TEACHING WHAT THE STUDENT LACKS: REASONING-MODE TRANSPORT FOR ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) trains a student model on its own reasoning trajectories with feedback from a stronger teacher model. However, OPD does not explicitly address imbalanced coverage of reasoning modes at student-visited states. We introduce Reasoning-Mode Transport On-Policy Distillation (RMT-OPD) to prioritize under-covered teacher modes. At states visited along student trajectories that fail final-answer verification, RMT-OPD samples multiple next steps from the student and an answer-conditioned teacher. It uses shared semantic embeddings and optimal transport to compare their local distributions and assign mode deficits. These deficits weight teacher steps during maximum-likelihood training, while subsequent states remain student-generated. RMT-OPD requires only teacher-generated text, without access to teacher logits. Under an idealized discrete-mode model, the weighted deficit equals the teacher probability mass missing from the student. On four mathematical reasoning benchmarks, RMT-OPD performs best among the compared methods and improves Avg4 by 1.7 percentage points over uniform weighting under matched per-state sampling. Fixed-state analyses show smaller teacher–student mode-coverage gaps for ours than MOPD, while gains persist with a larger teacher and on code reasoning tasks.

1 INTRODUCTION

Distilling the reasoning capabilities of large language models (LLMs) into smaller students can reduce deployment costs (Gu et al., 2024; Ko et al., 2024; Shing et al., 2025). Sequence-level distillation transfers these capabilities by training the student to imitate complete teacher-generated reasoning trajectories (Kim & Rush, 2016). During its own generation, the student may reach intermediate states absent from those trajectories. Sequence-level imitation provides no direct teacher feedback at these states. On-policy distillation (OPD) addresses this mismatch by obtaining teacher supervision at student-visited prefixes (Agarwal et al., 2024). OPD thus specifies where teacher feedback is obtained, but leaves open which reasoning steps to emphasize at each state.

A student-visited state may admit several valid next steps that reflect different reasoning modes. For a mathematical problem, one step may set up an equation while another uses an invariant. At inference time, Self-Consistency samples diverse reasoning paths and aggregates their answers (Wang et al., 2023). SOPD provides step-level teacher corrections along student-generated trajectories (Sun et al., 2026). MOPD uses successful and failed student rollouts to give the teacher more context (Yu et al., 2026). Neither method explicitly compares which reasoning modes the teacher and student produce at the same state. A single correction may therefore reinforce a familiar mode rather than expand the student's coverage.

Identifying underrepresented teacher modes requires comparing local teacher and student reasoning distributions. Optimal transport provides a way to align distributions without pairing samples in advance. Lohit and Jones apply it to teacher–student feature distributions (Lohit & Jones, 2022). MultiLevelOT aligns logit distributions at the token and sequence levels across different tokenizers (Cui et al., 2025b). These methods build their transport objectives from teacher features or logits. Our black-box teacher instead returns sampled reasoning steps as text. Different steps can express the same mode in different words, so comparing their wording alone is insufficient. Existing optimal-transport objectives therefore cannot directly measure reasoning-mode coverage from text at a student-visited state.

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.

To make this comparison using only generated text, we propose Reasoning-Mode Transport On-Policy Distillation (RMT-OPD). RMT-OPD collects states visited along student trajectories that fail final-answer verification. At each state, we independently sample several next reasoning steps from the current student and a teacher given the correct answer. We embed these steps in a shared semantic space to form empirical local distributions for both models. Optimal transport aligns the distributions, and the resulting coupling yields a deficit for each teacher mode that reflects how well the student covers it. These deficits give greater weight to teacher steps from under-covered modes during standard maximum-likelihood training. The teacher steps are not executed, so subsequent states continue to come from the student's own trajectories. As the student changes, fresh rollouts update the local comparisons and the modes that receive stronger supervision.

Our contributions are as follows:

- **A reasoning-mode coverage perspective for on-policy distillation.** We view teacher knowledge at a student-visited state as a distribution over valid reasoning modes and prioritize those under-covered by the current student.
- **Reasoning-Mode Transport On-Policy Distillation.** RMT-OPD samples local teacher and student reasoning steps and uses optimal transport to give greater weight to teacher modes the student under-covers. It accesses the teacher only through text generation and trains the student with standard maximum likelihood.
- **Empirical evidence for mode-level supervision.** RMT-OPD performs best among the compared methods on four mathematical benchmarks, with a 1.7-point Avg4 gain over uniform weighting under matched per-state sampling.

2 METHOD

RMT-OPD starts from student trajectories that fail final-answer verification and treats their visited prefixes as on-policy states for supervision. At each state, it independently samples local reasoning steps from the current student and a gold-conditioned teacher. It embeds these steps in a shared semantic space and uses optimal transport to identify teacher modes with weak student support. The resulting deficits give teacher steps from under-covered modes more weight than those from already-covered modes and determine the strength of supervision at each state (Figure 1). Teacher steps are not executed, so subsequent states remain student-generated.

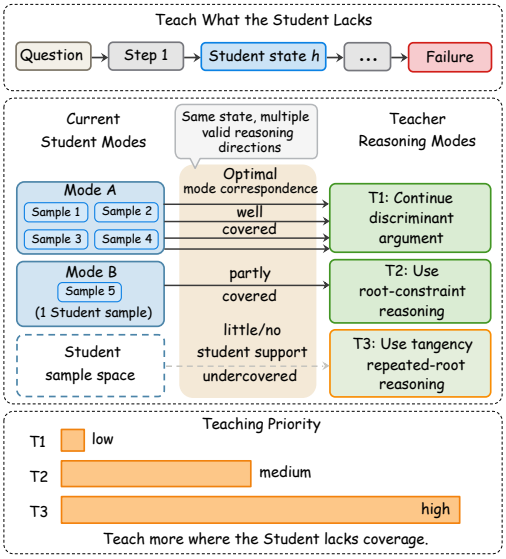

Figure 1: **Coverage-aware teaching principle.** RMT-OPD compares local teacher and student reasoning steps at a student-visited failure state. Teacher modes with weaker student support receive greater weight during training.

2.1 PROBLEM FORMULATION

Let x denote a training problem, $y^\star$ its gold answer, and π_θ the student policy. Conventional sequence-level distillation trains on teacher-generated trajectories, which may not cover the intermediate states visited by the student. OPD instead provides teacher supervision at student-visited states, but existing black-box methods typically use a single teacher target and cannot represent multiple valid local reasoning modes. At each such state h, we consider the student's local distribution $\pi_\theta^{\text{step}}(\cdot \mid h)$ and the gold-conditioned teacher distribution $q_T^{\text{step}}(\cdot \mid h, y^\star)$. Comparing these local distributions allows us to identify teacher reasoning modes that the student under-covers and give their corresponding steps greater weight during maximum-likelihood training.

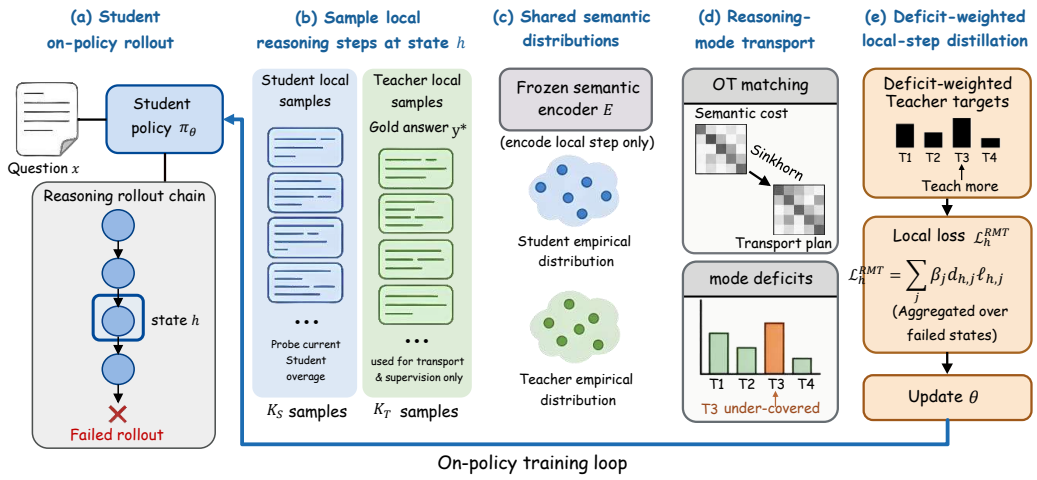

Figure 2: **Overview of the complete RMT-OPD training procedure.** At states from failed student rollouts, RMT-OPD samples local student and gold-conditioned teacher steps, aligns their empirical semantic distributions with optimal transport, and uses mode-level deficits to weight local teacher supervision. The updated student then generates fresh on-policy trajectories in the next training round.

2.2 On-Policy Local Reasoning Distributions

Failure-state collection. For each training problem x, the student samples a complete trajectory $\tau^S = (s_1^S, \ldots, s_L^S) \sim \pi_\theta(\cdot \mid x)$, where each s_l^S is a natural reasoning step. A task-specific verifier $V(x, \tau^S) \in \{0, 1\}$ identifies failed trajectories. For each failure, we collect the visited states $h_l = (x, s_{1:l}^S)$ for $l = 1, \ldots, L$ and denote the resulting set by $\mathcal{H}^-$.

Local reasoning-step sampling. For each $h \in \mathcal{H}^-$, we independently sample K_S student steps $a_{h,i}^S \sim \pi_\theta^{\text{step}}(\cdot \mid h)$ and K_T teacher steps $a_{h,j}^T \sim q_T^{\text{step}}(\cdot \mid h, y^\star)$. Each sample ends at the first natural reasoning boundary. The teacher is a gold-conditioned black-box model and may correct an erroneous prefix instead of continuing it. Only local teacher steps are used as training targets; complete continuations of teacher candidates and their final-answer verification are reserved for mechanism evaluation. Teacher steps are not executed and do not alter the original student trajectory, so all subsequent states remain fully student-generated.

