DIAL-OPD: LEARNING MORE FROM FEWER TOKENS IN ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) supervises student-generated trajectories with token-level teacher signals, and its sampled-token variant has become the practical default by avoiding the cost of full-vocabulary probabilities. Yet we find that supervising more tokens is not better: because the student's capacity is finite, training on a carefully chosen subset can significantly outperform full-token OPD. Effective OPD therefore hinges not on how much supervision to provide, but on which tokens deserve it. Existing disagreement-based criteria, however, are scale-blind: the log-ratio reward treats a token the teacher strongly endorses and one that both models assign negligible probability as equally informative, causing supervision to concentrate on *low-low* tokens that hinder learning. We propose **DIAL-OPD**, a token-selection method that bridges log-probability and probability spaces by weighting reward magnitude with the logarithmic mean of teacher and student probabilities. The parameter β controls this weighting, and the highest-scoring tokens are retained for training. We evaluate DIAL-OPD against 9 baselines across 4 teacher–student pairs and 7 mathematical reasoning benchmarks, showing that retaining only 40% of tokens, it outperforms Vanilla OPD and its full-token variants by up to **5.25** percentage points in mean accuracy, while doubling Vanilla OPD's AIME25 Pass@16 from 13.33% to 26.67%. It also outperforms the strongest token-selection baseline at matched retention ratios, with up to an **18**% relative improvement in mean accuracy. Even with a 4B teacher, DIAL-OPD surpasses the strongest full-token baseline using an 8B teacher at both student scales, **showing that effective supervision allocation can outweigh teacher scaling.** Selection analyses suggest that moderate β achieves the best balance between suppressing low-low tokens and preserving useful disagreements. Token-level studies reveal that DIAL-OPD filters high-reward tokens with limited reasoning value while preserving supervision essential to reasoning correctness.

1 INTRODUCTION

On-policy distillation (OPD) has become a standard component of LLM post-training (Qwen Team, 2025; Lu & Lab, 2025; DeepSeek-AI, 2026). By supervising student-generated trajectories, it reduces exposure bias relative to fixed-response distillation (Agarwal et al., 2024; Gu et al., 2026) and provides dense token-level guidance beyond outcome-based reinforcement learning (DeepSeek-AI, 2025). Full-vocabulary OPD minimizes the reverse KL between student and teacher distributions, but transferring and storing vocabulary-wide teacher probabilities incurs substantial communication and memory costs (Zhu et al., 2025; Sheng et al., 2024). Sampled-token OPD is thus the common practical choice: it uses student-sampled tokens to obtain an unbiased single-sample Monte Carlo estimate of the reverse KL and performs policy-gradient optimization with token-level rewards (Kimi Team, 2026; GLM-5 Team, 2026; Xiaomi LLM-Core Team, 2026).

In sampled-token OPD, supervision relies on a single-sample Monte Carlo estimate and is therefore noisy, raising a basic question: does more supervision necessarily improve learning? We study this question with Qwen3-4B as the teacher and Qwen3-0.6B as the student. Randomly selecting only a subset of tokens outperforms full-token Vanilla OPD at every tested retention ratio from 40% to 90%, with 80% retention improving MATH-500 accuracy by **8.2** percentage points (Figure 1). However, when the student is enlarged to match the 4B teacher, this advantage disappears (Figure 2). These results indicate that the benefit stems from the student's finite capacity: because the student cannot

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

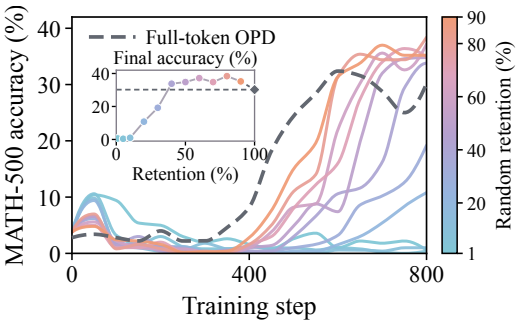

Figure 1: **Supervision density has a non-monotonic effect.** Random selection (1–90%) vs. full-token OPD (100%), 0.6B student; inset: final accuracy.

