching and target matching, respectively. This loss matches the total probability assigned to $\mathcal{C}_t$ and its allocation among the candidates, while leaving the distribution within the complement unspecified. More details are provided in Appendices A and D.

Table 1: Accuracy (%): Avg@4 for MATH-500 and OlympiadBench, Avg@16 otherwise; Avg. is the equal-weight macro average. Best/second-best student scores are **bold**/underlined.

Method	AIME24	AIME25	AIME26	MATH	AMC23	Olymp.	HMMT Feb26	HMMT Nov25	Avg.
Teacher	61.88	46.67	51.46	94.55	93.59	74.96	31.44	43.13	62.21
Student	11.88	8.13	10.42	72.70	46.56	42.17	6.25	3.75	25.23
SFT	22.71	19.58	16.25	81.15	59.06	50.00	13.26	8.75	33.85
GRPO	21.04	18.13	9.38	77.30	51.09	45.73	13.83	4.38	30.11
OPD	33.95	25.62	21.87	86.40	71.09	57.01	18.56	15.00	41.19
AOPD	31.46	24.17	23.54	**87.70**	73.13	50.35	19.51	15.42	40.66
EOPD	35.83	24.17	22.71	86.85	71.72	57.42	19.32	15.83	41.73
Rank-OPD	35.42	23.75	25.00	86.60	72.03	57.49	21.02	14.79	42.01
Rank-OPD-T	**36.04**	**27.08**	**26.46**	87.45	**76.25**	**58.12**	**21.21**	**16.88**	**43.69**
Δ Rank-OPD vs. OPD	+1.47	−1.87	+3.13	+0.20	+0.94	+0.48	+2.46	−0.21	+0.82
Δ Rank-OPD-T vs. OPD	+2.09	+1.46	+4.59	+1.05	+5.16	+1.11	+2.65	+1.88	+2.50

5 Experiments

5.1 Experimental Setup

Implementation details. We train Qwen3-1.7B-Non-Thinking using Qwen3-4B-Instruct-2507 as the teacher (Yang et al., 2025) on English DAPO-Math-17K (Yu et al., 2025) with 4× NVIDIA RTX 6000D GPUs. All methods use batch size 128, learning rate 10^{-6}, and maximum response lengths of 8k/16k tokens for training/evaluation; further details are in Appendix A.

Evaluation protocol. We report Avg@16 on AIME 2024/2025/2026, AMC 2023, HMMT February 2026, and HMMT November 2025, and Avg@4 on MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024) and OlympiadBench (He et al., 2024), using top-$p = 0.7$ (Holtzman et al., 2019). Avg@N is mean correctness over N responses, averaged equally across datasets. We additionally use 64 responses per problem to evaluate Pass@k ($k \leq 64$) (Chen et al., 2021).

Baselines. We compare teacher-response supervised fine-tuning (**SFT**); **GRPO** with verifiable rewards (Shao et al., 2024); standard on-policy distillation (**OPD**) (Agarwal et al., 2024; Gu et al., 2024); entropy-aware **EOPD** (Jin et al., 2026); and advantage-aware **AOPD** (Jia et al., 2026).

5.2 Main Results

Average accuracy. Both ranking-aware variants outperform all baselines on average: Rank-OPD and Rank-OPD-T achieve 42.01% and 43.69%, improving over OPD by 0.82% and 2.50%, respectively. Rank-OPD-T leads on seven of eight benchmarks, with its largest gains on AMC 2023 (+5.16%) and AIME 2026 (+4.59%), while AOPD leads on MATH-500. These results show that ranking structure benefits both selective supervision and dense target refinement.

Diversity under increased sampling. Figure 3 shows that both ranking-aware variants outperform OPD across all three benchmarks and sampling budgets. On AIME 2026, Rank-OPD leads through $k = 32$, while Rank-OPD-T performs best at $k = 64$. On HMMT February 2026, Rank-OPD gains a clearer advantage at larger budgets. On AMC 2023, the ranking-aware variants remain comparable to EOPD, with the methods converging as k increases. Results for more benchmarks are reported in Appendix C.1.

5.3 Why Does Transferring Ranking Structure Work?

To understand how supervision that transfers ranking structure contributes to student performance, we analyze training dynamics and the token positions at which ranking constraints are active. We then examine these positions quantitatively and qualitatively to connect the observed behavior to the design of Rank-OPD.

Ranking transfer preserves room for exploration. Figure 4(a) shows that OPD remains low-entropy, whereas Rank-OPD, EOPD, and Rank-OPD-T rise early and stabilize at higher levels. Rank-OPD-T combines the largest entropy increase with the best average accuracy (Table 1), a pattern consistent with greater room for student exploration.

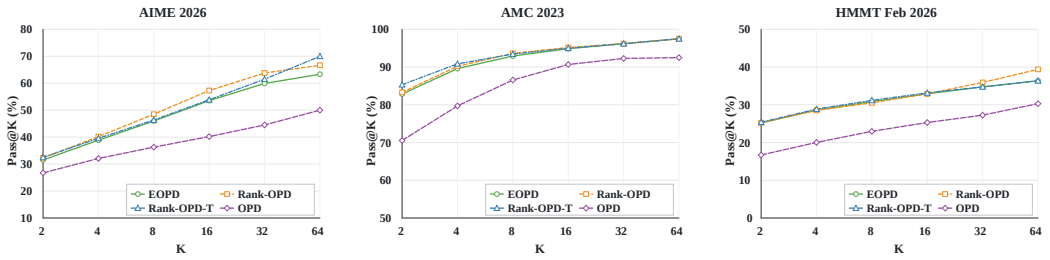

Figure 3: Pass@k curves for EOPD, Rank-OPD, Rank-OPD-T, and OPD on **(a)** AIME 2026, **(b)** AMC 2023, and **(c)** HMMT February 2026, with sampling budgets up to 64.

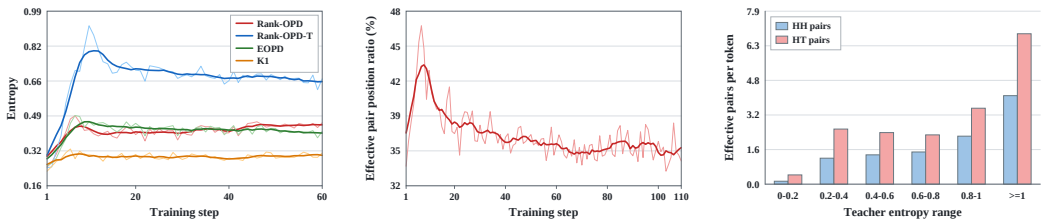

Figure 4: Training dynamics and selective supervision: **(a)** student entropy over the first 60 training steps; **(b)** the fraction of response positions containing an active ranking pair over training; and **(c)** the mean number of active HH/HT pairs per token in each teacher-entropy bin over the first 60 steps. Measurement definitions appear in Appendix C.4.

Ranking supervision concentrates on a subset of positions. We count response-token positions containing at least one effective pair, defined here as a selected pair whose student gap violates its margin. Figure 4(b) shows that this fraction declines after an early peak and stabilizes around 35%. Thus, in the later training regime, roughly 65% of positions contribute no pairwise ranking gradient. This sparsity highlights a key advantage of Rank-OPD: it concentrates supervision on the few token decisions that still require correction. We next examine where the active pairs occur and what decisions they constrain.

Quantitative analysis: active-pair density and teacher uncertainty. Over the first 60 training steps, we group response positions by teacher entropy and measure the mean number of active pairs per token, separately for head–head (HH) and head–tail (HT) relations. Figure 4(c) shows that active-pair density generally increases with teacher entropy and rises sharply in the highest-entropy bin. Moreover, HT pairs outnumber HH pairs in every bin, underscoring the key role of HT constraints in complementing within-head ordering with supervision across the head–tail boundary. Thus, teacher uncertainty is associated with denser ranking supervision, but this aggregate pattern does not reveal what token-level decisions the active constraints encode. We next examine individual positions to characterize these decisions qualitatively.

Qualitative analysis: three decision patterns. Inspection of representative active positions reveals three roles of ranking supervision: reinforcing an under-separated preference, correcting a reversed preference, and preserving flexibility among plausible continuations. Figure 5(a)–(c) illustrates these patterns through teacher and student probabilities, the head–tail partition, and the resulting gradient directions.

(a) Agreement with insufficient separation. The teacher assigns 0.99 probability to "with" and less than 0.01 to "where," while the student assigns 0.90 and 0.09. Although the order is correct, the student gap is below the adaptive margin ($m = 4.689$), so the active HT constraint reinforces it. **(b) Correcting a reversed preference.** The teacher prefers "So" (0.35) to "From" (< 0.01), whereas the student assigns less than 0.01 to "So" and 0.40 to "From." The HT constraint pushes "So" up and "From" down. **(c) Multiple plausible continuations.** The teacher distributes probability across "Alternatively" (0.37), "However" (0.20), and "But" (0.17), while the student instead favors "However" (0.41), followed by "So" (0.25), "But" (0.15), and "Let" (0.12)." The smaller margin ($m = 1.105$) requires limited separation and deactivates once satisfied, preserving relative preferences without copying the exact allocation.



Figure 5: Selective ranking supervision and set-mass control. Colors denote ■ teacher head, ■ teacher tail, ■ student head, and ■ student tail probabilities. **(a)** An agreed but under-separated preference. **(b)** A reversed preference. **(c)** Multiple plausible continuations. **(d)** Mass control preserves more student probability on the teacher's Top-K candidates.