Semantic distribution construction. To compare teacher and student steps without requiring shared tokenization or internal representations, we map each newly generated step a to

$$z = \frac{E(a)}{\|E(a)\|_2}, \tag{1}$$

where E is a frozen semantic encoder. We encode only the new step, excluding the problem and preceding prefix, to prevent shared context from dominating the similarity. The resulting empirical distributions are

$$\mu_h^S = \sum_{i=1}^{K_S} \alpha_i \delta_{z_{h,i}^S}, \qquad \mu_h^T = \sum_{j=1}^{K_T} \beta_j \delta_{z_{h,j}^T}, \tag{2}$$

where δ_z is a Dirac mass, $\alpha_i = 1/K_S$, and $\beta_j = 1/K_T$.

2.3 Reasoning-Mode Transport and Deficit

Reasoning-mode correspondence. For student sample i and teacher sample j at state h, we define the semantic transport cost as

$$C_{h,ij} = 1 - \cos\left(z_{h,i}^S, z_{h,j}^T\right). \tag{3}$$

Following the Sinkhorn formulation (Cuturi, 2013), we obtain their entropy-regularized optimal transport coupling by solving

$$\Pi_h^\star = \arg\min_{\Pi \in U(\alpha,\beta)} \left[\langle \Pi, C_h \rangle - \varepsilon H(\Pi)\right], \tag{4}$$

where $U(\alpha, \beta)$ is the set of couplings with marginals α and β, and $H(\Pi) = -\sum_{i,j} \Pi_{ij} \log \Pi_{ij}$.

Reasoning-mode deficit. For each teacher mode j, we define its deficit as

$$d_{h,j} = \frac{1}{\beta_j} \sum_i \Pi^\star_{h,ij} C_{h,ij}. \tag{5}$$

This quantity is the transport cost per unit of teacher mass assigned to mode j; a larger value indicates weaker coverage by the student. The total deficit at state h is

$$D_h = \sum_j \beta_j d_{h,j} = \langle \Pi^\star_h, C_h \rangle . \tag{6}$$

Thus, $\beta_j d_{h,j}$ measures the contribution of teacher mode j to the total mismatch D_h at state h.

2.4 Deficit-Weighted Training

Absolute deficit weighting. For each teacher step $a^T_{h,j}$, we define the length-normalized cross-entropy loss as

$$\ell_{h,j} = -\frac{1}{|a^T_{h,j}|} \log \pi_\theta(a^T_{h,j} \mid h). \tag{7}$$

RMT-OPD weights this loss by the absolute contribution of each teacher mode:

$$\mathcal{L}_h^{\mathrm{RMT}} = \sum_j \beta_j d_{h,j} \ell_{h,j} = D_h \, \mathbb{E}_{j \sim q_D(\cdot|h)}[\ell_{h,j}], \tag{8}$$

where $q_D(j \mid h) = \beta_j d_{h,j} / D_h$ for $D_h > 0$. Here, q_D determines the relative teaching priority within a state, while D_h controls the total supervision assigned to that state. The overall objective is

$$\mathcal{L}_{\mathrm{RMT}} = \frac{1}{|\mathcal{H}^-|} \sum_{h \in \mathcal{H}^-} \mathcal{L}_h^{\mathrm{RMT}}. \tag{9}$$

We retain the absolute weights $\beta_j d_{h,j}$ because normalizing them within each state would remove D_h and give nearly aligned and strongly mismatched states the same total weight.

Iterative on-policy updates. At each update, the current student generates fresh trajectories. RMT-OPD collects their failure states, reconstructs the local reasoning distributions and transport deficits, and updates the student using $\mathcal{L}_{\mathrm{RMT}}$. Repeating this process keeps supervision focused on the errors of the current student. Figure 2 illustrates the complete training pipeline. Algorithm 1 in Appendix B summarizes its implementation.

2.5 Theoretical Interpretation

We give two results that connect the transport deficit to the RMT-OPD training objective.

Proposition 1 (Global-to-Mode Transport Decomposition)**.** *At each state h, the total transport deficit decomposes into the weighted deficits of the teacher modes:*

$$D_h = \langle \Pi^\star_h, C_h \rangle = \sum_j \beta_j d_{h,j}. \tag{10}$$

Thus, $\beta_j d_{h,j}$ measures the contribution of teacher mode j to the total mismatch at state h. These contributions are also the weights used by the RMT-OPD training objective.

Missing teacher mass. For a more concrete interpretation, consider unregularized transport over a finite mode set $\mathcal{Z} = \{1, \ldots, M\}$. Let β_j and p_j be the teacher and student masses on mode j, respectively, with zero cost between identical modes and unit cost between different modes.

Corollary 1 (Missing Teacher Mass)**.** *In this idealized setting, the weighted deficit of teacher mode j is*

$$\beta_j d_j^{(0)} = (\beta_j - p_j)_+, \tag{11}$$

where $(x)_+ = \max(x, 0)$ and the superscript (0) denotes the unregularized limit.

When $p_j < \beta_j$, the mode weight equals the teacher probability mass missing from the student; when $p_j \geq \beta_j$, it is zero. The continuous Sinkhorn deficit used by RMT-OPD provides a geometry-aware extension of this interpretation. Proofs and additional theoretical analyses are provided in Appendix E.

3 Experiments

3.1 Experimental Setup

Tasks and model settings. Our main experiments use Qwen3-4B non-thinking as the student and Qwen3-4B-Non-Thinking-RL-Math-Step500 as the teacher in a controlled same-size setting (Team, 2025). We train on the 57K subset of DeepMath-103K (He et al., 2025) with difficulty level at least 6 and evaluate on AIME 2024, AIME 2025, HMMT February 2025, and HMMT November 2025. To evaluate applicability across model scales and reasoning domains, we pair the same student with a Qwen3-14B teacher for cross-scale mathematical reasoning and code reasoning on the 131-problem LiveCodeBench v6 benchmark (Jain et al., 2025). Following the MOPD code-training protocol (Yu et al., 2026), we use the same 131 problem prompts for on-policy training and evaluation. For each problem, 50% of the available unit tests are retained for training-time rollout verification, while evaluation uses the complete test suite. RMT-OPD accesses the teacher only through text generation.

Baselines. We compare RMT-OPD with the base student and teacher, Teacher SFT (Kim & Rush, 2016), OPD (Agarwal et al., 2024), SOPD (Sun et al., 2026), and MOPD (Yu et al., 2026). For Table 1, we use the OPD and SOPD results reported by SOPD (Sun et al., 2026), which follows the same student–teacher setting, training pool, benchmarks, and evaluation protocol.

We reproduce MOPD using the authors' public implementation and compare it with RMT-OPD over three independent training seeds under matched training and evaluation protocols. SOPD is excluded from mechanism, cross-scale, and code analyses because its training implementation and checkpoints are not publicly available.

Training configuration. Each run performs 10 on-policy updates with 1,024 prompts per update, and only failed student rollouts trigger local distribution probing. At each visited state, we independently sample five next-step candidates from the student and five from the teacher. We encode only the newly generated steps with the frozen `BAAI/bge-large-en-v1.5` encoder (Xiao et al., 2024) and compute Sinkhorn transport with cosine distance and $\varepsilon = 0.05$. Training uses the absolute transport-deficit weights defined in Section 2.4, and only teacher target tokens contribute to the supervised loss. Final-answer verification of student rollouts selects failure states for training, whereas complete continuations of teacher candidates are generated and verified only for mechanism evaluation. Neither the teacher-candidate continuations nor their verification outcomes enter the training objective.

Evaluation protocols. For mathematical reasoning, we generate 32 independent responses per problem, verify final answers with Math-Verify, and report mean accuracy on each benchmark and their unweighted average (Avg4). For code reasoning, we report Mean@8 and Pass@8 on LiveCodeBench v6. For mechanism analysis, we compare all models on fixed failure states from the initial student and against the same verified teacher candidates. We report the teacher-to-student support gap, optimal-transport gap, and distributional shape gap, with full protocol and statistical details provided in Appendix B.

Table 1: **Main results on four mathematical reasoning benchmarks.** We report mean accuracy (%) over 32 responses per problem. Bold and underlined values indicate the best and second-best results, respectively. RMT-OPD reports mean ± standard deviation over three training seeds.

Method	AIME 2024 ↑	AIME 2025 ↑	HMMT Feb. ↑	HMMT Nov. ↑	Avg4 ↑
Qwen3-4B-RL-Math Teacher	58.0	54.6	32.5	38.9	46.0
Qwen3-4B Student	21.5	21.9	10.0	8.0	15.4
OPD	61.9	57.0	32.5	39.6	47.7
MOPD	69.6	66.1	<u>40.4</u>	50.9	56.8
SOPD	<u>71.8</u>	<u>67.4</u>	38.9	<u>52.7</u>	<u>57.7</u>
RMT-OPD	$\mathbf{74.4} \pm 0.6$	$\mathbf{70.0} \pm 0.7$	$\mathbf{41.7} \pm 0.5$	$\mathbf{55.1} \pm 0.6$	$\mathbf{60.3} \pm 0.4$

3.2 Main Results

Table 1 reports the mean accuracy on four mathematical reasoning benchmarks. The base student obtains an Avg4 score of 15.4, compared with 46.0 for the teacher. Standard OPD reaches 47.7, while SOPD and MOPD reach 57.7 and 56.8, respectively. RMT-OPD achieves the highest Avg4 score of 60.3, exceeding OPD, SOPD, and MOPD by 12.6, 2.6, and 3.5 percentage points, respectively.

RMT-OPD achieves the highest mean accuracy on all four benchmarks. On AIME 2024 and AIME 2025, it reaches 74.4 and 70.0, respectively, improving over SOPD by 2.6 percentage points in both cases. On HMMT February, it achieves 41.7, compared with 40.4 for MOPD, while on HMMT November it reaches 55.1, compared with 52.7 for SOPD. Across the four benchmarks, the improvement over the strongest baseline ranges from 1.3 to 2.6 percentage points. We next use ablations to isolate the contributions of multi-target supervision and transport-based weighting.

3.3 Ablation Study

We use ablations to separate the effects of additional teacher samples, state-level weighting, and mode-level deficits. Except for Single Target, which uses one local teacher step, all variants use $K_S = K_T = 5$. Uniform assigns equal weights, Nearest selects the closest teacher target, and Centroid uses the centroid gap. Pointwise Deficit weights each teacher target by its minimum semantic distance to a student sample, $d^{\mathrm{PD}}_{h,j} = \min_i C_{h,ij}$, without computing a transport coupling. Global OT applies one state-level transport weight, whereas RMT-OPD uses mode-level deficits.