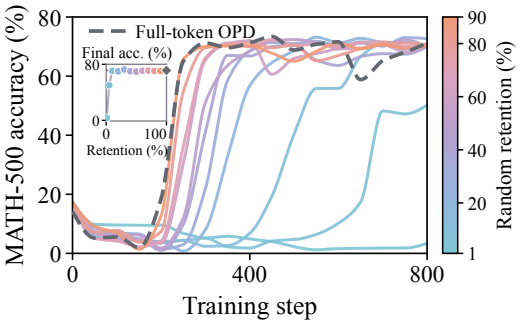

Figure 2: **The non-monotonic effect persists without a teacher–student size gap.** Random selection vs. full-token OPD, 4B student; inset: final accuracy.

absorb supervision from all tokens equally well, concentrating on the most effective token signals is better than training on all token signals. Thus, *more supervision does not necessarily improve learning in OPD*; the key is to select the most useful supervision rather than use all of it.

Random selection does not distinguish tokens by their learning value, motivating a more principled selection criterion. We first consider using teacher–student disagreement as such a criterion (Xu et al., 2026; Shao et al., 2026; Yu et al., 2026). In sampled-token OPD, the log-ratio reward compares the probabilities that the teacher and student assign to each sampled token and weights its policy-gradient term, making its magnitude a natural measure of disagreement. We find that selecting tokens with the largest absolute rewards outperforms both selecting those with the smallest rewards and random selection across various retention ratios (Figure 3). These results support larger reward magnitudes as a proxy for greater token learning value. However, this strategy exhibits prolonged stagnation early in training, and its final performance does not consistently exceed full-token Vanilla OPD. Examining the log-ratio reward reveals a key blind spot: **it measures relative disagreement but ignores absolute probability scale.** For example, student–teacher probabilities of $(0.08, 0.8)$ and $(10^{-4}, 10^{-3})$ yield the same reward, although the former reflects strong teacher support and the latter negligible probability under both models. We call tokens in the latter regime *low-low tokens*. Prior work identifies reliability issues in low-probability generation tails (Holtzman et al., 2020) and shows that suppressing rewards where both models assign low probability improves OPD training (Zhao et al., 2026a), motivating closer scrutiny of these tokens. We find that (i) when selecting tokens with the largest absolute rewards at 10% retention, low-low tokens account for only **20.6**% of selected tokens but **49.8**% of total absolute reward (Figure 4); (ii) replacing selected low-low tokens with non-low-low tokens at the same retention ratio mitigates early training stagnation and improves final accuracy across various retention ratios (Figure 5). Selecting tokens solely by reward magnitude can therefore concentrate supervision on tokens that hinder learning. *To address this blind spot, token selection should consider both the relative disagreement captured by reward magnitude and the absolute probability scale at which it occurs.*

Building on these findings, we propose **DIAL-OPD**, a token-selection method that bridges log-probability and probability spaces. We use the logarithmic mean of the probabilities assigned to each token by the teacher and student to incorporate probability scale into reward-based selection. A tunable parameter β controls the strength of probability weighting, reducing the priority of low-low tokens for $\beta > 0$. The resulting criterion provides a continuous dial: $\beta = 0$ recovers selection based solely on reward magnitude, $\beta = 1$ corresponds to selection by the absolute probability gap, and intermediate values geometrically interpolate between the two, while $\beta > 1$ further strengthens probability weighting. DIAL-OPD selects the highest-scoring tokens within each response at a specified retention ratio and applies the original OPD loss only to the selected tokens.

Experiments show that incorporating probability scale mitigates early training stagnation and improves final accuracy, with extensive β sweeps favoring moderate weighting. We evaluate DIAL-OPD against 9 baselines across 4 teacher–student pairs and 7 mathematical reasoning benchmarks. Retaining only **40**% of tokens, DIAL-OPD surpasses Vanilla OPD and the strongest OPD variant, which train on all response tokens, in mean accuracy across all settings. Notably, it doubles Vanilla OPD's AIME25 Pass@16 from **13.33**% to **26.67**%. At **20**% retention, DIAL-OPD exceeds the

strongest token-selection baseline at the same ratio by up to **5.23** points in mean accuracy and **10.33** points on GSM8K Avg@4. With a 4B teacher, it also surpasses the strongest full-token reference using an 8B teacher by up to **1.76** mean accuracy points, *demonstrating that effective supervision allocation can outweigh teacher scaling.* Selection analyses suggest that moderate weighting suppresses low-low tokens, whereas excessive weighting may discard useful disagreements. Token-level studies reveal that DIAL-OPD filters high-reward tokens with limited reasoning value while preserving supervision essential to reasoning correctness.