5.4 Ablation Studies

Rank-OPD: relation types and set-mass control. Ranking within the teacher's Top-K support does not by itself prevent probability mass from drifting outside the candidate set (Figure 5(d)). We therefore add set-mass control and compare relation variants using $\lambda_C = 0.1$. As shown in Table 2(a), mass control alone achieves 34.70%. Adding HT relations improves the score to 35.74%, and further adding HH relations reaches 38.06%, a total gain of 3.36% points. The complete objective also outperforms HH+HT without mass control (36.66%) by 1.40%, whereas increasing λ_C to 0.2 slightly reduces the score to 37.59%. These results confirm the complementary roles suggested by Figure 5: HT separates preferred head tokens from tail alternatives, HH transfers ordering within the head, and set-mass control preserves candidate coverage.

Rank-OPD: ranking margin. We compare fixed margins $m \in \{1, 5, 9\}$ with the gap-adaptive rule. Performance decreases as the fixed margin grows, from 37.61% at $m = 1$ to 36.30% and 34.47% at $m = 5$ and $m = 9$, respectively. This suggests that large fixed margins over-separate weak teacher preferences, consistent with Figure 4(c), where ranking constraints are most active at high-entropy positions with multiple plausible continuations. Conversely, a uniformly small margin provides insufficient separation for confident preferences. The gap-adaptive rule balances these cases by assigning pair-specific margins capped by the teacher gap. With $m_{\min} = 1$ and $m_{\max} = 5$, it performs best at $\tau_\Delta = 0.4$, achieving 38.06% and outperforming the strongest fixed-margin baseline by 0.45%. The non-monotonic trend highlights the importance of calibrating the adaptation rate, as overly gradual saturation may weaken separation for confident preferences. Overall, these results support adapting the margin to preference confidence. Top-p sensitivity is reported in Appendix C.3.

Rank-OPD-T: probability-transport budget. We vary $\rho \in \{0, 0.05, 0.10, 0.20\}$ with the relation graph, severity mapping, and loss weights fixed. At $\rho = 0$, transport vanishes and the internal objective recovers the original teacher FKL target on the modeled support (conditions in Appendix D.5), obtaining 38.86%. A modest budget of $\rho = 0.05$ achieves the best result, 40.15%, improving over the unmodified target by 1.29%. This shows that violation-driven transport contributes beyond dense matching to the original teacher distribution. The improvement is not monotonic in transport strength. Increasing ρ to 0.10 lowers the score to 39.07%, although it remains 0.21% above $\rho = 0$. At $\rho = 0.20$, performance falls to 38.69%, which is 0.17% below the no-transport baseline

Table 2: Ablation results: mean Avg@16 (%) over AIME 2024, AIME 2025, AMC 2023, and HMMT February 2026. Full results: Appendix C.2.

(a) Relations and mass

Configuration	Avg.
Mass only	34.70
HH+HT (no mass)	36.66
HT + mass (0.1)	35.74
HH+HT + mass (0.1)	**38.06**
HH+HT + mass (0.2)	37.59

(b) Ranking margin

Configuration	Avg.
Fixed $m = 1$	37.61
Fixed $m = 5$	36.30
Fixed $m = 9$	34.47
Adaptive $\tau_\Delta = 0.3$	37.74
Adaptive $\tau_\Delta = 0.4$	**38.06**
Adaptive $\tau_\Delta = 0.5$	37.32

(c) Rank-OPD-T

ρ	Avg.
0 (FKL)	38.86
0.05	**40.15**
0.10	39.07
0.20	38.69

and 1.46% below the optimum. These results support a conservative donor budget: modest transport strengthens preferences associated with ranking violations while retaining most of the teacher allocation, whereas excessive redistribution can distort useful probability structure even though conditional mass is conserved.

6 Related Work

Knowledge Distillation. Knowledge distillation (KD) transfers a teacher's knowledge through soft targets or generated sequences (Hinton et al., 2015; Kim & Rush, 2016; Sanh et al., 2020; Jiao et al., 2020; Zhao et al., 2022). In autoregressive generation, offline KD learns from fixed dataset or teacher-generated prefixes, whereas the student conditions on its own prefixes at inference, creating exposure bias (Bengio et al., 2015). On-policy distillation (OPD) reduces this mismatch by querying the teacher along student-generated trajectories and providing token-level feedback at states the student actually visits (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024; Lu & Lab, 2025). Recent work further improves OPD through alternative KL-based objectives and optimization strategies (Wu et al., 2025; Jung et al., 2025; Jin et al., 2026; Jia et al., 2026; Xing et al., 2026; Ziheng et al., 2026; Xie et al., 2026; Xu et al., 2026; Yu et al., 2026; Li et al., 2026; Song & Zheng). Across these formulations, teacher guidance is represented through numerical differences between token probabilities. Building on OPD, our method makes the teacher's local ranking structure explicit by transferring decision-relevant relations among candidate tokens at student-generated prefixes.

Reasoning Post-Training. Chain-of-thought supervised fine-tuning and reasoning-trace distillation transfer reasoning ability by imitating complete teacher-generated solutions (Magister et al., 2023; Hsieh et al., 2023; Guo et al., 2025). Reinforcement learning with verifiable rewards (RLVR), including GRPO, instead lets the student explore its own reasoning trajectories (Shao et al., 2024; Guo et al., 2025); however, answer-level rewards are sparse, making credit assignment across long chains difficult and motivating process-level supervision (Cobbe et al., 2021; Lightman et al., 2024). OPD combines on-policy student exploration with dense teacher feedback at every student-visited prefix, making it particularly suitable for reasoning post-training (Agarwal et al., 2024; Jin et al., 2026; Li et al., 2026). Our method further organizes this token-level feedback around relative preferences among candidate tokens, thereby providing structured supervision while retaining flexibility among plausible continuations.

7 Conclusion

In this work, we study how teacher preferences guide token-level decisions in on-policy distillation. Existing probability-matching objectives do not explicitly compare candidate tokens, making it unclear which preferences need correction and how much separation is required. To address this, we introduce Rank-OPD, which selectively transfers decision-relevant ranking relations using teacher-gap-adaptive margins and candidate-set mass control. We further develop Rank-OPD-T, which converts violations of the same relations into bounded probability transport to construct a ranking-refined dense target. Empirically, both approaches improve mathematical reasoning while preserving flexibility among plausible continuations, and our analyses show that the resulting supervision concentrates on reversed or insufficiently separated preferences. Overall, our results demonstrate that the teacher's local ranking structure can support effective on-policy knowledge transfer, either through selective relational constraints or through ranking-refined distributional targets.

AI Use Statement

In this work, we used generative AI tools solely for language polishing and grammatical correction. We did not use generative AI tools for generating research ideas, conducting literature reviews, developing methods, writing or debugging code, performing experiments, analyzing data, creating figures, or formulating scientific claims. All AI-assisted edits were carefully reviewed by the authors to ensure accuracy and to preserve the original meaning. The authors take full responsibility for the final content of this work, including all text, claims, and artifacts.

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. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/e995f98d56967d946471af29d7bf99f1-Paper.pdf.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pp. 89–96, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102363. URL https://doi.org/10.1145/1102351.1102363.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Zhe Ding, Hui Ji, Su Pan, and Miao Chi. Distilling large language models for code generation via ranking supervision. *Neural Networks*, 205:109437, 2027. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2026.109437. URL https://www.sciencedirect.com/science/article/pii/S0893608026008956.

Arvid Frydenlund, Gagandeep Singh, and Frank Rudzicz. Language modelling via learning to rank. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 10636–10644, 2022.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

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

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, Bangkok, Thailand, August 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, 2021. URL https://arxiv.org/abs/2103.03874.

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

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

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.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.372. URL https://aclanthology.org/2020.findings-emnlp.372/.

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. *arXiv preprint arXiv:2603.07079*, 2026.

Seongryong Jung, Suwan Yoon, DongGeon Kim, and Hwanhee Lee. Todi: Token-wise distillation via fine-grained divergence control. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8089–8102, 2025.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

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

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

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6121–6133, 2020.

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.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1773–1781, 2023.

Sashank Reddi, Rama Kumar Pasumarthi, Aditya Menon, Ankit Singh Rawat, Felix Yu, Seungyeon Kim, Andreas Veit, and Sanjiv Kumar. Rankdistil: Knowledge distillation for ranking. In *International Conference on Artificial Intelligence and Statistics*, pp. 2368–2376. PMLR, 2021.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL https://arxiv.org/abs/1910.01108.

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.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models, 2026. *URL https://arxiv. org/abs/2604.00626*.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5737–5755, 2025.

Zhengpeng Xie, Li Lyna Zhang, Zeke Xie, and Mao Yang. Trust region policy distillation. *arXiv preprint arXiv:2607.04751*, 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.

Guo Yu, Wenlin Liu, Yulan Hu, Hao-Xuan Ma, Jun-Peng Jiang, and Han-Jia Ye. Dense supervision, sparse updates: On the sparsity and geometry of on-policy distillation. *arXiv preprint arXiv:2606.13657*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, juncai liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru 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, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 113222–113244. Curran Associates, Inc., 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a4277440d50f1f15d2cb4c14f7e0c0d2-Paper-Conference.pdf.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11943–11952. IEEE, 2022.

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

A Implementation Details

A.1 Training Configuration

We use Qwen3-4B-Instruct-2507 as the frozen teacher and Qwen3-1.7B-Non-Thinking as the student. All methods are trained on the English DAPO-Math-17K corpus using four NVIDIA RTX 6000D GPUs. Table 3 summarizes the shared configuration.

Table 3: Training configuration shared by the main experiments.