Table 2 shows that uniform multi-target supervision improves Avg4 from 57.8 to 58.6, but remains 1.7 percentage points below RMT-OPD under the same per-state sampling configuration. The compute comparison in Table 10 shows that Uniform uses slightly more teacher generation over training than RMT-OPD ($2.91\times$ versus $2.82\times$ the SOPD reference), so a larger teacher-query budget does not explain RMT-OPD's advantage. Nearest reaches 58.5, while Pointwise Deficit reaches 59.1. This suggests that emphasizing teacher targets farther from the student's sampled continuations can help even without a transport coupling. Global OT reaches 59.3 by assigning one weight per state, whereas RMT-OPD reaches 60.3 using mode-wise deficits. Together, these results suggest that the transport coupling provides information beyond additional samples, pointwise distance, or a single state-level weight.

3.4 Mechanism Analysis

We evaluate whether RMT-OPD improves the student's coverage of teacher reasoning modes. Figure 3 compares the models on the same initial-student failure states and against the same verified teacher candidates. Compared with MOPD, RMT-OPD reduces the average support, optimal-transport, and distributional shape gaps by 30.9%, 28.2%, and 34.8%, respectively. These reductions are consistent across all four mathematical benchmarks. The larger shape-gap reduction is consistent with a change in the student's reasoning-mode composition, beyond a shift in its semantic center.

Table 2: **Ablation of reasoning-mode supervision.** All multi-target variants use $K_S = K_T = 5$, while Single Target uses one local teacher step. Bold and underlined values indicate the best and second-best results.

Variant	AIME 24 ↑	AIME 25 ↑	HMMT Feb. ↑	HMMT Nov. ↑	Avg4 ↑
Single Target	71.9	67.5	39.1	52.8	57.8
Uniform	72.7	68.2	39.8	53.7	58.6
Nearest	72.9	67.9	39.6	53.5	58.5
Centroid	73.1	68.6	40.0	53.9	58.9
Pointwise Deficit	73.3	68.8	40.2	54.1	59.1
Global OT	<u>73.5</u>	<u>69.0</u>	<u>40.4</u>	<u>54.3</u>	<u>59.3</u>
RMT-OPD	**74.4**	**70.0**	**41.7**	**55.1**	**60.3**

Table 3: **Independent evaluation of reasoning-mode alignment.** BGE is used to construct the training signal; E5 and the frozen LLM judge are used only for evaluation. Brackets report 95% problem-level cluster-bootstrap confidence intervals.

Evaluation	Base Student	MOPD	RMT-OPD
BGE OT gap ↓	0.221 [0.211, 0.231]	0.171 [0.163, 0.179]	**0.123** [0.117, 0.129]
E5 OT gap ↓	0.236 [0.225, 0.247]	0.189 [0.180, 0.198]	**0.145** [0.138, 0.152]
LLM-judged coverage ↑	0.462 [0.438, 0.486]	0.579 [0.556, 0.602]	**0.692** [0.670, 0.714]

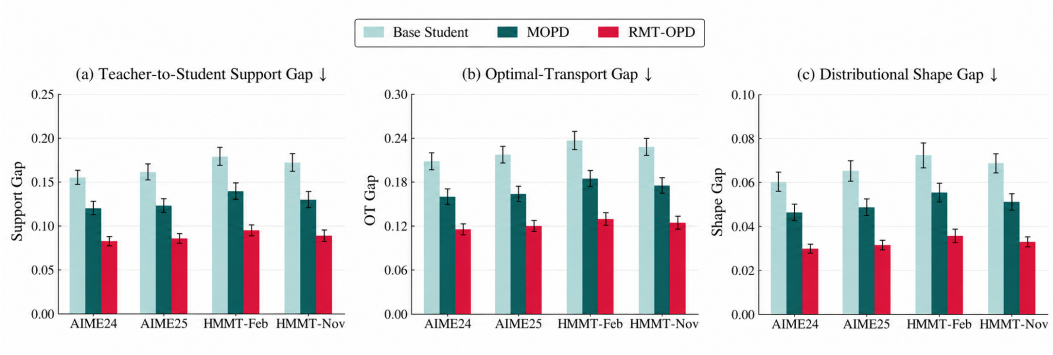

Figure 3: **Teacher–student reasoning-mode alignment.** We compare the initial student, MOPD, and RMT-OPD on the same fixed failure states collected from the initial student. Lower values indicate better alignment; error bars denote 95% confidence intervals from problem-level cluster bootstrap.

Table 3 checks whether this pattern holds beyond the BGE representation used during training. With the independent E5 encoder, RMT-OPD has a lower OT gap than MOPD; the frozen LLM judge also finds greater coverage of teacher reasoning modes. These evaluations support the coverage interpretation without relying solely on the training encoder. The evaluation protocols are detailed in Appendix C.

Teacher-target reliability. Training does not filter individual teacher steps by their final outcomes. To examine whether high-deficit targets are less reliable, we rank the 16 teacher candidates at each fixed probe state by their deficits before outcome filtering. We then extend each step to a complete continuation and check its final answer with Math-Verify. Table 4 reports the verified continuation rate for each deficit quartile. The rate is 81.8% in the highest-deficit quartile and 82.4% in the lowest-deficit quartile, with overlapping 95% confidence intervals. This audit does not establish equivalence between quartiles, but shows no clear decline in continuation reliability among high-deficit targets.

Table 4: **Teacher-continuation reliability by transport-deficit quartile.** Candidates are grouped by deficit within each fixed probe state before outcome filtering. Brackets report 95% problem-level cluster-bootstrap confidence intervals.

Deficit quartile	Verified continuation rate ↑
Q1 (lowest)	82.4% $[80.5, 84.2]$
Q2	82.9% $[81.0, 84.7]$
Q3	82.1% $[80.2, 83.9]$
Q4 (highest)	81.8% $[79.9, 83.6]$
Overall	82.3% $[81.3, 83.2]$

Table 5: **Cross-scale mathematical reasoning.** We use a Qwen3-14B teacher and a Qwen3-4B student and report mean accuracy (%) over 32 responses per problem. Teacher Access indicates the information used during distillation. Bold and underlined values indicate the best and second-best results, respectively.

Method	Teacher Access	AIME 2024 ↑	AIME 2025 ↑	HMMT Feb. ↑	HMMT Nov. ↑	Avg4 ↑
Qwen3-4B Student	–	21.5	21.9	10.0	8.0	15.4
Qwen3-14B Teacher	–	64.8	60.6	36.9	44.7	51.8
Teacher SFT	Text	58.6	54.1	32.4	40.3	46.4
OPD	Logits	64.2	59.7	35.7	44.0	50.9
MOPD	Logits + peers	<u>69.4</u>	<u>65.2</u>	<u>39.8</u>	<u>49.7</u>	<u>56.0</u>
RMT-OPD	**Text only**	**72.7**	**68.3**	**42.9**	**53.6**	**59.4**

Table 6: **Code reasoning on LiveCodeBench v6.** We evaluate a Qwen3-4B student with a Qwen3-14B teacher. Teacher Access indicates the information used during distillation. Bold and underlined values indicate the best and second-best results, respectively.

Method	Teacher Access	Mean@8 ↑	Pass@8 ↑	AST Transport Gap ↓
Qwen3-4B Student	–	28.80	49.36	0.214
Qwen3-14B Teacher	–	46.75	65.16	–
Teacher SFT	Text	46.05	56.67	0.188
OPD	Logits	60.92	67.30	0.161
MOPD	Logits + peers	<u>61.92</u>	<u>67.35</u>	<u>0.149</u>
RMT-OPD	**Text only**	**65.40**	**71.20**	**0.108**

3.5 Generalization and Efficiency

Cross-scale mathematical reasoning. We pair the Qwen3-4B student with a Qwen3-14B teacher to evaluate RMT-OPD across model scales. The methods use different forms of teacher access, as indicated in Table 5. RMT-OPD uses only generated text and achieves an Avg4 score of 59.4, compared with 56.0 for MOPD and 50.9 for OPD. It leads on all four benchmarks, exceeding MOPD by 3.3, 3.1, 3.1, and 3.9 percentage points on AIME 2024, AIME 2025, HMMT February, and HMMT November, respectively. These gains persist in the evaluated larger-teacher setting. Because teacher access and generation costs differ across methods, this table is a performance comparison rather than a compute-matched test.

Code reasoning. We evaluate whether the same RMT-OPD framework applies to code reasoning on LiveCodeBench v6. As shown in Table 6, RMT-OPD reaches 65.40 Mean@8 and 71.20 Pass@8, exceeding MOPD by 3.48 and 3.85 percentage points, respectively. Its AST transport gap is also lower, at 0.108 compared with 0.149 for MOPD. The improvements in both evaluation scores and the structural gap are consistent with the mathematical results, although the two domains use different evaluation measures.

Efficiency. The sampling audit identifies $K = 5$ as a practical operating point. Increasing K from 5 to 8 improves Avg4 by only 0.4 percentage points. Teacher generation volume is normalized to the reported SOPD step-level generation protocol, increasing from $2.82\times$ to $4.47\times$. Wall-clock cost

is normalized to RMT-OPD with $K = 1$, increasing from $1.78\times$ to $2.36\times$. Table 9 reports the full accuracy–cost comparison.

4 Related Work

Language-model and reasoning distillation. Sequence-level distillation trains students on complete teacher-generated sequences (Kim & Rush, 2016). Later language-model methods refine how students learn from teacher distributions. MiniLLM uses reverse KL (Gu et al., 2024), while DistiLLM combines skew KL with adaptive off-policy training (Ko et al., 2024). TAID gradually interpolates between student and teacher distributions during training (Shing et al., 2025). For reasoning tasks, students have been trained on teacher-generated rationales (Ho et al., 2023; Hsieh et al., 2023) and structured reasoning signals (Wen et al., 2025). Self-Consistency illustrates the diversity of reasoning paths at inference time by sampling multiple chains and aggregating their answers (Wang et al., 2023). Studies of personalized explanations and reasoning trajectories (Saha et al., 2023; Chen et al., 2025; Yang et al., 2026) further examine how the value of teacher supervision depends on the characteristics of the student being trained and on the particular trajectories selected for training.