2 PRELIMINARIES

On-policy distillation (OPD) aligns a student policy π_θ with a fixed teacher π_T on student-generated responses (Agarwal et al., 2024; Gu et al., 2026). For $x \sim \mathcal{D}$, $y \sim \pi_\theta(\cdot \mid x)$, and $c_t = (x, y_{<t})$, OPD minimizes the reverse KL between their next-token distributions over vocabulary $\mathcal{V}$:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x,y}\left[\sum_{t=1}^{|y|} D_{\mathrm{KL}}(\pi_\theta(\cdot \mid c_t) \,\|\, \pi_T(\cdot \mid c_t))\right], \quad D_{\mathrm{KL}}(P\|Q) = \sum_{v\in\mathcal{V}} P(v)\log\frac{P(v)}{Q(v)}. \quad (1)$$

Transferring and storing full-vocabulary teacher distributions incurs substantial communication and memory costs; sampled-token OPD reduces these costs by querying only the teacher log-probabilities of student-generated tokens (Zhu et al., 2025; Sheng et al., 2024). With the rollout context distribution held fixed during differentiation, the reverse-KL descent direction takes the policy-gradient reinforcement learning (RL) form (Williams, 1992; Sutton et al., 1999):

$$g_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)}\left[\sum_{t=1}^{|y|} \underbrace{\log\frac{\pi_T(y_t \mid c_t)}{\pi_\theta(y_t \mid c_t)}}_{r_t^{\mathrm{OPD}}} \nabla_\theta \log \pi_\theta(y_t \mid c_t)\right]. \quad (2)$$

Student rollouts provide an unbiased Monte Carlo estimate of g_{OPD}, with reward r_t^{OPD} treated as a stop-gradient weight (Lu & Lab, 2025; Ko et al., 2026; Zhao et al., 2026a; derivation in Appendix C). This formulation is adopted by recent open-source LLMs (GLM-5 Team, 2026; Kimi Team, 2026; Xiaomi LLM-Core Team, 2026) and is the focus of this work.

3 RETHINKING SUPERVISION ALLOCATION IN ON-POLICY DISTILLATION

3.1 MORE SUPERVISION DOES NOT IMPLY BETTER LEARNING

Dense supervision is a central advantage of OPD, providing teacher guidance at every generated token (Lu & Lab, 2025). To test the implicit expectation that **denser supervision improves learning**, we conduct a controlled experiment comparing full-token training with uniform random token selection under a fixed update budget.

Experimental setup. We train Qwen3-0.6B with a Qwen3-4B teacher (Qwen Team, 2025) on DeepScaleR (Luo et al., 2025). Under a fixed update budget, we compare uniform random selection within each response ($\rho \in \{1\%, 5\%, 10\%, 20\%, \ldots, 90\%\}$) with full-token Vanilla OPD ($\rho = 100\%$). All tokens remain in context; the OPD loss is averaged over selected tokens in each batch. We track MATH-500 accuracy throughout training (Hendrycks et al., 2021; Lightman et al., 2023); implementation and evaluation details appear in Appendix B.1.

Results. Surprisingly, random token selection outperforms full-token Vanilla OPD at all tested retention ratios from 40% to 90%, reaching 38.4% accuracy at 80% versus 30.2%, a gain of **8.2** percentage points (Figure 1). However, this benefit may reflect the smaller student's limited capacity to utilize dense supervision from a larger teacher. To test this explanation, we use a Qwen3-4B student with the same teacher. Figure 2 shows that (i) 10% retention nearly matches full-token Vanilla OPD (70.6% vs. 71.2%); and (ii) 30% retention outperforms it (72.6%). Together, these results show that **more supervision does not necessarily improve learning**, even at equal teacher–student model size. *Since random selection already improves learning, how can we identify the token subset that maximizes learning performance under a fixed update budget?*

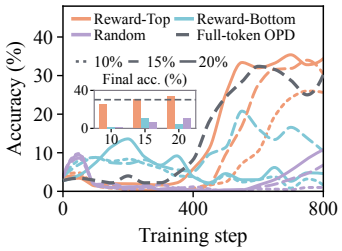

Figure 3: Reward-based selection; inset: final accuracy.