Setting	Value
Teacher	Qwen3-4B-Instruct-2507
Student	Qwen3-1.7B-Non-Thinking
Training data	English DAPO-Math-17K
Hardware	$4\times$ NVIDIA RTX 6000D
Global batch size	128
Learning rate	1×10^{-6}
Maximum prompt length	2,048
Maximum response length	8,192
Optimizer / scheduler	AdamW / constant learning-rate schedule
Warmup / weight decay	0 steps (0%) / 0.01
Gradient accumulation / clipping	Dynamic token-based accumulation / global norm 1.0
Numerical precision	BF16 mixed precision with FSDP
Rollout temperature / Top-p	1.0 / 1.0
Rollout refresh frequency	Every training step; rollout count per prompt is method-specific (Table 6)
Training steps / epochs	110 steps / 1 epoch
Random seeds	42

Student responses are sampled on-policy. At every valid response position, teacher and student scores are computed from the same causal prefix; prompt, padding, and invalid prediction positions are excluded from the response-token set $\mathcal{R}$. Teacher candidate scores, pair indices, margins, severities, and transported targets are detached from the computation graph. Teacher probabilities used for head selection and candidate-set mass retain their full-vocabulary normalization. Rollout temperature is distinct from the unit scoring temperature used by the losses.

A.2 Evaluation Configuration and Metrics

The evaluation covers AIME 2024, AIME 2025, AIME 2026, MATH-500, AMC 2023, OlympiadBench, HMMT February 2026, and HMMT November 2025. Table 4 records the protocol used by the main text.

Table 4: Evaluation configuration.

Setting	Value
Maximum response length	16,384
Decoding Top-p	0.7
Decoding temperature	1.0
Samples for the main accuracy table	16; 4 for MATH-500 and OlympiadBench
Samples for Pass@k	64
Pass@k budgets	$k \in \{2, 4, 8, 16, 32, 64\}$

Prompt and chat template. Each evaluation prompt begins with "Solve the following math problem step by step.", followed by the problem statement and an instruction requiring the final answer on its own line in the form $\boxed{\textit{answer}}$. The prompt is rendered with the model's native chat template using `add_generation_prompt=True` and `enable_thinking=False`.

For a benchmark with n problems, let N be the number of sampled responses for each problem and let c_i be the number judged correct for problem i. We report

$$\text{Avg@}N = \frac{100}{n}\sum_{i=1}^{n}\frac{c_i}{N}, \qquad \widehat{\text{Pass@}k} = \frac{100}{n}\sum_{i=1}^{n}\left(1 - \frac{\binom{N-c_i}{k}}{\binom{N}{k}}\right), \tag{A.1}$$

where $\binom{N-c_i}{k} = 0$ when $N - c_i < k$. Avg@N is mean sample-level correctness, whereas Pass@k estimates the probability that at least one of k samples is correct. The macro average in Table 1 weights the eight benchmarks equally.

A.3 Algorithmic Flow of Rank-OPD and Rank-OPD-T

Algorithms 1 and 2 give the end-to-end training updates for Rank-OPD and Rank-OPD-T. Both methods sample student trajectories, score the same causal prefixes with the teacher, and construct the same cumulative-mass relation graph. Their final supervision stages differ: Rank-OPD directly optimizes pairwise hinges, whereas Rank-OPD-T converts margin violations into a detached transported target.

Algorithm 1 Rank-OPD training update for one on-policy batch

Input: Prompts $\mathcal{B}$, rollout policy $p_{\bar{\theta}}$, student p_θ, frozen teacher q
$K, \rho_H, \delta_p, P_{\max}, m_{\min}, m_{\max}, \tau_\Delta, \gamma, \beta_C, \lambda_C$
Output: Updated student parameters θ

Stage 1: Collect on-policy prefixes and teacher candidates

Sample responses $y \sim p_{\bar{\theta}}(\cdot \mid x)$ for prompts $x \in \mathcal{B}$
Construct valid response positions $\mathcal{R}$ and prefixes $s_t = (x, y_{<t})$
Initialize $\mathcal{L} \leftarrow 0$
for each $t \in \mathcal{R}$ **do**
  $\mathcal{C}_t \leftarrow \text{TopK}_K(q(\cdot \mid s_t))$, sorted by decreasing q_t
  $Q_t \leftarrow \sum_{v\in\mathcal{C}_t} q_t(v)$
  if $Q_t \geq \rho_H$ **then**
    $b_t \leftarrow \min\{r : \sum_{u=1}^{r} q_t(v_{t,u}) \geq \rho_H\}$
  else
    $b_t \leftarrow |\mathcal{C}_t|$
  end if

Stage 2: Construct selective ranking supervision

  $\mathcal{P}_t \leftarrow \{(i,j) : i < j,\ i \leq b_t,\ \log[q_t(v_{t,i})/q_t(v_{t,j})] > \delta_p\}$
  Sample $\widehat{\mathcal{P}}_t \subseteq \mathcal{P}_t$ uniformly, with $|\widehat{\mathcal{P}}_t| \leq P_{\max}$
  for each $(i,j) \in \widehat{\mathcal{P}}_t$ **do**
    $\Delta^T_{t,ij} \leftarrow \log[q_t(v_{t,i})/q_t(v_{t,j})]$
    $m_{t,ij} \leftarrow \phi(\Delta^T_{t,ij})$ using Equation (6)
  end for
  $\ell^{\text{rank}}_t \leftarrow |\widehat{\mathcal{P}}_t|^{-1}\sum_{(i,j)\in\widehat{\mathcal{P}}_t}[m_{t,ij} - g^S_{t,ij}]_+$
  Set $\ell^{\text{rank}}_t \leftarrow 0$ when $\widehat{\mathcal{P}}_t = \varnothing$
  $P_t \leftarrow \sum_{v\in\mathcal{C}_t} p_t(v)$
  $\ell^{\text{set}}_t \leftarrow [\text{logit}(\beta_C Q_t) - \text{logit}(P_t)]_+$
  $\mathcal{L} \leftarrow \mathcal{L} + \ell^{\text{rank}}_t + \lambda_C \ell^{\text{set}}_t$
end for

Stage 3: Optimize the student

$\theta \leftarrow \text{OptimizerStep}\big(\theta, \mathcal{L}/|\mathcal{R}|\big)$

Algorithm 2 Rank-OPD-T training update for one on-policy batch

Input: Prompts $\mathcal{B}$, rollout policy $p_{\tilde{\theta}}$, student p_θ, frozen teacher q
Relation parameters from Algorithm 1; $\rho, \tau_R, \epsilon_\nu, \lambda_{\text{mass}}, \lambda_{\text{tr}}$
Output: Updated student parameters θ

Stage 1: Construct the shared on-policy relation graph
Sample student responses and construct valid prefixes $\{s_t : t \in \mathcal{R}\}$
Initialize $\mathcal{L} \leftarrow 0$
for each $t \in \mathcal{R}$ **do**
  Construct $\mathcal{C}_t, Q_t, b_t, \widehat{\mathcal{P}}_t$, and $m_{t,ij}$ as in Algorithm 1
  $P_t \leftarrow \sum_{v \in \mathcal{C}_t} p_t(v)$
  $r_t^T(i) \leftarrow q_t(v_{t,i})/Q_t$, $r_t^S(i) \leftarrow p_t(v_{t,i})/P_t$
Stage 2: Build the violation-guided transport target
  for each $(i,j) \in \widehat{\mathcal{P}}_t$ **do**
    $d_{t,ij} \leftarrow [m_{t,ij} - g^S_{t,ij}]_+$
    $\nu_{t,ij} \leftarrow 1 - \exp(-d_{t,ij}/\tau_R)$
  end for
  $\mathcal{E}_t^+ \leftarrow \{(i,j) \in \widehat{\mathcal{P}}_t : \nu_{t,ij} > \epsilon_\nu\}$
  Initialize $\Delta r_t(k) \leftarrow 0$ for every $k \in \{1, \dots, |\mathcal{C}_t|\}$
  for each donor j appearing in $\mathcal{E}_t^+$ **do**
    $S_{t,j} \leftarrow \sum_{u:(u,j)\in\mathcal{E}_t^+} \nu_{t,uj}$
    for each $(i,j) \in \mathcal{E}_t^+$ **do**
      $\delta_{t,ij} \leftarrow \rho r_t^T(j)\nu_{t,ij}/\max(1, S_{t,j})$
      $\Delta r_t(i) \leftarrow \Delta r_t(i) + \delta_{t,ij}$
      $\Delta r_t(j) \leftarrow \Delta r_t(j) - \delta_{t,ij}$
    end for
  end for
  $r_t^{T,*} \leftarrow \text{sg}[r_t^T + \Delta r_t]$
Stage 3: Match internal allocation and total candidate mass
  $\mathcal{L}_{\text{mass},t} \leftarrow -Q_t \log P_t - (1 - Q_t)\log(1 - P_t)$
  $\mathcal{L}_{\text{tr},t} \leftarrow -Q_t \sum_i r_t^{T,*}(i) \log r_t^S(i)$
  $\mathcal{L} \leftarrow \mathcal{L} + \lambda_{\text{mass}}\mathcal{L}_{\text{mass},t} + \lambda_{\text{tr}}\mathcal{L}_{\text{tr},t}$
end for
$\theta \leftarrow \text{OptimizerStep}\big(\theta, \mathcal{L}/|\mathcal{R}|\big)$

For exact relation construction, let the teacher candidates be sorted so that $q_t(v_{t,1}) \geq \cdots \geq q_t(v_{t,k_t})$. After the Top-$p$ boundary b_t is obtained from Equation (5), near-ties are removed with threshold $\delta_p \geq 0$:

$$\mathcal{P}_t = \left\{ (i,j) : 1 \leq i < j \leq k_t,\ i \leq b_t,\ \log \frac{q_t(v_{t,i})}{q_t(v_{t,j})} > \delta_p \right\}. \tag{A.2}$$

Thus $\mathcal{P}_t$ contains head–head and head–tail relations and excludes tail–tail relations. If $Q_t < \rho_H$, the returned candidate set is treated as the head. We either retain all valid relations or sample uniformly without replacement up to a cap $P_{\max}$; the realized set is denoted $\widehat{\mathcal{P}}_t$. An empty pair set contributes zero ranking loss but does not disable the set-mass branch.