On-policy reasoning distillation. On-policy distillation moves teacher supervision to student-generated trajectories. GKD uses teacher token distributions at states visited by the student (Agarwal et al., 2024). Fu et al. identify unreliable teacher guidance on some student-generated prefixes as a failure mode of sampled-token OPD (Fu et al., 2026). More recent methods change how feedback is provided at these states. SOPD obtains a teacher-generated reasoning step at each student step using only textual teacher responses (Sun et al., 2026). MOPD uses successful and failed student rollouts to provide the teacher with richer context for its feedback (Yu et al., 2026). These methods vary in the granularity and context of teacher feedback, but none explicitly compares the teacher's and student's local reasoning-mode coverage.

Optimal-transport-based knowledge distillation. Optimal transport has also been used to compare teacher and student distributions. Lohit and Jones align feature distributions (Lohit & Jones, 2022), while KNOT matches student label predictions to a weighted combination of teacher predictions (Bhardwaj et al., 2022). SinKD applies the Sinkhorn distance to predictive distributions (Cui et al., 2025a). MultiLevelOT aligns logit distributions at the token and sequence levels across different tokenizers (Cui et al., 2025b), and SelKD uses inverse optimal transport for selective transfer of teacher categories (Shi et al., 2025). These methods operate on features, logits, or predicted probabilities. Generated reasoning steps pose a different comparison problem because the same reasoning mode can be expressed in different words. After mapping steps into a shared semantic space, optimal transport can align their empirical distributions without pairing individual samples in advance. RMT-OPD uses this alignment to compare teacher- and student-generated steps at the same student-visited state, then prioritizes teacher modes that the student under-covers.

5 Conclusion

In this paper, we study black-box on-policy distillation from the perspective of reasoning-mode coverage. We propose RMT-OPD, which constructs local teacher and student reasoning distributions from text samples at student-visited failure states and uses optimal transport to identify under-covered teacher modes. Our analysis decomposes the state-level transport discrepancy into mode-level deficits. It relates each weighted deficit to the teacher probability mass missing from the student under an idealized discrete-mode model. RMT-OPD achieves the best performance among the compared methods on all four mathematical reasoning benchmarks. The ablation and fixed-state analyses support the role of deficit-weighted mode supervision beyond the benefit of drawing multiple teacher samples, while the cross-scale and code experiments show that the gains extend to a larger teacher and a different reasoning domain. These findings suggest that black-box on-policy distillation can move beyond single local corrections toward distribution-level supervision that targets the reasoning modes the student lacks.

AI USE STATEMENT

Generative AI tools assisted with drafting and revising portions of the manuscript, organizing the technical exposition, and checking notation and LaTeX presentation. They also helped refine explanations of the method and theoretical analysis and summarize experimental results supplied by the authors. The authors reviewed the AI-assisted text and take responsibility for the final manuscript, including its citations, mathematical statements, figures, and reported values. Generative AI was not used to generate data, train models, run experiments, or evaluate the reported results.

ETHICS STATEMENT

This work studies reasoning distillation on mathematical and programming benchmarks and does not involve human participants. RMT-OPD uses final-answer verification to identify failed trajectories, but this signal does not certify the validity of every intermediate reasoning step. Likewise, improved benchmark accuracy does not establish that the trained student is reliable or safe in other domains, particularly high-stakes applications. We do not evaluate broader safety properties or potential misuse; deployment in such settings would require separate assessment and appropriate safeguards.

REPRODUCIBILITY STATEMENT

Section 2 specifies the method and training objective. Section 3.1 reports the models, datasets, baselines, and evaluation protocols. Appendix B provides implementation and evaluation details, while Appendix E contains the assumptions and proofs.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Rishabh Bhardwaj, Tushar Vaidya, and Soujanya Poria. KNOT: knowledge distillation using optimal transport for solving NLP tasks. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pp. 4801–4820. International Committee on Computational Linguistics, 2022. URL https://aclanthology.org/2022.coling-1.425.

Xinghao Chen, Zhijing Sun, Wenjin Guo, Miaoran Zhang, Yanjun Chen, Yirong Sun, Hui Su, Yijie Pan, Dietrich Klakow, Wenjie Li, and Xiaoyu Shen. Unveiling the key factors for distilling chain-of-thought reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pp. 15094–15119. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-ACL.782. URL https://doi.org/10.18653/v1/2025.findings-acl.782.

Xiao Cui, Yulei Qin, Yuting Gao, Enwei Zhang, Zihan Xu, Tong Wu, Ke Li, Xing Sun, Wengang Zhou, and Houqiang Li. Sinkd: Sinkhorn distance minimization for knowledge distillation. *IEEE Trans. Neural Networks Learn. Syst.*, 36(7):11887–11901, 2025a. doi: 10.1109/TNNLS.2024.3501335. URL https://doi.org/10.1109/TNNLS.2024.3501335.

Xiao Cui, Mo Zhu, Yulei Qin, Liang Xie, Wengang Zhou, and Houqiang Li. Multi-level optimal transport for universal cross-tokenizer knowledge distillation on language models. In Toby

Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 23724–23732. AAAI Press, 2025b. doi: 10.1609/AAAI.V39I22.34543. URL https://doi.org/10.1609/aaai.v39i22.34543.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2292–2300, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=mZn2Xyh9Ec.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *CoRR*, abs/2603.25562, 2026. doi: 10.48550/ARXIV.2603.25562. URL https://doi.org/10.48550/arXiv.2603.25562.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *CoRR*, abs/2504.11456, 2025. doi: 10.48550/ARXIV.2504.11456. URL https://doi.org/10.48550/arXiv.2504.11456.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 14852–14882. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.830. URL https://doi.org/10.18653/v1/2023.acl-long.830.

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 L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, volume ACL 2023 of *Findings of ACL*, pp. 8003–8017. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.507. URL https://doi.org/10.18653/v1/2023.findings-acl.507.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Xavier Carreras, and Kevin Duh (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 1317–1327. The Association for Computational Linguistics, 2016. doi: 10.18653/V1/D16-1139. URL https://doi.org/10.18653/v1/d16-1139.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace (eds.), *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Suhas Lohit and Michael Jones. Model compression using optimal transport. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pp. 3645–3654. IEEE, 2022. doi: 10.1109/WACV51458.2022.00370. URL https://doi.org/10.1109/WACV51458.2022.00370.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. Can language models teach? teacher explanations improve student performance via personalization. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/c6afe9a5d1e1068796d32613ddca1ab7-Abstract-Conference.html.

Liangliang Shi, Zhengyan Shi, and Junchi Yan. Selkd: Selective knowledge distillation via optimal transport perspective. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=H4iVLvRusn.

Makoto Shing, Kou Misaki, Han Bao, Sho Yokoi, and Takuya Akiba. TAID: temporally adaptive interpolated distillation for efficient knowledge transfer in language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=cqsw28DuMW.

Changhui Sun, Lanbo Liu, Hang Lei, Tong Ling, Jiahang Xie, Zhiyong Zheng, Yujia Wang, Hao Liu, Feng Xiao, Lu Liu, Yanlong Du, Zifeng Cheng, Ziwei Jiang, and Qing Gu. Step-level on-policy distillation: Interpolating between on-policy distillation and supervised fine-tuning. *CoRR*, abs/2608.16333, 2026. doi: 10.48550/ARXIV.2608.16333. URL https://doi.org/10.48550/arXiv.2608.16333.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533, 2022. doi: 10.48550/ARXIV.2212.03533. URL https://doi.org/10.48550/arXiv.2212.03533.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Xiangyu Wen, Junhua Huang, Zeju Li, Min Li, Jianyuan Zhong, Zhijian Xu, Mingxuan Yuan, Yongxiang Huang, and Qiang Xu. Reasoning scaffolding: Distilling the flow of thought from

llms. *CoRR*, abs/2509.23619, 2025. doi: 10.48550/ARXIV.2509.23619. URL https://doi.org/10.48550/arXiv.2509.23619.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (eds.), *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pp. 641–649. ACM, 2024. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878.

Yuming Yang, Mingyoung Lai, Wanxu Zhao, Xiaoran Fan, Zhiheng Xi, Mingqi Wu, Chiyue Huang, Jun Zhao, Haijun Lv, Jian Tong, Yunhua Zhou, Yicheng Zou, Qipeng Guo, Tao Gui, Qi Zhang, and Xuanjing Huang. Which reasoning trajectories teach students to reason better? A simple metric of informative alignment. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 42123–42150. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.1950. URL https://doi.org/10.18653/v1/2026.acl-long.1950.

Weichen Yu, Xiaomin Li, Yizhou Zhao, Xiaoze Liu, Ruowang Zhang, Haixin Wang, Yinyi Luo, Chen Henry Wu, Gaurav Mittal, Matt Fredrikson, and Yu Hu. Multi-rollout on-policy distillation via peer successes and failures. *CoRR*, abs/2605.12652, 2026. doi: 10.48550/ARXIV.2605.12652. URL https://doi.org/10.48550/arXiv.2605.12652.

A Notation

Table 7 summarizes the principal notation used in RMT-OPD.

B Detailed Experimental Setup

This section describes the tasks, model configurations, training procedure, and evaluation protocols. The main experimental configuration is summarized in Table 8.

B.1 Tasks, Models, and Baselines

Tasks and datasets. For mathematical reasoning, we use the approximately 57K-problem subset of DeepMath-103K with difficulty level at least 6. We evaluate on AIME 2024, AIME 2025, HMMT February 2025, and HMMT November 2025. Final answers are checked using Math-Verify.

For code reasoning, we follow the LiveCodeBench v6 protocol used by MOPD (Yu et al., 2026). The benchmark contains 131 competition-level programming problems with compiler feedback and unit tests. We use the same problem prompts during on-policy training and evaluation. For each problem, 50% of the available unit tests are retained for training-time rollout verification, while evaluation uses the complete test suite. We report Mean@8 and Pass@8.

Teacher–student settings. The controlled mathematical setting uses Qwen3-4B non-thinking as the student and Qwen3-4B-Non-Thinking-RL-Math-Step500 as the teacher. The cross-scale mathematical and code experiments use a Qwen3-14B teacher and a Qwen3-4B student.

RMT-OPD accesses the teacher through text generation. It does not use teacher logits, hidden states, gradients, or internal attention representations.