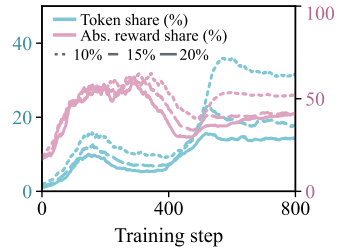

Figure 4: Low-low token and absolute reward shares (Reward-Top).

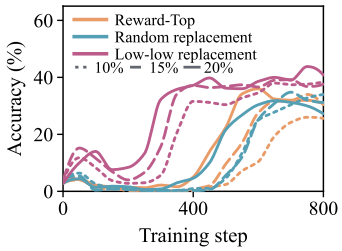

Figure 5: **Low-low replacement improves learning.**

3.2 Which Tokens Are Worth Learning From?

Random selection improves learning without distinguishing token value, motivating a more principled criterion. Teacher–student disagreement offers a natural starting point: recent work improves OPD by using disagreement or token update impact to guide selection (Xu et al., 2026; Shao et al., 2026; Yu et al., 2026). Full-vocabulary OPD measures disagreement by KL divergence; sampled-token OPD uses a log-ratio reward that weights each token's policy-gradient term (Equation 2). We therefore examine whether reward magnitude identifies tokens worth prioritizing.

Experimental setup. At 10%, 15%, and 20% retention per response, we compare *Reward-Top* (largest absolute rewards) with (i) *Reward-Bottom* (smallest), testing whether larger rewards are more useful; and (ii) *Random* (uniform sampling), testing whether reward-based selection outperforms uninformed selection. All methods use matched retention ratios, the original OPD loss, and the training and evaluation settings of Section 3.1.

Results. Reward-Top outperforms Reward-Bottom and Random at 10%, 15%, and 20% retention, reaching 25.6%, 30.4%, and 34.4% accuracy, respectively (Figure 3). However, (i) all three runs show little improvement during the first 350 training steps, indicating that large rewards do not readily translate into learning gains; and (ii) Reward-Top does not consistently outperform full-token Vanilla OPD. Thus, reward magnitude helps distinguish token learning value but is insufficient for reliable selection. *These limitations raise a question: what information does reward magnitude miss when assessing token learning value?*

3.3 Low–Low Supervision Can Hinder Learning

To understand these limitations, we examine the log-ratio reward: it captures relative disagreement but ignores absolute probability scale. Student–teacher probabilities of $(0.08, 0.8)$ and $(10^{-4}, 10^{-3})$ both yield $\log 10$, despite strong teacher support in the former and negligible probabilities under both models in the latter, which we term *low-low tokens*. Prior work identifies unreliable low-probability generation tails (Holtzman et al., 2020), while PowerOPD improves training stability and performance by suppressing low-low rewards (Zhao et al., 2026a). Accordingly, we test (i) whether Reward-Top overemphasizes low-low tokens and (ii) whether training on them hinders learning.

Experimental setup. Following Section 3.2, we study Reward-Top at 10%, 15%, and 20% retention, defining low-low tokens by student and teacher probabilities both at most 1%. (i) To quantify their concentration, we track their shares of selected tokens and total absolute reward throughout training. (ii) To test whether they hinder learning, we replace selected low-low tokens with the highest-reward unselected non-low-low tokens (*Low-low replacement*). As a control, *Random replacement* removes the same number of randomly selected tokens and refills by reward rank without excluding low-low tokens. Both preserve the retention ratio and OPD loss (Appendix B.3).