For Rank-OPD-T, retain edges with $\nu_{t,ij} > \epsilon_\nu$, $0 \leq \epsilon_\nu < 1$, and set all other severities to zero before applying Equation (9). All donor budgets are computed from the original r_t^T, and transfers are applied simultaneously. If the retained edge set is empty or $\rho = 0$, then $r_t^{T,*} = r_t^T$.

Numerically stable evaluation. For a proper nonempty candidate set, define $g_t^M = \text{LSE}_{v\in\mathcal{C}_t} z_t(v) - \text{LSE}_{v\notin\mathcal{C}_t} z_t(v)$. The mass loss and conditional probabilities are evaluated as

$$\begin{aligned} \mathcal{L}_{\text{mass},t} &= Q_t \,\text{softplus}(-g_t^M) + (1 - Q_t)\,\text{softplus}(g_t^M), \\ \log r_t^S(i) &= z_t(v_{t,i}) - \text{LSE}_{u\in\mathcal{C}_t} z_t(u), \\ \log r_t^T(i) &= \log q_t(v_{t,i}) - \text{LSE}_{j=1}^{k_t} \log q_t(v_{t,j}). \end{aligned} \tag{A.3}$$

When $\mathcal{C}_t = \mathcal{V}$, the mass branch is identically zero. Invalid empty candidate returns are excluded using the same valid-position mask for every method.

A.4 Baselines and Reproduction Configuration

Table 5 summarizes the baselines used in the main comparison. Table 6 separates their method-specific settings from the shared configuration in Tables 3 and 4.

Table 5: Overview of the evaluated baselines.

Method	Policy	Trajectory source	Training signal
SFT	Off-policy	Teacher responses	Token-level cross-entropy on teacher-generated solutions
GRPO	On-policy	Student rollouts	Verifiable outcome reward with group-relative policy optimization
OPD	On-policy	Student rollouts	Token-level teacher supervision on student-visited prefixes; $k1$ estimator
EOPD	On-policy	Student rollouts	Entropy-aware combination of reverse and forward KL (Jin et al., 2026)
AOPD	On-policy	Student rollouts	Policy-gradient updates for positive advantages and teacher matching for non-positive advantages (Jia et al., 2026)

Table 6: Method-specific reproduction settings. Shared settings follow Table 3. Rollout reuse denotes the number of actor optimization epochs per rollout batch.

Method	Objective or Routing	Key Reproduction Settings
SFT	Response-token cross-entropy on one teacher response per prompt	Teacher decoding uses temperature 1.0, top-$p = 1.0$, no top-K truncation, and a maximum response length of 8,192 tokens.
GRPO	Group-relative policy gradient with rule-based mathematical rewards $\{+1, -1, 0\}$	$G = 8$ rollouts; advantages are centered and normalized by the group standard deviation; reuse $= 1$.
OPD	Sampled-token k_1 estimator of $D_{\mathrm{KL}}(p_\theta \Vert q)$, with detached advantage $A_t = \log q(y_t \mid s_t) - \log p_{\text{old}}(y_t \mid s_t)$	One rollout with sampled-token support only. Sampled log-probabilities are lower-bounded at -10, and the k_1 estimator is clipped to $[-10, 10]$. No teacher-support importance sampling; reuse $= 1$.
EOPD	Sampled-token k_1 surrogate at every valid token, combined with top-16 FKL when the full-vocabulary teacher entropy satisfies $H_q \geq 1.0$	Top-16 teacher probabilities are not renormalized. The k_1 and gated FKL terms both have weight 1.0 and use the all-valid-token mean denominator. One rollout; reuse $= 1$.
AOPD	k_1 policy-gradient update when $q(y_t \mid s_t) - p_{\text{old}}(y_t \mid s_t) > 0$; otherwise, top-16 truncated FKL	The two mutually exclusive branches have weights $1.0/1.0$ and share the all-valid-token mean denominator; advantages are not normalized. Top-16 teacher probabilities are not renormalized, and the FKL branch is divided by 16. One rollout; reuse $= 1$.
Rank-OPD	Adaptive HH/HT pairwise ranking with candidate-set mass control	Candidate and relation parameters: $K = 16$, $\rho_H = 0.9$, $\delta_p = 0$, and $P_{\text{max}} = 1024$. Adaptive-margin parameters: $m_{\text{min}} = 1$, $m_{\text{max}} = 5$, $\tau_\Delta = 0.4$, and $\gamma = 1$. Set-mass parameters: $\beta_C = 1$ and $\lambda_C = 0.1$.
Rank-OPD-T	Violation-guided target transport with candidate-set mass matching	Uses the same candidate, relation, and adaptive-margin settings as Rank-OPD. Transport parameters: $\rho = 0.05$, $\tau_R = 0.5$, and $\epsilon_\nu = 0$. Loss weights: $\lambda_{\text{mass}} = \lambda_{\text{tr}} = 1$.

The OPD result in the main comparison uses the sampled-token reverse-KL configuration above. A separate forward-KL control uses the unnormalized teacher Top-16 support, is optimized by direct backpropagation without a policy-gradient ratio, and is not the reported OPD baseline.

B Limitations

Although the proposed methods operate only on a restricted candidate set, constructing and evaluating pairwise relations introduces additional computation compared with standard OPD. In particular, the number of candidate relations can grow quadratically with the selected Top-K set, and the relation filtering, margin computation, and probability transport steps add implementation overhead beyond a conventional token-level divergence. This overhead remains manageable under the small candidate sets used in our experiments, but may become more noticeable when larger candidate sets are required.Further improvements in relation selection and computational efficiency therefore remain an important direction for future work.

C Additional Results

C.1 Additional Pass-at-k Results

Figure 6 complements Figure 3 with the remaining competition benchmarks evaluated using 64 samples per problem.

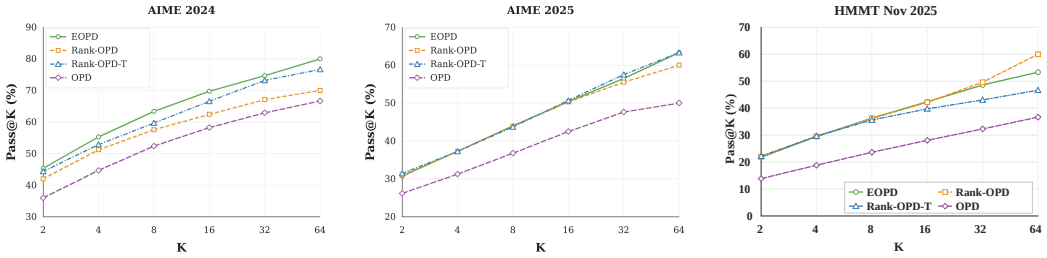

Figure 6: Additional Pass@k curves for EOPD, Rank-OPD, Rank-OPD-T, and OPD on **(a)** AIME 2024, **(b)** AIME 2025, and **(c)** HMMT November 2025, with sampling budgets up to 64.

C.2 Detailed Ablation Results

The main text reports the mean Avg@16 over AIME 2024, AIME 2025, AMC 2023, and HMMT February 2026. Tables 7–9 retain this four-benchmark definition for the Avg. column and expose the individual benchmark results. All entries use the first reported value (Avg@16); the accompanying Pass@k values are omitted.

Table 7: Detailed ablation of relation types and candidate-set mass control. Entries are Avg@16 (%).

Relations	λ_C	AIME24	AIME25	AMC23	HMMT Feb26	Avg.
Mass only	0.1	29.79	23.54	69.38	16.10	34.70
HH+HT	0	29.79	27.71	70.00	19.13	36.66
HT	0.1	30.63	24.17	68.28	19.89	35.74
HH+HT	0.1	35.42	23.75	72.03	21.02	**38.06**
HH+HT	0.2	30.21	26.46	74.38	19.32	37.59

Table 8: Detailed ablation of the ranking margin. Adaptive variants use $m_{\min} = 1$ and $m_{\max} = 5$. Entries are Avg@16 (%).

Margin	Parameter	AIME24	AIME25	AMC23	HMMT Feb26	Avg.
Fixed	$m = 1$	31.88	23.96	75.47	19.13	37.61
Fixed	$m = 5$	33.13	24.58	68.75	18.75	36.30
Fixed	$m = 9$	27.71	23.75	69.38	17.05	34.47
Adaptive	$\tau_\Delta = 0.3$	34.38	24.58	71.72	20.27	37.74
Adaptive	$\tau_\Delta = 0.4$	35.42	23.75	72.03	21.02	**38.06**
Adaptive	$\tau_\Delta = 0.5$	32.29	27.50	70.94	18.56	37.32

Table 9: Detailed ablation of the donor transport budget in Rank-OPD-T. Entries are Avg@16 (%).

ρ	AIME24	AIME25	AMC23	HMMT Feb26	Avg.
0 (FKL)	33.33	27.71	74.53	19.89	38.86
0.05	36.04	27.08	76.25	21.21	**40.15**
0.10	33.96	26.04	75.62	20.64	39.07
0.20	31.46	25.83	77.03	20.45	38.69

C.3 Top-p Ablation

Here Top-p denotes the cumulative teacher-mass threshold ρ_H used to form the head decision set, not the decoding Top-p in Table 4. All other training and evaluation settings are held fixed.

Table 10: Ablation of the head-coverage threshold ρ_H. Entries are Avg@16 (%).