Baselines. We compare RMT-OPD with the initial student, the teacher, Teacher SFT, standard OPD, SOPD, and MOPD. The Uniform, Nearest, Centroid, and Global OT variants are used only in the ablation study.

For the controlled mathematical comparison, we use the OPD and SOPD results reported in the SOPD evaluation. This evaluation uses the same student–teacher pair, training pool, mathematical

Table 7: Summary of the principal notation used in RMT-OPD.

Symbol	Meaning
Problem and policy	
x, $y^\star$, V	Training problem, gold answer, and binary final-answer verifier.
π_θ	Student policy parameterized by θ.
$q_T^{\mathrm{step}}(\cdot \mid h, y^\star)$	Gold-conditioned teacher distribution over local reasoning steps at state h.
τ^S, s_l^S	Complete student trajectory and its l-th natural reasoning step.
h_l, $\mathcal{H}^-$	Student-visited state $h_l = (x, s_{1:l}^S)$ and the set of states collected from failed trajectories.
Local reasoning distributions	
K_S, K_T	Numbers of local reasoning steps sampled from the student and teacher, respectively.
i, j	Indices of student and teacher samples, respectively.
$a_{h,i}^S$, $a_{h,j}^T$	Student and teacher local reasoning steps sampled at state h.
E, $z_{h,i}^S$, $z_{h,j}^T$	Frozen semantic encoder and the normalized semantic representations of student and teacher steps.
μ_h^S, μ_h^T	Empirical student and teacher reasoning distributions at state h.
α_i, β_j	Probability masses assigned to student sample i and teacher sample j.
Reasoning-mode transport	
$C_{h,ij}$, C_h	Semantic transport cost between student sample i and teacher sample j, and the corresponding cost matrix.
$U(\alpha, \beta)$	Set of transport couplings with student and teacher marginals α and β.
$\Pi_h^\star$	Entropy-regularized optimal transport coupling at state h.
ε, $H(\Pi)$	Entropic regularization coefficient and entropy of transport coupling Π.
$d_{h,j}$	Transport deficit per unit mass of teacher mode j at state h.
D_h	Total state-level transport deficit, given by $D_h = \sum_j \beta_j d_{h,j}$.
Deficit-weighted training	
$\ell_{h,j}$	Length-normalized cross-entropy loss for teacher step $a_{h,j}^T$.
$q_D(j \mid h)$	Normalized deficit distribution that determines relative teaching priority within state h.
$\mathcal{L}_h^{\mathrm{RMT}}$	Deficit-weighted training loss at state h.
$\mathcal{L}_{\mathrm{RMT}}$	Overall RMT-OPD objective averaged over collected failure states.
Mechanism-evaluation metrics	
$\bar{z}_h^S$, $\bar{z}_h^T$	Semantic centroids of the student and teacher empirical distributions.
$G_{\mathrm{support}}(h)$	Teacher-to-student support gap at state h.
$G_{\mathrm{OT}}(h)$	empirical local reasoning distributions.
$G_{\mathrm{shape}}(h)$	Distributional shape gap after removing the contribution of centroid displacement.
$G_{\mathrm{judge}}(h)$	Teacher-mode coverage measured by the frozen LLM strategy judge.

benchmarks, and evaluation protocol. Analyses that require trained checkpoints or newly generated outputs do not include SOPD.

We reproduce MOPD using the authors' public implementation. Its default configuration samples eight student rollouts per prompt and constructs the peer-conditioned teacher context using two successful trajectories and one failed trajectory.

For LiveCodeBench, the Teacher SFT row corresponds to the GKD (TS) baseline reported by MOPD (Yu et al., 2026), which trains the student by maximum likelihood on teacher-generated completions.

B.2 Training and Implementation

On-policy rollout and state collection. At each update, the current student generates one complete reasoning trajectory $\tau^S \sim \pi_\theta(\cdot \mid x)$ for every training prompt x. We partition the trajectory as $\tau^S = (s_1^S, \ldots, s_L^S)$ using natural reasoning boundaries. Explicit step markers, solution markers,

and paragraph boundaries may terminate a step. A 1,024-token safety cap is used when no natural boundary is generated.

A task-specific verifier evaluates the final answer. Local reasoning-distribution sampling is performed when $V(x, \tau^S) = 0$. For each failed trajectory, we retain all student-visited natural reasoning states $h_l = (x, s^S_{1:l})$ for $l = 1, \ldots, L$.

Local reasoning-distribution sampling. At each retained state h, we independently sample K_S local student steps and K_T local teacher steps. Student candidates follow $a^S_{h,i} \sim \pi^{\text{step}}_\theta(\cdot \mid h)$, and teacher candidates follow $a^T_{h,j} \sim q^{\text{step}}_T(\cdot \mid h, y^\star)$.

The teacher observes the gold answer and may correct an error in the current student prefix. Each candidate ends at the first natural reasoning boundary. Teacher steps are not executed, so later states continue to come from the original student trajectory. Complete teacher-candidate continuations and their final-answer verification are used only for mechanism evaluation.

Semantic representation and optimal transport. We encode only the newly generated reasoning step using the frozen semantic encoder `BAAI/bge-large-en-v1.5`. Each representation is normalized as $z = E(a)/\|E(a)\|_2$.

The student and teacher empirical distributions use uniform masses, with $\alpha_i = 1/K_S$ and $\beta_j = 1/K_T$. The semantic transport cost is $C_{h,ij} = 1 - \cos(z^S_{h,i}, z^T_{h,j})$. We compute the transport plan using entropy-regularized Sinkhorn optimal transport. The teacher-mode deficit is $d_{h,j} = \beta_j^{-1} \sum_i \Pi^\star_{h,ij} C_{h,ij}$.

Pointwise-deficit ablation. To test whether pointwise semantic distance accounts for the gains without distribution-level transport, we define $d^{\text{PD}}_{h,j} = \min_i C_{h,ij}$ and optimize $\mathcal{L}^{\text{PD}}_h = \sum_{j=1}^{K_T} \beta_j d^{\text{PD}}_{h,j} \ell_{h,j}$. Pointwise Deficit uses the same encoder and training and evaluation settings as RMT-OPD, with $K_S = K_T = 5$ samples per state. It replaces the transport-derived deficit with each teacher target's minimum distance to a student sample, without computing a transport coupling or introducing an additional hyperparameter.

Student optimization. RMT-OPD uses standard teacher-forced maximum-likelihood learning:

$$\mathcal{L}_{\text{RMT}} = \frac{1}{|\mathcal{H}^-|} \sum_{h \in \mathcal{H}^-} \sum_{j=1}^{K_T} \beta_j d_{h,j} \ell_{h,j}. \tag{12}$$

Each training run performs 10 on-policy updates with 1,024 prompts per update. Local correction targets are processed using packed batching. Only teacher target tokens contribute to the supervised loss; student trajectory tokens are used as context.

Implementation and compute. RMT-OPD is implemented in PyTorch. Student generation and black-box teacher sampling are executed asynchronously using vLLM (Kwon et al., 2023). Student optimization uses fully sharded data parallelism and FlashAttention-2 (Dao, 2024). Model weights and training use bfloat16 precision.

The controlled mathematical experiments are conducted on one node with eight NVIDIA H100 80GB GPUs. The cross-scale and code experiments use the same software stack, with teacher generation and student optimization assigned to separate GPU worker groups.

B.3 Evaluation Protocols

Mathematical evaluation. For each mathematical benchmark, we independently generate 32 responses per problem using temperature 1.0, top-$p = 1.0$, and a maximum response length of 32,768 tokens. For problem i and response j, the evaluation seed is $42 + 32i + j$.

Math-Verify checks symbolic equivalence between the generated and reference answers. We report mean accuracy on each benchmark and their unweighted average, Avg4.

Table 8: Summary of the main training and evaluation configuration.

Setting	Configuration	Setting	Configuration
Math training data	DeepMath difficulty ≥ 6, approximately 57K	Math benchmarks	AIME24/25, HMMT25-Feb/Nov
Student	Qwen3-4B non-thinking	Controlled teacher	Qwen3-4B-Non-Thinking-RL-Math-Step500
Cross-scale teacher	Qwen3-14B	Code benchmark	LiveCodeBench v6
Training updates	10	Prompts per update	1,024
Main rollout	$T = 1.0$, top-$p = 1.0$	Maximum response length	32,768
Local samples	$K_S = K_T = 5$	Local sampling	$T_{\text{local}} = 0.7$, top-$p = 1.0$
Maximum local-step length	1,024	Semantic encoder	`BAAI/bge-large-en-v1.5`
Embedding dimension	1,024	Transport cost	$1 - \cos(\cdot, \cdot)$
Sinkhorn regularization	$\varepsilon = 0.05$	Sinkhorn stopping	100 iterations, tolerance 10^{-6}
Optimizer	AdamW, learning rate 2×10^{-6}	Precision and clipping	bf16, $\|\nabla\|_2 \leq 1.0$
Math evaluation	32 responses per problem	Probe evaluation	$K_S^{\text{eval}} = K_T^{\text{eval}} = 16$
Training seeds	3	Hardware	8× NVIDIA H100 80GB
Generation backend	vLLM	Training backend	FSDP and FlashAttention-2

Algorithm 1 Reasoning-Mode Transport On-Policy Distillation (RMT-OPD)

Require: Training set $\mathcal{D} = \{(x, y^\star)\}$; student π_{θ_0}; gold-conditioned teacher q_T^{step}; verifier V; frozen encoder E; sample counts K_S, K_T; OT coefficient ε; number of updates N_{upd}
Ensure: Updated student $\pi_{\theta_{N_{\text{upd}}}}$

for $u = 0, \ldots, N_{\text{upd}} - 1$ **do**
  Sample a prompt batch $\mathcal{B}_u \subseteq \mathcal{D}$
  Initialize $\mathcal{H}_u^- \leftarrow \emptyset$
  for all $(x, y^\star) \in \mathcal{B}_u$ **do**
    Sample $\tau^S \sim \pi_{\theta_u}(\cdot \mid x)$
    if $V(x, \tau^S) = 0$ **then**
      Add its visited states $h_l = (x, s_{1:l}^S)$ and associated answer $y_h^\star$ to $\mathcal{H}_u^-$
    end if
  end for
  for all $h \in \mathcal{H}_u^-$ **do**
    Sample $\{a_{h,i}^S\}_{i=1}^{K_S}$ and $\{a_{h,j}^T\}_{j=1}^{K_T}$ from the student and gold-conditioned teacher
    Encode only the new steps with E to form μ_h^S and μ_h^T
    Compute C_h, $\Pi_h^\star$, $d_{h,j}$, and D_h as defined above
    Compute $\ell_{h,j}$ and $\mathcal{L}_h^{\text{RMT}}$ as defined above
  end for
  Compute $\mathcal{L}_{\text{RMT}} = |\mathcal{H}_u^-|^{-1} \sum_{h \in \mathcal{H}_u^-} \mathcal{L}_h^{\text{RMT}}$
  Update θ_u using $\mathcal{L}_{\text{RMT}}$ to obtain θ_{u+1}
end for

MOPD and RMT-OPD are trained using three independent training seeds, and the main results report the seed-level mean. SOPD is included through its published values.