ρ_H	AIME24	AIME25	AMC23	HMMT Feb26	Avg.
0.70	31.88	25.42	71.25	19.89	37.11
0.80	32.50	27.08	73.28	19.70	**38.14**
0.90	35.42	23.75	72.03	21.02	38.06
0.95	32.92	25.21	72.97	20.27	37.84

C.4 Margin Dynamics During Training

Figure 7 reports the mean adaptive pairwise margin over sampled valid pairs at each training step. The margin decreases sharply during the first few steps and then fluctuates around a comparatively stable level near 3.7, with no sustained upward drift over the remaining training window. This behavior is consistent with the on-policy prefix distribution changing most rapidly at the beginning of training and becoming more stable thereafter.

For the diagnostics in Figure 4, an active pair satisfies $g^S_{t,ij} < m_{t,ij}$. The active-pair position ratio is

$$\frac{1}{|\mathcal{R}|}\sum_{t\in\mathcal{R}} \mathbb{I}(|\mathcal{A}_t| > 0)\,, \qquad \mathcal{A}_t = \{(i,j) \in \widehat{\mathcal{P}}_t : g^S_{t,ij} < m_{t,ij}\}. \tag{C.1}$$

The entropy-binned analysis reports the number of active HH and HT pairs per valid response token in each teacher-entropy bin. Pair availability and pair activity are distinct: a position may contain valid relations while all of them already satisfy their margins.

C.5 Case Studies with the 1.7B Student

We compare seven Qwen3-1.7B-Non-Thinking models on the same OlympiadBench problem and the same sampled response. The example isolates a consequential modeling choice: the two removed balls come from separate bags and are therefore independent. Each box retains the decisive reasoning steps and final answer; repetitive intermediate enumeration and self-checking are marked by an ellipsis. Markdown headings, lists, and mathematical expressions are typeset for readability. Checkpoint paths and other evaluation metadata are omitted.

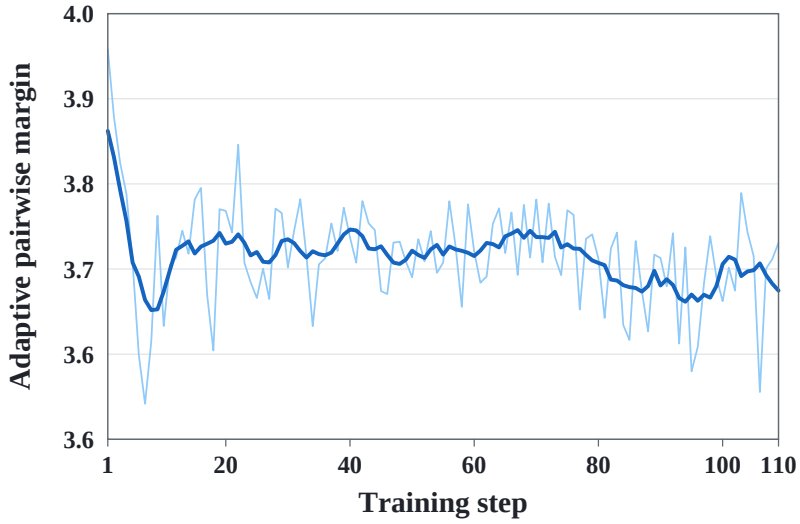

Figure 7: Adaptive pairwise margin over 110 training steps. The light curve shows the per-step mean and the dark curve shows its smoothed trend.

Problem

Solve the following math problem step by step.
Billy and Crystal each have a bag of 9 balls. The balls in each bag are numbered from 1 to 9. Billy and Crystal each remove one ball from their own bag. Let b be the sum of the numbers on the balls remaining in Billy's bag. Let c be the sum of the numbers on the balls remaining in Crystal's bag. Determine the probability that b and c differ by a multiple of 4.
The final answer must be on its own line in the exact form $\boxed{answer}$, where only the answer is placed inside the braces.

Benchmark: OlympiadBench **Ground-truth answer:** $\frac{7}{27}$

SFT

Answer process.

Step 3: Count the number of valid pairs
Billy and Crystal each remove one ball from their own bag.
Total number of possible outcomes:
- Billy has 9 choices (balls 1 to 9)
- Crystal has 9 choices (balls 1 to 9)

Total: $9 \times 9 = 81$ possible pairs.
We want the number of pairs (x, y) such that $|x - y| \equiv 0 \pmod 4$, i.e., $x \equiv y \pmod 4$
We can compute the number of such pairs.

Step 4: Count how many numbers from 1 to 9 are congruent to each residue mod 4
Numbers from 1 to 9:
1, 2, 3, 4, 5, 6, 7, 8, 9
Residues mod 4:
- 1 mod 4: 1, 5 → two numbers
- 2 mod 4: 2, 6 → two numbers
- 3 mod 4: 3, 7 → two numbers
- 0 mod 4: 4, 8 → two numbers

. . .

Wait — that's 8 numbers, which is correct.

. . .

There are 4 residue classes: 0,1,2,3
So total number of valid pairs:

$$4 \times 4 = 16$$

Step 5: Total number of possible pairs

$$9 \times 9 = 81$$

So the probability that $|x - y| \equiv 0 \pmod 4$ is:

$$\frac{16}{81}$$

Final Answer

$$\boxed{\frac{16}{81}}$$

Incorrect step: The response assigns the residue-1 class as $\{1, 5\}$, omitting 9, and consequently reports $4 \times 4 = 16$ favorable pairs.

Result. Incorrect; the extracted answer is $\frac{16}{81}$.
Analysis. The response correctly treats the two draws as independent and uses $9 \times 9 = 81$ total outcomes, but it initially omits the ball numbered 9 from the residue-1 class. This changes the favorable count from $3^2 + 2^2 + 2^2 + 2^2 = 21$ to 16.

GRPO

Answer process.

Step 2: Total sum of all balls
The sum of numbers from 1 to 9 is:

$$\text{Total sum} = \frac{9 \times 10}{2} = 45$$

- Billy removes one ball, so his remaining sum is $b = 45 - x$, where x is the number on the ball he removed.
- Crystal also removes one ball, so her remaining sum is $c = 45 - y$, where y is the number on the ball she removed.

So:

$$b = 45 - x, \quad c = 45 - y$$

We want:

$$|b - c| \equiv 0 \mod 4 \Rightarrow |(45 - x) - (45 - y)| \equiv 0 \mod 4 \Rightarrow |y - x| \equiv 0 \mod 4$$

So we need:

$$|y - x| \equiv 0 \mod 4 \Rightarrow y \equiv x \mod 4$$

Step 3: Count the number of pairs (x, y) such that $x \equiv y \mod 4$
We need to count the number of pairs (x, y) where x and y are distinct numbers from 1 to 9, and $x \equiv y \mod 4$.
Let's list the numbers from 1 to 9 and group them by their modulo 4:

- Numbers $\equiv 0$ mod 4: 4, 8 $\rightarrow$ 2 numbers
- Numbers $\equiv 1$ mod 4: 1, 5, 9 $\rightarrow$ 3 numbers
- Numbers $\equiv 2$ mod 4: 2, 6 $\rightarrow$ 2 numbers
- Numbers $\equiv 3$ mod 4: 3 $\rightarrow$ 1 number

. . .

Now, we count the number of pairs (x, y) where $x \equiv y \mod 4$:

- For each residue class, the number of pairs is:
- 0 mod 4: $2 \times 1 = 2$ (since we can pair each of the 2 numbers with each of the 2 numbers)
- 1 mod 4: $3 \times 2 = 6$
- 2 mod 4: $2 \times 1 = 2$
- 3 mod 4: $1 \times 0 = 0$ (no numbers $\equiv 3$ mod 4 to pair with others)

Total number of valid pairs:

$$2 + 6 + 2 + 0 = 10$$

But wait — we need to consider that **x** and **y** are **distinct**. So we must subtract the cases where $x = y$.
But in our setup, **x** and **y** are the numbers removed by Billy and Crystal, so they must be **distinct**.
So we need to count the number of ordered pairs (x, y) where $x \neq y$ and $x \equiv y \mod 4$.

. . .

- For 0 mod 4: 2 numbers. Number of ordered pairs where $x \equiv y \mod 4$ and $x \neq y$: $2 \times 1 = 2$
- For 1 mod 4: 3 numbers. Number of ordered pairs where $x \equiv y \mod 4$ and $x \neq y$: $3 \times 2 = 6$
- For 2 mod 4: 2 numbers. Number of ordered pairs where $x \equiv y \mod 4$ and $x \neq y$: $2 \times 1 = 2$
- For 3 mod 4: 1 number. No pairs where $x \equiv y \mod 4$ and $x \neq y$: 0

Total number of valid pairs:

$$2 + 6 + 2 = 10$$

Step 4: Total number of possible pairs
Billy and Crystal each remove one ball from their own bag, so there are:

$$9 \times 8 = 72 \text{ possible pairs}$$

Step 5: Compute the probability

$$\text{Probability} = \frac{\text{Number of valid pairs}}{\text{Total number of pairs}} = \frac{10}{72} = \frac{5}{36}$$

Final Answer

$$\boxed{\frac{5}{36}}$$

Incorrect step: The response writes the residue-3 class as $\{3\}$ rather than $\{3, 7\}$ and also imposes $x \neq y$, yielding $10/72$.

Result. Incorrect; the extracted answer is $\frac{5}{36}$.
Analysis. GRPO incorrectly couples the two bags by requiring the removed values to be distinct, producing a denominator of $9 \times 8 = 72$. It also omits 7 from the residue-3 class in its initial count. Both choices contradict the independence of the two bags.

OPD

Answer process.

$$|b - c| \equiv 0 \pmod 4 \Rightarrow b_{\text{Crystal}} - b_{\text{Billy}} \equiv 0 \pmod 4$$

So the difference between the numbers removed by Crystal and Billy must be divisible by 4.
Let $x = b_{\text{Billy}}$, $y = b_{\text{Crystal}}$, where x and y are distinct numbers from 1 to 9 (since each removes one ball).
We are to compute the probability that $y - x \equiv 0 \pmod 4$.

Step 2: Count total possible outcomes
Billy and Crystal each remove one ball from their own bag. The balls are distinct (since each removes one), so the total number of possible outcomes is:

$$9 \times 8 = 72$$

Because Billy has 9 choices, and Crystal has 8 remaining choices (since one ball was removed from her own bag).
We now count the number of such outcomes where $y - x \equiv 0 \pmod 4$, i.e., $y \equiv x \pmod 4$.

Step 3: Count favorable outcomes
We need to count the number of pairs (x, y) such that $x \in \{1, 2, \ldots, 9\}$, $y \in \{1, 2, \ldots, 9\}$, $x \neq y$, and $x \equiv y \pmod 4$.
But note: x and y are the numbers removed from their respective bags. Since the balls are removed from different bags, the values x and y are independent (i.e., the choice of Billy's ball and Crystal's ball are independent), and we are only concerned with their **mod 4 equivalence**.

...

Residue classes modulo 4
- $\equiv 0 \pmod 4$: 4, 8 → 2 numbers
- $\equiv 1 \pmod 4$: 1, 5, 9 → 3 numbers
- $\equiv 2 \pmod 4$: 2, 6 → 2 numbers
- $\equiv 3 \pmod 4$: 3, 7 → 2 numbers

...

Now, for each residue class, we can compute how many pairs (x, y) such that $x \equiv y \pmod 4$, and $x \neq y$.
But note: the values x and y are **not** necessarily distinct in the sense that x and y are from different bags, but the problem says "each removes one ball", so the values are distinct (since one ball is removed from each bag). So $x \neq y$, because Billy and Crystal removed different balls.
So we are counting ordered pairs (x, y) such that $x \in \{1, \ldots, 9\}$, $y \in \{1, \ldots, 9\}$, $x \neq y$, and $x \equiv y \pmod 4$.

...

- Total pairs with $x \equiv y \pmod 4$: $\sum_{r=0}^{3} (n_r \times n_r)$
- Subtract the cases where $x = y$: for each r, n_r values of $x \equiv r$, so n_r pairs where $x = y$

So total favorable pairs:

$$\sum_{r=0}^{3} (n_r^2 - n_r) = \sum_{r=0}^{3} n_r(n_r - 1)$$

Now compute:
- $n_0 = 2 \to 2(2 - 1) = 2$
- $n_1 = 3 \to 3(3 - 1) = 6$
- $n_2 = 2 \to 2(2 - 1) = 2$
- $n_3 = 2 \to 2(2 - 1) = 2$

Sum: $2 + 6 + 2 + 2 = 12$
So there are **12** favorable ordered pairs (x, y) such that $x \equiv y \pmod 4$, $x \neq y$

...

Step 4: Compute probability
Total number of possible outcomes: $9 \times 8 = 72$
So probability:

$$\frac{12}{72} = \frac{1}{6}$$

Final Answer

$$\boxed{\frac{1}{6}}$$

Incorrect step: After acknowledging separate bags, the response nevertheless assumes $x \neq y$ and uses $9 \times 8 = 72$ total outcomes.

Result. Incorrect; the extracted answer is $\frac{1}{6}$.
Analysis. Although the response briefly notes that the balls come from different bags, it still enforces $x \neq y$ and removes the nine equal-value outcomes. The resulting count $12/72$ is therefore a without-replacement calculation for what is actually a pair of independent draws.

AOPD

Answer process.

- Billy's remaining sum: $b = 45 - x$, where x is the number removed by Billy.
- Crystal's remaining sum: $c = 45 - y$, where y is the number removed by Crystal.

But since both remove one ball, the sum of the balls removed is $x + y$, and total remaining sum is:

$$b + c = (45 - x) + (45 - y) = 90 - (x + y)$$

We want to compute the probability that $|b - c| \equiv 0 \pmod 4$

That is:

$$|b - c| \equiv 0 \pmod 4$$
$$|(45 - x) - (45 - y)| \equiv 0 \pmod 4 \Rightarrow |y - x| \equiv 0 \pmod 4$$

So we are to compute the probability that $|x - y| \equiv 0 \pmod 4$, where x and y are distinct integers from the set $\{1, 2, \ldots, 9\}$, with x removed by Billy, y by Crystal.

But we must be careful: Billy and Crystal remove **one ball each**, so the values x and y are distinct, and the removed balls are from the same set (both 1 to 9).

So total number of possible outcomes:

- Billy removes one of 9 balls: 9 choices
- Crystal removes one of the remaining 8 balls: 8 choices
- So total outcomes = $9 \times 8 = 72$

Now, for each such pair of removed values (x, y), compute $|x - y| \mod 4$, and count how many such pairs have $|x - y| \equiv 0 \pmod 4$.

We want:

$$\text{Number of pairs } (x, y), \text{ distinct, } |x - y| \equiv 0 \pmod 4$$
$$\Rightarrow x \equiv y \pmod 4$$

...

$$\sum_{r=0}^{3} (|S_r|^2 - |S_r|) = \sum_{r=0}^{3} |S_r|^2 - \sum_{r=0}^{3} |S_r|$$

Now, compute $|S_r|$ for each residue:

- Residue 0: {4,8} → 2
- Residue 1: {1,5,9} → 3
- Residue 2: {2,6} → 2
- Residue 3: {3,7} → 2

So:

- $|S_0|^2 - |S_0| = 2^2 - 2 = 4 - 2 = 2$
- $|S_1|^2 - |S_1| = 3^2 - 3 = 9 - 3 = 6$
- $|S_2|^2 - |S_2| = 2^2 - 2 = 4 - 2 = 2$
- $|S_3|^2 - |S_3| = 2^2 - 2 = 2$

Sum total:

$$2 + 6 + 2 + 2 = 12$$

So total number of pairs (x, y) such that $x \equiv y \pmod 4$, $x \neq y$, and x and y are distinct, and both removed from their respective bags?

Wait — no! This counts the number of **ordered pairs** (x, y) where both are in the same residue class, and Billy removes x, Crystal removes y, and $x \neq y$

But does this include all such cases where the sum b and c differ by a multiple of 4?

Yes — because we showed earlier that $|b - c| \equiv 0 \pmod 4$ iff $|x - y| \equiv 0 \pmod 4$, and since $x \equiv y \pmod 4$, this is true.

So the total number of such favorable outcomes is 12?

But wait — the total number of possible outcomes is $9 \times 8 = 72$

So probability is $\frac{12}{72} = \frac{1}{6}$

...

Final Answer

$$\boxed{\frac{1}{6}}$$

Incorrect step: The response assumes that Crystal has only eight choices after Billy removes x, as if both draws came from one shared bag.

Result. Incorrect; the extracted answer is $\frac{1}{6}$.

Analysis. AOPD repeatedly assumes that Crystal chooses from eight remaining values after Billy's draw. This treats two separate bags as a shared pool, excludes valid events with $x = y$, and leads to the incorrect ratio $12/72$.

EOPD

Answer process.
Then the sum of the remaining balls in Billy's bag is:

$$b = 45 - x$$

The sum of the remaining balls in Crystal's bag is:

$$c = 45 - y$$

We are to compute the probability that:

$$|b - c| \equiv 0 \pmod 4$$

Substitute:

$$|b - c| = |(45 - x) - (45 - y)| = |y - x|$$

So we are looking for the probability that:

$$|y - x| \equiv 0 \pmod 4$$

So the problem reduces to: **What is the probability that the absolute difference between the numbers removed by Billy and Crystal is divisible by 4?**
Note: Billy and Crystal each remove one ball from the set {1, 2, ..., 9}. We assume that the balls are removed **without replacement**, and each ball is equally likely to be removed.

Step 2: Total number of ways to remove one ball each
Billy has 9 choices for his ball, and Crystal has 8 choices for her ball (since one ball is removed from the bag). So total possible pairs:

$$9 \times 8 = 72$$

We now count how many such pairs (x, y) satisfy $|x - y| \equiv 0 \pmod 4$
We can compute this by iterating over all possible values of x and y, and counting when $|x - y| \equiv 0 \pmod 4$
But since x and y are from 1 to 9, and distinct, we can do this by checking all pairs with $x \neq y$, and count how many have $|x - y| \equiv 0 \pmod 4$

Step 3: Count pairs where $|x - y| \equiv 0 \pmod 4$
We consider all pairs (x, y) with $x \neq y$, and $|x - y| \equiv 0 \pmod 4$
We will go through values of x from 1 to 9, and for each x, count the number of y such that $y \equiv x \pmod 4$, and $y \neq x$
Note: $y \in \{1, 2, \ldots, 9\}$, and $y \neq x$
Let's group numbers from 1 to 9 by their residue modulo 4:
- $\equiv 0 \pmod 4$: 4, 8 → 2 numbers
- $\equiv 1 \pmod 4$: 1, 5, 9 → 3 numbers
- $\equiv 2 \pmod 4$: 2, 6 → 2 numbers
- $\equiv 3 \pmod 4$: 3, 7 → 2 numbers

So:
- Residue 0: 2 numbers
- Residue 1: 3 numbers
- Residue 2: 2 numbers
- Residue 3: 2 numbers

For each x, we want $y \equiv x \pmod 4$, and $y \neq x$

...