Fixed probe-state evaluation. For each mathematical benchmark, the frozen initial Qwen3-4B student generates eight trajectories per problem. We retain trajectories with incorrect final answers and select one non-terminal natural reasoning state from each failed trajectory. Since each benchmark contains 30 problems, this procedure produces at most 240 probe states per benchmark.

The probe states remain fixed across all evaluated models and checkpoints. At each state, we sample $K_S^{\text{eval}} = 16$ student candidates and $K_T^{\text{eval}} = 16$ teacher candidates. The teacher candidates are fixed and shared across the evaluated student models.

Each teacher candidate is extended into a complete continuation and checked using Math-Verify. The mechanism analysis reports the teacher-to-student support gap, optimal-transport gap, and distributional shape gap. Confidence intervals are computed using problem-level cluster bootstrap.

Cross-scale and code evaluation. The cross-scale mathematical evaluation uses a Qwen3-14B teacher and a Qwen3-4B student. The teacher-access level of each method is reported in the corresponding table.

For LiveCodeBench v6, training-time rollout verification uses the reduced unit-test sets, whereas the reported evaluation uses the complete test suite. We generate eight programs per problem. Mean@8 is the average fraction of programs that pass all unit tests. Pass@8 is the fraction of problems for which at least one generated program passes all tests.

We additionally report the AST transport gap. For a program p, we parse its abstract syntax tree and construct an AST-node-type count vector $a(p)$. The structural cost between programs p_i and p_j is

$$C_{\mathrm{AST}}(p_i, p_j) = 1 - \frac{\sum_t \min(a_{i,t}, a_{j,t})}{\sum_t \max(a_{i,t}, a_{j,t})}. \tag{13}$$

This structural cost is used as the teacher–student optimal-transport cost matrix. Programs that cannot be parsed are assigned the maximum structural cost.

C Additional Mechanism Analysis

We evaluate whether RMT-OPD improves the student's coverage of teacher reasoning modes. All comparisons use the fixed probe states and shared verified teacher candidates described in Appendix B.

C.1 Reasoning-Mode Coverage Metrics

For a fixed reasoning state h, let the student and teacher empirical distributions be

$$\mu_h^S = \sum_i \alpha_i \delta_{z_{h,i}^S}, \qquad \mu_h^T = \sum_j \beta_j \delta_{z_{h,j}^T}, \tag{14}$$

with semantic cost

$$C_{h,ij} = 1 - \cos(z_{h,i}^S, z_{h,j}^T). \tag{15}$$

Teacher-to-student support gap. We define

$$G_{\mathrm{support}}(h) = \sum_j \beta_j \min_i C_{h,ij}. \tag{16}$$

For each teacher mode, this metric identifies the closest student candidate and averages the resulting cost according to the teacher mass. A lower support gap indicates that more teacher modes have similar candidates in the student support. This metric does not account for the amount of student probability mass assigned to each mode.

Optimal-transport gap. We compute

$$G_{\mathrm{OT}}(h) = \langle \Pi_h^\star, C_h \rangle, \tag{17}$$

where $\Pi_h^\star$ is the Sinkhorn transport plan between the student and teacher empirical distributions. Unlike the support gap, G_{OT} accounts for both support correspondence and probability-mass allocation. A lower OT gap indicates empirical local reasoning distributions.

Distributional shape gap. Let

$$\bar{z}_h^S = \sum_i \alpha_i z_{h,i}^S, \qquad \bar{z}_h^T = \sum_j \beta_j z_{h,j}^T \tag{18}$$

be the student and teacher semantic centroids. We define

$$G_{\mathrm{shape}}(h) = G_{\mathrm{OT}}(h) - \frac{1}{2} \left\| \bar{z}_h^S - \bar{z}_h^T \right\|_2^2. \tag{19}$$

The second term measures the discrepancy explained by the displacement between the two centroids. The remaining shape gap measures distributional mismatch beyond this centroid displacement. Its lower-bound interpretation is provided in Appendix E.

C.2 Teacher–Student Reasoning-Mode Alignment

Figure 3 compares the three models on the same probe states and against the same verified teacher candidates. The average support gap decreases from 0.166 for the Base Student to 0.129 for MOPD and 0.089 for RMT-OPD. Relative to MOPD, RMT-OPD reduces the support gap by 30.9%.

The average OT gap decreases from 0.221 for the Base Student to 0.171 for MOPD and 0.123 for RMT-OPD, corresponding to a 28.2% reduction relative to MOPD. The average shape gap decreases from 0.067 to 0.050 and 0.033, respectively, with RMT-OPD reducing the gap by 34.8% relative to MOPD. The consistent reductions across the three metrics indicate improved teacher-mode coverage and probability-mass alignment. The larger reduction in the shape gap is consistent with a change in the composition of the student reasoning distribution beyond a shift in its semantic centroid.

C.3 Independent Mechanism Validation

We evaluate the same fixed probe states and candidate sets using an independent semantic encoder and a frozen LLM strategy judge. Table 3 reports the results in the main text; the evaluation protocols are described below.

Independent semantic encoder. We use `intfloat/e5-large-v2` (Wang et al., 2022) to re-encode the fixed student and teacher candidates. E5 is not used for training, candidate selection, or teacher weighting. We recompute the cosine cost and OT gap while keeping the probe states and candidate sets unchanged. Under this encoder, the average OT gap is 0.236 for the Base Student, 0.189 for MOPD, and 0.145 for RMT-OPD.

Frozen LLM strategy judge. We use Llama-3.3-70B-Instruct (Team, 2024) with temperature 0. For each teacher reasoning action T_j, the judge receives the problem, the current state, the teacher action, and the set of student candidates. It determines whether at least one student candidate follows the same high-level reasoning strategy as the teacher action. The output is restricted to `MATCH` or `NO_MATCH`.

The student candidates are randomly permuted three times, and the final decision is determined by majority vote. We define the judge-based coverage as

$$G_{\text{judge}}(h) = \sum_j \beta_j \mathbf{1}\left[T_j \text{ is matched by at least one student candidate}\right]. \tag{20}$$

RMT-OPD achieves a coverage score of 0.692, compared with 0.579 for MOPD and 0.462 for the Base Student.

The method ordering is consistent across the training encoder, the independent encoder, and the frozen strategy judge. These results provide independent evidence that the improvement in reasoning-mode alignment is not limited to the semantic representation used during RMT-OPD training.

D Additional Experimental Results

This section examines the effects of local sampling scale, computational budget, and training progress. We first study the trade-off between reasoning-distribution estimation and generation cost.

D.1 Effect of Reasoning-Distribution Sample Size and Computational Cost

We vary the local sample size while keeping $K_S = K_T = K$. Table 9 reports the resulting mathematical reasoning performance, distributional gaps, and generation costs.

Increasing K consistently improves both reasoning accuracy and teacher–student distribution alignment. With $K = 1$, the local distributions reduce to a single student–teacher sample pair, resulting in an average accuracy of 57.9, a support gap of 0.141, and an OT gap of 0.178. Increasing K to 3 raises the accuracy to 59.4 and reduces the two gaps to 0.106 and 0.143. At $K = 5$, the accuracy reaches 60.3, while the support and OT gaps decrease further to 0.089 and 0.123.

Table 9: Effect of local reasoning-distribution sample size on performance, alignment, and computational cost. Bold denotes the main configuration.

Method	Math Avg. ↑	Support Gap ↓	OT Gap ↓	Triggered Prompts ↓	Teacher Samples per State	Teacher Tokens per Triggered State	Teacher Gen. Volume† ↓	Total Generation‡ ↓	Wall Time§ ↓
SOPD (reference)	57.7	–	–	100%	1	$29.0 \rightarrow 45.2$	$1.00\times$	$1.00\times$	–
RMT-OPD, $K = 1$	57.9	0.141	0.178	59.5%	1	36.7	$0.60\times$	$1.10\times$	$1.00\times$
RMT-OPD, $K = 3$	59.4	0.106	0.143	57.5%	3	109.8	$1.73\times$	$2.23\times$	$1.43\times$
RMT-OPD, $K = 5$	**60.3**	**0.089**	**0.123**	**56.4%**	**5**	**182.3**	$2.82\times$	$3.32\times$	$1.78\times$
RMT-OPD, $K = 8$	60.7	0.083	0.117	55.9%	8	290.4	$4.47\times$	$4.97\times$	$2.36\times$

† Teacher generation volume is normalized by the SOPD reference. ‡ Total generation includes the main student rollout, local student sampling, and teacher sampling. § Wall time is normalized within the RMT-OPD configurations, with $K = 1$ set to $1.00\times$.

Table 10: Compute-matched comparison at $K_S = K_T = 5$. Uniform-$K5$ and RMT-OPD use the same per-state sampling and optimization configurations.

Method	Math Avg. ↑	Triggered Prompts	Student Samples per State	Teacher Samples per State	Teacher Tokens per Triggered State	Teacher Gen. Volume ↓	Total Generation ↓	Wall Time ↓
SOPD (reference)	57.7	100%	–	1	$29.0 \rightarrow 45.2$	$1.00\times$	$1.00\times$	–
Budget-Matched Uniform-$K5$	58.6	58.3%	5	5	181.7	$2.91\times$	$3.43\times$	$1.81\times$
RMT-OPD, $K = 5$	**60.3**	56.4%	5	5	182.3	$2.82\times$	$3.32\times$	$1.78\times$

Increasing K from 5 to 8 improves the average accuracy by only 0.4 percentage points and reduces the support and OT gaps by 0.006 each. The evaluated performance thus exhibits diminishing returns beyond five local samples.