Now total number of pairs where $|x - y| \equiv 0 \pmod 4$:

$$2(res0) + 6(res1) + 2(res2) + 2(res3) = 12$$

...

So the probability is:

$$\frac{12}{72} = \frac{1}{6}$$

...

We have 12 such ordered pairs.
So yes.
Final answer:

$$\boxed{\frac{1}{6}}$$

Incorrect step: The response explicitly adopts "without replacement" sampling and computes the denominator as $9 \times 8 = 72$.

Result. Incorrect; the extracted answer is $\frac{1}{6}$.
Analysis. EOPD makes the same structural error as OPD and AOPD: it assumes a without-replacement draw from one shared set. Its residue-class counting is internally consistent under that false premise, but the denominator and favorable set are both wrong for independent bags.

Rank-OPD

Answer process.

We are given that Billy and Crystal each have a bag of 9 balls numbered from 1 to 9.

- Billy removes one ball from his bag.
- Crystal removes one ball from her bag.
- Let b be the sum of the numbers on the balls remaining in Billy's bag.
- Let c be the sum of the numbers on the balls remaining in Crystal's bag.

We are to find the probability that b and c **differ by a multiple of 4**, i.e., $|b - c| \equiv 0 \pmod 4$.

Step 1: Total sum of numbers from 1 to 9

$$\text{Sum} = 1 + 2 + \cdots + 9 = \frac{9 \cdot 10}{2} = 45$$

Each person removes one ball, so the total sum of the balls removed is:

- Billy removes one ball: say x, so Billy's remaining sum is $45 - x$
- Crystal removes one ball: say y, so Crystal's remaining sum is $45 - y$

So:

- $b = 45 - x$
- $c = 45 - y$

We want $|b - c| \equiv 0 \pmod 4$

Compute $b - c$:

$$b - c = (45 - x) - (45 - y) = y - x$$

So we want $|y - x| \equiv 0 \pmod 4$, i.e., $|y - x|$ is divisible by 4.

So the problem reduces to: **What is the probability that the absolute difference between the numbers removed by Billy and Crystal is divisible by 4?**

Step 2: Total number of possible outcomes

Billy and Crystal each remove one ball from 1 to 9.

- Billy removes one ball: 9 choices
- Crystal removes one ball: 9 choices

So total number of possible outcomes: $9 \times 9 = 81$

Step 3: Favorable outcomes

We want the number of pairs (x, y), where $x \in \{1, 2, \ldots, 9\}$, $y \in \{1, 2, \ldots, 9\}$, such that $|x - y| \equiv 0 \pmod 4$

That is, $|x - y|$ is divisible by 4.

So we want $x \equiv y \pmod 4$

So we count the number of pairs (x, y) such that $x \equiv y \pmod 4$

For each residue class modulo 4, we count how many numbers from 1 to 9 fall into that class.

Step 4: Count numbers in each residue class modulo 4

Numbers from 1 to 9:

- Residue 0 mod 4: 4, 8 → 2 numbers
- Residue 1 mod 4: 1, 5, 9 → 3 numbers
- Residue 2 mod 4: 2, 6 → 2 numbers
- Residue 3 mod 4: 3, 7 → 2 numbers

...

Step 5: For each residue class, number of pairs where $x \equiv y \pmod 4$

For residue 0: 2 choices for x, 2 choices for y → $2 \times 2 = 4$ pairs For residue 1: $3 \times 3 = 9$ pairs For residue 2: $2 \times 2 = 4$ pairs For residue 3: $2 \times 2 = 4$ pairs

Total favorable outcomes:

$$4 + 9 + 4 + 4 = 21$$

Step 6: Total possible outcomes

$9 \times 9 = 81$

Step 7: Probability

$$\text{Probability} = \frac{21}{81} = \frac{7}{27}$$

Final Answer

$$\boxed{\frac{7}{27}}$$

Result. Correct; the extracted answer is $\frac{7}{27}$.

Analysis. Rank-OPD correctly models the removals as 81 ordered outcomes. It identifies residue-class sizes $(2, 3, 2, 2)$ modulo 4, obtains $2^2 + 3^2 + 2^2 + 2^2 = 21$ favorable pairs, and reduces $21/81$ to $7/27$.

Rank-OPD-T

Answer process.

We want to compute the probability that $b - c \equiv 0 \pmod 4$

So compute:

$$b - c = (45 - x) - (45 - y) = y - x$$

So we want:

$$y - x \equiv 0 \pmod 4$$

That is:

$$y \equiv x \pmod 4$$

We are to find the probability that the two numbers x and y, selected uniformly at random from the set {1,2,...,9}, are congruent modulo 4.

Note: The balls are removed **without replacement** from their own bags, and both remove one ball each.

So, we are choosing:

- Billy removes one ball from {1 to 9} → possible values $x \in \{1, \ldots, 9\}$
- Crystal removes one ball from {1 to 9} → possible values $y \in \{1, \ldots, 9\}$

But the values are **distinct** because they are removed from the same bag, and we assume that both remove different balls (not specified otherwise, but in reality, since each removes one ball, and the total balls are 9, they are removing from the same pool — but balls are distinct, so removing one from Billy's bag and one from Crystal's bag — but are the removals independent? Are they disjoint or overlapping?

Wait — the problem says:

> *Billy and Crystal each remove one ball from their own bag.*

So:

- Billy removes one ball from his own 9-ball bag → one ball is taken out of his bag.
- Crystal removes one ball from her own 9-ball bag → one ball is taken out of her bag.

But there is no mention of the balls being shared or being removed from each other — so they are **independent** choices.

. . .

So we can compute the number of pairs $(x, y) \in \{1, \ldots, 9\} \times \{1, \ldots, 9\}$ such that $x \equiv y \pmod 4$, and divide by 81 (since there are 9 choices for x, 9 for y).

So total number of ordered pairs: $9 \times 9 = 81$

We now count how many such pairs satisfy $x \equiv y \pmod 4$

That is, for each residue class modulo 4, count how many numbers from 1 to 9 fall into that class, and then square the count (since each value x with residue r, y with residue r gives one such pair).

. . .

Residues:

- Residue 0: 4, 8 → 2 numbers
- Residue 1: 1, 5, 9 → 3 numbers
- Residue 2: 2, 6 → 2 numbers
- Residue 3: 3, 7 → 2 numbers

. . .

So for each residue class:

- Number of values $x \equiv r \pmod 4$: $r = 0, 1, 2, 3$: counts are 2, 3, 2, 2

Now, number of pairs (x, y) with $x \equiv y \pmod 4$ is:

$$\sum_{r=0}^{3} (\text{count}_r) \times (\text{count}_r) = 2^2 + 3^2 + 2^2 + 2^2 = 4 + 9 + 4 + 4 = 21$$

Total number of pairs: $9 \times 9 = 81$

So the probability is:

$$\frac{21}{81} = \frac{7}{27}$$

. . .

Final Answer:

$$\boxed{\frac{7}{27}}$$

Result. Correct; the extracted answer is $\frac{7}{27}$.

Analysis. Rank-OPD-T explicitly revisits whether the draws should be coupled, resolves that they come from separate bags, and then applies the same correct $21/81$ residue-class calculation. The self-correction is verbose but reaches the correct probabilistic model and answer.

D Supplementary Theoretical Analysis

Throughout this section, we fix a valid student prefix and a nonempty teacher candidate set $\mathcal{C}$, and omit the position subscript. Teacher and student logits are finite on a shared effective vocabulary $\mathcal{V}$. Unless a boundary case is stated explicitly, $\mathcal{C}$ is a proper subset of $\mathcal{V}$, so $0 < P, Q < 1$. Relations, margins, severities, and transport budgets are treated as stop-gradient quantities. Write $\mathcal{E} = \widehat{\mathcal{P}}$ and $\mathcal{E}^+ = \{(i,j) \in \mathcal{E} : \nu_{ij} > \epsilon_\nu\}$, where $0 \le \epsilon_\nu < 1$, and assume $\rho \in [0,1]$.

D.1 Teacher Compatibility of the Adaptive Margin

Proposition 1 (Monotonicity and teacher compatibility). *Let $0 < m_{\min} \le m_{\max}$ and $\tau_\Delta, \gamma > 0$. The mapping ϕ in Equation (6) is nondecreasing on $(0,\infty)$ and satisfies*

$$0 < \phi(d) \le \min\{d, m_{\max}\}. \tag{D.1}$$

If $r^S = r^T$, every retained pair has zero violation and the transported target equals r^T.

Proof. Define the unclipped branch

$$h(d) = m_{\min} + (m_{\max} - m_{\min})\left[1 - \exp(-(d/\tau_\Delta)^\gamma)\right]. \tag{D.2}$$

For $d > 0$,

$$h'(d) = (m_{\max} - m_{\min}) \exp[-(d/\tau_\Delta)^\gamma] \, \frac{\gamma d^{\gamma-1}}{\tau_\Delta^\gamma} \ge 0. \tag{D.3}$$

Both d and $h(d)$ are positive and nondecreasing, so their pointwise minimum is positive and nondecreasing. The definition also gives $\phi(d) \le d$ and $\phi(d) \le m_{\max}$. If $r^S = r^T$, conditional normalization cancels in every pair ratio, giving $g_{ij}^S = \Delta_{ij}^T \ge m_{ij}$. Hence all deficits, severities, and transfers are zero, and Equation (10) yields $r^{T,*} = r^T$. □

The nominal parameter $m_{\min}$ is not a lower bound on the final clipped margin: for sufficiently small positive teacher gaps, $\phi(d) = d < m_{\min}$. At full teacher matching $p = q$, the set hinge also vanishes because $P = Q \ge \beta_C Q$. This establishes a feasible teacher-matching point, but neither uniqueness nor convergence to that point.

D.2 Gradients of Rank-OPD

Fix the realized pair set $\widehat{\mathcal{P}}$ and choose the zero subgradient at hinge boundaries. An active pair contributes -1 at its preferred endpoint and $+1$ at its lower-ranked endpoint, up to the pair-averaging factor, which proves Equation (8). The ranking gradient is therefore zero-sum over candidate logits and vanishes outside $\mathcal{C}$.

For the set term, define

$$z_{\mathcal{C}} = \log \sum_{v \in \mathcal{C}} e^{z(v)}, \qquad z_{\bar{\mathcal{C}}} = \log \sum_{v \notin \mathcal{C}} e^{z(v)}, \qquad g^M = z_{\mathcal{C}} - z_{\bar{\mathcal{C}}} = \operatorname{logit}(P). \tag{D.4}$$

For $0 < \beta_C Q < 1$, its derivative is

$$\frac{\partial \ell^{\text{set}}}{\partial z(v)} = \mathbb{I}(P < \beta_C Q) \begin{cases} -p(v)/P, & v \in \mathcal{C}, \\ +p(v)/(1-P), & v \notin \mathcal{C}. \end{cases} \tag{D.5}$$

Thus an active set constraint raises candidate logits relative to the complement under gradient descent. The full per-position Rank-OPD gradient is the ranking gradient plus λ_C times Equation (D.5).

D.3 Feasibility and Boundedness of Rank Transport

Proposition 2 (Donor budget). *For any donor j, let $S_j = \sum_{u:(u,j)\in\mathcal{E}^+} \nu_{uj}$. The total outflow defined by Equation (9) satisfies*

$$D_j := \sum_{i:(i,j)\in\mathcal{E}^+} \delta_{ij} = \rho r^T(j) \frac{S_j}{\max(1, S_j)} \le \rho r^T(j). \tag{D.6}$$

Proof. All edges leaving j share the same denominator. Summing their transfers gives the equality, and $S_j / \max(1, S_j) \le 1$ gives the bound. □

Proposition 3 (Valid transported target)**.** *If r^T is a probability distribution and $0 \le \rho \le 1$, then*

$$\sum_{k=1}^{|\mathcal{C}|} r^{T,*}(k) = 1, \qquad r^{T,*}(k) \ge 0. \tag{D.7}$$

If $F = \sum_{(i,j)\in\mathcal{E}^+} \delta_{ij}$ is the total transported mass, then

$$F \le \rho, \qquad \|r^{T,*} - r^T\|_1 \le 2F \le 2\rho. \tag{D.8}$$

Proof. Summing Equation (10) over k, each edge contributes one positive and one negative term, so the total remains one. For any candidate k, Equation (D.6) limits its outflow to $\rho r^T(k)$, while its inflow is nonnegative. Therefore

$$r^{T,*}(k) \ge r^T(k) - \rho r^T(k) = (1-\rho) r^T(k) \ge 0. \tag{D.9}$$

Moreover, $F = \sum_j D_j \le \rho \sum_j r^T(j) = \rho$. Each edge changes two coordinates by δ_{ij}, so the triangle inequality gives $\|r^{T,*} - r^T\|_1 \le 2F \le 2\rho$. □

A candidate may be both a winner and a donor, so opposing flows can cancel and the realized target displacement can be strictly smaller than $2F$. Because all transport occurs within the conditional distribution, multiplying $r^{T,*}$ by the unchanged teacher coverage Q preserves the unnormalized candidate-set mass.

D.4 Candidate-Set Mass Matching

Proposition 4 (Optimum and gradient of the mass loss)**.** *Let $M^S = \sigma(g^M)$ and fix $M^T \in (0,1)$. Equation (11) satisfies*

$$\frac{\partial \mathcal{L}_{\text{mass}}}{\partial g^M} = M^S - M^T, \qquad \frac{\partial^2 \mathcal{L}_{\text{mass}}}{\partial (g^M)^2} = M^S(1 - M^S) > 0. \tag{D.10}$$

Its unique finite optimum in the scalar set logit is $g^M = \text{logit}(M^T)$, equivalently $M^S = M^T$.

Proof. Substitute $M^S = \sigma(g^M)$ into the binary cross-entropy and differentiate. Strict convexity and the first-order condition give the stated optimum. □

The binary cross-entropy equals a Bernoulli KL plus the teacher binary entropy, so its minimum is generally nonzero. Applying the chain rule gives the full-vocabulary gradient

$$\frac{\partial \mathcal{L}_{\text{mass}}}{\partial z(v)} = (P - Q) \begin{cases} p(v)/P, & v \in \mathcal{C}, \\ -p(v)/(1-P), & v \notin \mathcal{C}. \end{cases} \tag{D.11}$$

Thus the mass branch corrects both under-allocation and over-allocation to $\mathcal{C}$ independently of the conditional target.

D.5 Zero-Transport Limit and Its Relation to Forward KL

When $\rho = 0$, all transfers vanish and $r^{T,*} = r^T$. With $\lambda_{\text{mass}} = \lambda_{\text{tr}} = 1$, the per-position objective becomes

$$-\sum_{v\in\mathcal{C}} q(v) \log p(v) - (1-Q)\log(1-P). \tag{D.12}$$

Up to teacher-dependent entropy, Equation (D.12) is forward KL on the coarsened support containing each candidate in $\mathcal{C}$ and one aggregate complement category. Full-vocabulary forward KL admits the decomposition

$$D_{\text{KL}}(q\|p) = D_{\text{KL}}\big((Q, 1-Q)\,\|\,(P, 1-P)\big) + Q D_{\text{KL}}(r^T\|r^S) + (1-Q) D_{\text{KL}}\big(q(\cdot \mid \bar{\mathcal{C}})\|p(\cdot \mid \bar{\mathcal{C}})\big). \tag{D.13}$$

The zero-transport objective contains the first two terms, up to constants, but leaves the fine-grained distribution within $\bar{\mathcal{C}}$ unspecified. It equals full-vocabulary FKL only when $\mathcal{C} = \mathcal{V}$ (or when the omitted conditional term is constant with respect to the student). Unequal branch weights yield a reweighted decomposition rather than ordinary FKL.

D.6 Gradient Change Induced by Transport

The conditional loss has candidate-logit gradient

$$\frac{\partial \mathcal{L}_{\text{tr}}}{\partial z(v_k)} = Q\big(r^S(k) - r^{T,*}(k)\big), \tag{D.14}$$

and zero derivative for $v \notin \mathcal{C}$. Combining this with Equation (D.11), the complete per-position gradient is

$$\frac{\partial \ell^{\text{RT}}}{\partial z(v)} = \begin{cases} \lambda_{\text{mass}}(P-Q)r^S(k) + \lambda_{\text{tr}}Q(r^S(k) - r^{T,*}(k)), & v = v_k \in \mathcal{C}, \\ -\lambda_{\text{mass}}(P-Q)p(v)/(1-P), & v \notin \mathcal{C}. \end{cases} \tag{D.15}$$

Let $G_k^0 = Q(r^S(k) - r^T(k))$ denote the conditional-loss gradient without transport. Then

$$\begin{aligned} G_k - G_k^0 &= -Q\big(r^{T,*}(k) - r^T(k)\big) \\ &= -Q \sum_{j:(k,j)\in\mathcal{E}^+} \delta_{kj} + Q \sum_{i:(i,k)\in\mathcal{E}^+} \delta_{ik}. \end{aligned} \tag{D.16}$$

Each edge $i \leftarrow j$ therefore adds $-Q\delta_{ij}$ to the winner's gradient and $+Q\delta_{ij}$ to the donor's gradient. Under gradient descent, the edge increases the winner–donor logit gap by $2Q\delta_{ij}$ before multiplication by the learning rate and λ_{tr}. Summing Equation (D.16) over all candidates gives zero, so transport changes only relative conditional update directions at the current logits.

This zero-sum correction does not imply that a finite logit update preserves the student's candidate-set mass. For example, with $p = (1/2, 1/4, 1/4)$, $\mathcal{C} = \{1, 2\}$, and the candidate-only direction $h = (1, -1, 0)$,

$$\frac{d}{d\eta} P(z + \eta h)\Big|_{\eta=0} = (1-P)\sum_{v\in\mathcal{C}} p(v)h(v) = \frac{1}{16}. \tag{D.17}$$

Target-budget conservation and the effect of a student-logit update are therefore distinct; the mass branch is still needed to control P relative to Q.

D.7 Limits of the Guarantees

Directional transport does not guarantee that every teacher ordering is preserved in the final transported target. Consider $r^T = (0.40, 0.35, 0.25)$, candidate logits $(5, 0, 3)$, and a head containing all three candidates. Set $\delta_p = 0$, $m_{\text{min}} = 0.2$, $m_{\text{max}} = 2$, $\tau_\Delta = 0.7$, $\gamma = 1.3$, $\tau_R = 0.6$, $\rho = 0.3$, and $\epsilon_\nu = 0$. Only the pair $(2, 3)$ violates its margin, giving $\nu_{23} \approx 0.9962$, $\delta_{23} \approx 0.0747$, and

$$r^{T,*} \approx (0.40, 0.4247, 0.1753). \tag{D.18}$$

The flow correctly favors token 2 over token 3 but reverses the original order between tokens 1 and 2. The guarantees therefore concern validity, bounded displacement, and edge-wise update direction, not global order preservation.

All derivatives above are local to the detached surrogate used at one update. Recomputing rankings, severities, and transported targets after an update changes the next surrogate. With pair sampling, the guarantees apply to the realized edge set; unsampled violations are not directly corrected in that step. These position-level properties do not imply global convergence, generalization to unseen prefixes, or sequence-level performance improvements.