The additional samples increase generation cost. At $K = 5$, RMT-OPD uses $2.82\times$ the teacher generation volume and $3.32\times$ the total generation volume of the SOPD reference. Its wall time is $1.78\times$ that of the $K = 1$ configuration. Increasing K to 8 raises these values to $4.47\times$, $4.97\times$, and $2.36\times$, respectively.

Compared with $K = 1$, the $K = 5$ configuration reduces the support gap by 36.9% and the OT gap by 30.9%. Compared with $K = 8$, it retains nearly all of the accuracy gain while using approximately 37% less teacher generation and 25% less wall time. We therefore use $K = 5$ as the main configuration in the remaining experiments.

D.2 Compute-Matched Comparison

To separate the effect of transport-based weighting from that of additional sampling, we construct a compute-matched baseline called Budget-Matched Uniform-$K5$. It uses the same five student samples and five teacher samples per triggered state as RMT-OPD, together with the same step-length limit, optimization schedule, number of updates, and batching protocol. The only difference is that Uniform-$K5$ assigns equal weight $w_j^{\mathrm{uni}} = 1/K_T$ to all teacher targets instead of using their reasoning-mode deficits.

Table 10 shows that equal weighting does not recover the performance of RMT-OPD. Budget-Matched Uniform-$K5$ achieves an average accuracy of 58.6, whereas RMT-OPD reaches 60.3, a difference of 1.7 percentage points.

The two methods use the same sampling budget at every triggered state. Uniform-$K5$ has a slightly higher trigger rate because its student solves fewer prompts successfully. As a result, its teacher generation volume, total generation volume, and wall time are slightly higher than those of RMT-OPD. Therefore, the advantage of RMT-OPD cannot be attributed to a larger teacher-query budget. Under the same per-state sampling configuration, mode-deficit weighting provides more effective supervision than assigning equal weight to all teacher targets.

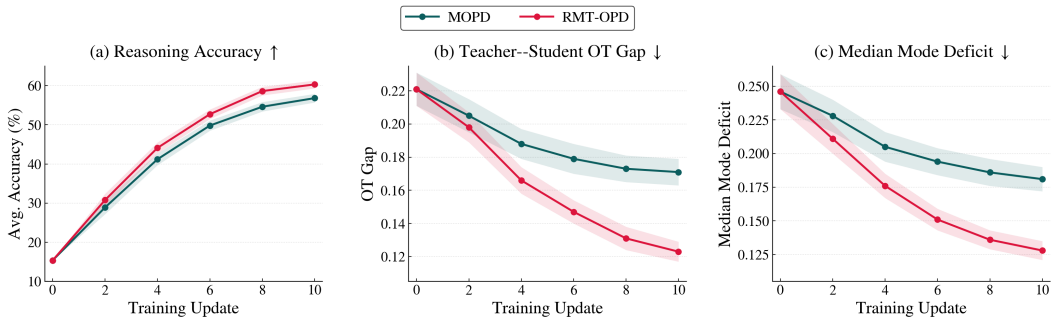

Figure 4: Training dynamics of MOPD and RMT-OPD. We report (a) average accuracy over the four mathematical reasoning benchmarks, (b) the average teacher–student OT gap, and (c) the median reasoning-mode deficit. Structural metrics are computed on the same fixed probe states and teacher candidates at every checkpoint. Shaded regions denote 95% problem-level cluster-bootstrap confidence intervals.

D.3 Training Dynamics

We track reasoning performance and distributional alignment at updates 0, 2, 4, 6, 8, and 10. MOPD and RMT-OPD start from the same student checkpoint. All structural metrics are evaluated using the same fixed probe states, teacher candidate sets, semantic encoder, and transport configuration.

Figure 4(a) shows that both methods improve reasoning accuracy throughout training, but RMT-OPD establishes an advantage from the early checkpoints. At update 2, its average accuracy is 30.8, compared with 28.9 for MOPD. The difference continues to increase, and the final accuracies reach 60.3 and 56.8, respectively.

Figure 4(b) shows the corresponding change in teacher–student distribution alignment. Both methods begin with an average OT gap of 0.221. MOPD reduces this value to 0.171, whereas RMT-OPD reduces it to 0.123. This corresponds to a 44.3% reduction from the initial checkpoint for RMT-OPD. At the final checkpoint, its OT gap is 28.2% lower than that of MOPD.

Figure 4(c) reports the median mode deficit over all teacher modes in the fixed probe states. RMT-OPD reduces this value from 0.246 to 0.128, a reduction of 48.0%. MOPD reduces it to 0.181, corresponding to a reduction of 26.4%. At the final checkpoint, the median deficit of RMT-OPD is 29.3% lower than that of MOPD.

Together, the three panels show a consistent training pattern: reasoning accuracy increases as the teacher–student OT gap and reasoning-mode deficit decrease. Compared with MOPD, RMT-OPD exhibits faster distributional alignment, a larger reduction in under-covered teacher modes, and higher final reasoning accuracy. These synchronized trends are consistent with the intended effect of reasoning-mode deficit supervision.

E Proofs and Additional Theoretical Analysis

This section provides additional theoretical results for the semantic transport formulation, verified teacher distributions, Sinkhorn approximation, deficit-weighted optimization, and iterative on-policy data collection.

E.1 Local Semantic Fidelity and Transport Consistency

The transport formulation assumes that the frozen semantic encoder preserves the local relationships between reasoning actions. We state this condition formally and relate the resulting embedding-space transport objective to the discrepancy in the original reasoning-action space.

For a fixed reasoning state h, let q_S denote the student reasoning-action distribution and let q_T^+ denote the teacher distribution conditioned on a verified correct continuation. Let $\rho(a, a')$ be a latent semantic distance between two reasoning actions. We use the normalized representation

$$e(a) = \frac{E(a)}{\|E(a)\|_2}. \tag{21}$$

Assumption 1 (Local semantic fidelity)**.** *For the student and verified teacher reasoning actions considered at state h, there exist constants $0 < m \le L < \infty$ such that*

$$m\rho(a, a') \le \|e(a) - e(a')\|_2 \le L\rho(a, a'). \tag{22}$$

Assumption 1 is local to the reasoning actions that occur in the current task. It requires the encoder to preserve semantic neighborhoods on this support rather than over the entire natural-language space.

Define the squared Wasserstein distance in the reasoning-action space as

$$W_{2,\rho}^2(q_S, q_T^+) = \inf_{\gamma \in \Gamma(q_S, q_T^+)} \mathbb{E}_{(A,B)\sim\gamma} \left[\rho^2(A, B)\right]. \tag{23}$$

Let the corresponding push-forward distributions in the semantic space be

$$\nu_S = e_{\#} q_S, \qquad \nu_T = e_{\#} q_T^+. \tag{24}$$

Theorem 1 (Transport fidelity in the semantic space)**.** *Under Assumption 1,*

$$mW_{2,\rho}(q_S, q_T^+) \le W_2(\nu_S, \nu_T) \le LW_{2,\rho}(q_S, q_T^+). \tag{25}$$

Because the semantic representations are normalized, the population transport objective based on cosine cost,

$$\mathcal{T}(\nu_S, \nu_T) = \inf_{\eta \in \Gamma(\nu_S, \nu_T)} \mathbb{E}_{(Z_S, Z_T)\sim\eta} \left[1 - \langle Z_S, Z_T \rangle\right], \tag{26}$$

satisfies

$$\frac{m^2}{2} W_{2,\rho}^2(q_S, q_T^+) \le \mathcal{T}(\nu_S, \nu_T) \le \frac{L^2}{2} W_{2,\rho}^2(q_S, q_T^+). \tag{27}$$

Proof. For any coupling $\gamma \in \Gamma(q_S, q_T^+)$, applying e to both components gives a coupling between ν_S and ν_T. Assumption 1 implies

$$m^2\rho^2(a, b) \le \|e(a) - e(b)\|_2^2 \le L^2\rho^2(a, b). \tag{28}$$

Taking expectations over the coupling and then minimizing over feasible couplings gives equation 25. For normalized vectors,

$$1 - \langle z, z' \rangle = \frac{1}{2}\|z - z'\|_2^2. \tag{29}$$

Substituting this identity into the transport objective gives Equation 27. □

Theorem 1 gives a conditional local consistency result: when the encoder preserves semantic distances on the relevant reasoning-action support, the embedding-space transport objective remains comparable to the underlying reasoning-distribution discrepancy. The independent-encoder and frozen-judge evaluations in Appendix C provide empirical checks of this representation assumption.

E.2 Verified Teacher Distributions for Mechanism Evaluation

The main training procedure uses gold-conditioned teacher steps without filtering them according to the final answer. For the mechanism analysis and qualitative case studies, we additionally extend the sampled teacher steps into complete continuations and retain those that produce a correct final answer.

Let

$$A_1^T, \ldots, A_{K_T}^T \overset{\text{i.i.d.}}{\sim} q_T(\cdot \mid h, y^\star) \tag{30}$$

be independent teacher samples at state h. Let $Z_j = V(A_j^T) \in \{0, 1\}$ indicate whether the complete continuation associated with A_j^T produces a correct final answer. Define $p_h = \Pr(Z_j = 1 \mid h) > 0$ and let

$$M = \sum_{j=1}^{K_T} Z_j \tag{31}$$

be the number of verified teacher samples.

Proposition 2 (Verifier filtering as success-conditioned sampling)**.** *Conditioned on $M > 0$, for any bounded function f, the empirical average over the verified teacher samples satisfies*

$$\mathbb{E}\left[\frac{1}{M}\sum_{j=1}^{K_T} Z_j f(A_j^T) \,\middle|\, M > 0\right] = \mathbb{E}_{A\sim q_T^+}[f(A)], \tag{32}$$

where

$$q_T^+(a \mid h) = q_T(a \mid h, y^\star, V = 1). \tag{33}$$

Moreover, as $K_T \to \infty$, the empirical verified distribution

$$\widehat{q}_T^+ = \frac{1}{M}\sum_{j:Z_j=1} \delta_{A_j^T} \tag{34}$$

converges almost surely to q_T^+.

Proof. Because the teacher samples are conditionally independent and identically distributed, every sample that passes the verifier follows the success-conditioned distribution q_T^+. By exchangeability, the expected empirical average of the retained samples therefore equals the expectation under q_T^+, which gives Equation 32.

For the convergence result, the strong law of large numbers gives

$$\frac{1}{K_T}\sum_{j=1}^{K_T} Z_j f(A_j^T) \longrightarrow p_h\, \mathbb{E}[f(A) \mid V = 1] \tag{35}$$

and

$$\frac{M}{K_T} \longrightarrow p_h \tag{36}$$

almost surely. Taking the ratio of these two quantities gives the stated convergence to q_T^+. □

The probability that none of the K_T teacher samples passes the verifier is

$$\Pr(M = 0) = (1 - p_h)^{K_T}. \tag{37}$$

Therefore, increasing the number of teacher samples decreases the probability of obtaining no verified teacher continuation at state h. This verified distribution is used only for mechanism evaluation; the main training objective continues to use step-only teacher sampling without final-outcome filtering.

E.3 Approximation Error of Sinkhorn Transport

RMT-OPD uses entropy-regularized Sinkhorn transport instead of solving the exact optimal-transport problem. We bound the difference between the transport cost produced by the regularized coupling and the exact OT cost.

Let $C_{ij} = 1 - \langle z_i^S, z_j^T \rangle$ be the semantic cost matrix. The exact OT cost is

$$W_0 = \min_{\Pi \in U(\alpha,\beta)} \langle \Pi, C \rangle, \tag{38}$$

where $U(\alpha, \beta)$ denotes the set of transport plans with student marginal α and teacher marginal β.

The entropy-regularized transport plan is

$$\Pi_\varepsilon = \underset{\Pi \in U(\alpha,\beta)}{\arg\min} \left\{ \langle \Pi, C \rangle - \varepsilon H(\Pi) \right\}, \tag{39}$$

where

$$H(\Pi) = -\sum_{i,j} \Pi_{ij} \log \Pi_{ij}. \tag{40}$$

We define the transport cost associated with the regularized coupling as

$$W_\varepsilon^{\text{cost}} = \langle \Pi_\varepsilon, C \rangle. \tag{41}$$

Theorem 2 (Bounded entropy-regularization error)**.** *For a transport problem with K_S student samples and K_T teacher samples,*

$$0 \le W_\varepsilon^{\text{cost}} - W_0 \le \varepsilon \log(K_S K_T). \tag{42}$$

Proof. Let Π_0 be an optimal solution to the exact transport problem in Equation 38. Since Π_ε minimizes the regularized objective in Equation 39,

$$W_\varepsilon^{\text{cost}} - \varepsilon H(\Pi_\varepsilon) \le W_0 - \varepsilon H(\Pi_0). \tag{43}$$

Rearranging gives

$$W_\varepsilon^{\text{cost}} - W_0 \le \varepsilon \left[H(\Pi_\varepsilon) - H(\Pi_0) \right]. \tag{44}$$

Any probability coupling with $K_S K_T$ entries satisfies

$$0 \le H(\Pi) \le \log(K_S K_T). \tag{45}$$

Therefore,

$$W_\varepsilon^{\text{cost}} - W_0 \le \varepsilon \log(K_S K_T). \tag{46}$$

Because W_0 is the minimum unregularized transport cost over all feasible couplings and Π_ε is feasible, $W_\varepsilon^{\text{cost}} \ge W_0$. Combining the two inequalities proves Equation 42. □

Theorem 2 shows that the Sinkhorn coupling provides a controlled approximation to exact optimal transport. The regularization parameter ε determines the trade-off between numerical stability and transport fidelity, while the resulting deviation in the unregularized transport cost is bounded by $\varepsilon \log(K_S K_T)$.

E.4 Centroid Lower Bound and Distributional Shape Gap

The complete OT discrepancy may partly reflect a displacement between the semantic centers of the student and teacher candidate distributions. We derive a centroid-based lower bound to separate this displacement from changes in the internal distributional structure.

Define the empirical semantic centroids as

$$\bar{z}_S = \sum_i \alpha_i z_i^S, \qquad \bar{z}_T = \sum_j \beta_j z_j^T. \tag{47}$$

Because the embeddings are normalized, the cosine transport cost is equivalent to one half of the squared Euclidean distance. Therefore, for any feasible coupling Π,

$$\langle \Pi, C \rangle = \frac{1}{2} \mathbb{E}_\Pi \left[\| Z_S - Z_T \|_2^2 \right]. \tag{48}$$

Applying Jensen's inequality to the regularized transport coupling gives

$$W_\varepsilon^{\text{cost}} \ge \frac{1}{2} \| \bar{z}_S - \bar{z}_T \|_2^2. \tag{49}$$

Equality holds when $Z_S - Z_T$ is constant almost surely under Π_ε. We therefore define the distributional shape gap as

$$\Delta_{\text{shape}} = W_\varepsilon^{\text{cost}} - \frac{1}{2} \| \bar{z}_S - \bar{z}_T \|_2^2 \ge 0. \tag{50}$$

The first term captures the complete optimal matching between the student and teacher distributions, while the second term retains only the displacement between their semantic centroids. Consequently, Δ_{shape} measures the part of the transport discrepancy that cannot be explained by a centroid shift.

If the student and teacher candidate distributions differ mainly by a common semantic displacement, then Δ_{shape} is close to zero. In contrast, missing teacher modes or changes in probability-mass allocation produce additional distributional deformation and increase the shape gap. The state-level metric $G_{\text{shape}}(h)$ reported in Figure 3 is the empirical version of Equation 50.

E.5 Optimization Effect of Deficit-Weighted Supervision

We next explain how reasoning-mode deficits affect the direction of the student update. For a fixed student state h, define the absolute weight of teacher mode j as

$$w_j = \beta_j d_j. \tag{51}$$

Let

$$D = \sum_j w_j \tag{52}$$

be the total state-level deficit. When $D > 0$, the corresponding normalized deficit distribution is

$$q_D(j) = \frac{w_j}{D} = \frac{\beta_j d_j}{D}. \tag{53}$$

For the teacher continuation a_j^T, define the length-normalized negative log-likelihood as

$$\ell_j(\theta) = -\frac{1}{|a_j^T|} \log \pi_\theta(a_j^T \mid h). \tag{54}$$

The local correction loss can then be written as

$$\mathcal{L}_h = \sum_j w_j \ell_j(\theta) = D\, \mathbb{E}_{j \sim q_D} \left[\ell_j(\theta) \right]. \tag{55}$$

Proposition 3 (Directional pressure on under-covered modes)**.** *Consider a mode-level softmax model in which the student assigns probability*

$$p_j = \frac{\exp(u_j)}{\sum_k \exp(u_k)} \tag{56}$$

to teacher mode j at state h. Using the normalized deficit distribution q_D as the target gives the mode-level cross-entropy

$$\overline{\mathcal{L}}_h = -\sum_j q_D(j) \log p_j. \tag{57}$$

Its gradient with respect to mode logit u_j is

$$\frac{\partial \overline{\mathcal{L}}_h}{\partial u_j} = p_j - q_D(j). \tag{58}$$

Proof. Equation 58 follows from the standard gradient of softmax cross-entropy. When $q_D(j) > p_j$, the gradient with respect to u_j is negative, so gradient descent increases the relative logit of mode j. When $q_D(j) < p_j$, the update decreases its relative logit.

Under the ideal discrete-mode setting from the main analysis,

$$\beta_j d_j^{(0)} = (\beta_j - p_j)_+. \tag{59}$$

Substituting this identity into Equation 53 gives

$$q_D(j) \propto (\beta_j - p_j)_+. \tag{60}$$

Therefore, the positive correction target is concentrated on modes for which $p_j < \beta_j$. □

Proposition 3 describes the local mode-level effect of deficit weighting. In the full language model, the optimization variable is the likelihood of a complete teacher continuation rather than a finite set of mode logits. However, the gradient of the actual local loss still decomposes as

$$\nabla_\theta \mathcal{L}_h = \sum_j w_j \nabla_\theta \ell_j(\theta). \tag{61}$$

Thus, w_j directly controls the contribution of each teacher continuation to the student update. Larger transport deficits assign stronger learning pressure to teacher modes that are insufficiently covered by the current student.

E.6 On-Policy State Distribution and Iterative Resampling

RMT-OPD collects new student rollouts after each model update instead of repeatedly training on a fixed set of old failure states. We relate this design to the mismatch between the failure-state distributions of successive student policies.

Let $d_\theta^-(h)$ denote the distribution of intermediate reasoning states induced by student policy π_θ, conditioned on the final trajectory being incorrect. Let $g(h)$ be a fixed correction functional satisfying $0 \le g(h) \le B$.

Suppose the correction data were collected from student π_{θ_r}, while the current student has been updated to $\pi_{\theta_{r+1}}$. The difference between the corresponding expected correction values satisfies

$$\left| \mathbb{E}_{h\sim d^-_{\theta_{r+1}}}[g(h)] - \mathbb{E}_{h\sim d^-_{\theta_r}}[g(h)] \right| \le B \ \mathrm{TV}\left(d^-_{\theta_{r+1}}, d^-_{\theta_r}\right), \tag{62}$$

where TV denotes total-variation distance.

Proof. For any function g with values in $[0, B]$ and any two distributions P and Q,

$$|\mathbb{E}_P[g] - \mathbb{E}_Q[g]| \le B \ \mathrm{TV}(P, Q). \tag{63}$$

Setting $P = d^-_{\theta_{r+1}}$ and $Q = d^-_{\theta_r}$ gives Equation 62. □

When an RMT-OPD update changes the student reasoning policy, the new failure-state distribution $d^-_{\theta_{r+1}}$ may differ from the distribution $d^-_{\theta_r}$ used to construct the previous correction data. Continuing to train on the old states would therefore reintroduce a state-distribution mismatch.

RMT-OPD instead repeats the following process after each update:

$$\pi_{\theta_r} \longrightarrow \text{fresh student rollouts} \longrightarrow d^-_{\theta_r} \longrightarrow \text{new transport-based correction data.} \tag{64}$$

This iterative collection procedure keeps teacher supervision aligned with the failure states that the current student still visits. Equation 62 shows that the potential error from reusing stale states grows with the discrepancy between successive failure-state distributions. The bound concerns this state-distribution mismatch rather than the global convergence of the complete training procedure.