Results. (i) *Low-low tokens carry disproportionate reward.* At 10%, 15%, and 20% retention, they constitute 20.59%, 13.49%, and 10.26% of selected tokens but contribute 49.80%, 41.43%, and 38.01% of total absolute reward, respectively (Figure 4). (ii) *Targeted replacement accelerates learning and improves accuracy.* Low-low replacement yields earlier gains and accuracies of 38.4%, 37.6%, and 40.8%, exceeding Reward-Top by 12.8, 7.2, and 6.4 percentage points, respec-

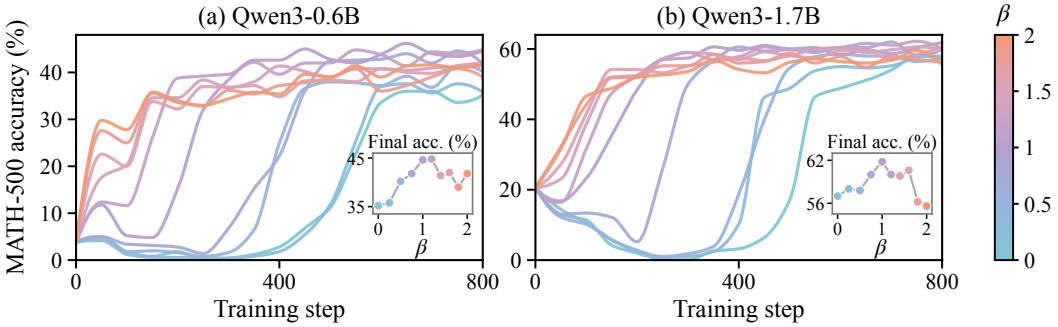

Figure 6: Learning dynamics of DIAL-OPD across β for Qwen3-0.6B and Qwen3-1.7B students, using a Qwen3-4B teacher and 40% token retention. Insets show final accuracy as a function of β.

tively (Figure 5). It also consistently outperforms random replacement, supporting targeted low-low removal. Thus, **low-low supervision can hinder learning despite large rewards**, indicating that *token selection should consider both relative disagreement and absolute probability scale.*

4 METHOD

Reward magnitude helps identify valuable tokens, yet low-low tokens contribute disproportionate absolute reward and can hinder learning. We therefore propose **DIAL-OPD**, a token-selection method that *preserves the relative-disagreement signal captured by reward magnitude while incorporating absolute probability scale to downweight low-low tokens.*

4.1 BRIDGING LOG-PROBABILITY AND PROBABILITY SPACES

Let $p_t = \pi_\theta(y_t \mid c_t)$ and $q_t = \pi_T(y_t \mid c_t)$ denote student and teacher probabilities for a sampled token. To make log-ratio-based selection probability-aware, we use the logarithmic mean (Bhatia, 2008), which averages probabilities along a uniform interpolation between $\log p_t$ and $\log q_t$:

$$L(p_t, q_t) = \int_0^1 p_t^{1-u} q_t^u \, du, \tag{3}$$

where $u \in [0, 1]$ parameterizes the interpolation and $p_t^{1-u} q_t^u$ is the corresponding probability. Since these probabilities lie between p_t and q_t, *low probabilities under both models yield a small logarithmic mean.* DIAL-OPD incorporates this probability scale by weighting the OPD reward magnitude with $L(p_t, q_t)^\beta$, where β controls the weighting strength:

$$\boxed{s_{\beta,t} = L(p_t, q_t)^\beta \left| r_t^{\mathrm{OPD}} \right|}, \qquad \beta \geq 0. \tag{4}$$

For $\beta > 0$, the small logarithmic mean of low-low tokens reduces their selection scores. Substituting $L(p_t, q_t) = (q_t - p_t)/(\log q_t - \log p_t)$ for $p_t \neq q_t$ connects relative and absolute disagreement:

$$s_{\beta,t} = \left(\frac{|q_t - p_t|}{|r_t^{\mathrm{OPD}}|} \right)^\beta \left| r_t^{\mathrm{OPD}} \right| = \left| r_t^{\mathrm{OPD}} \right|^{1-\beta} |q_t - p_t|^\beta. \tag{5}$$

At $\beta = 0$ and $\beta = 1$, selection uses the OPD reward magnitude and absolute probability gap, respectively. Intermediate values geometrically interpolate between them, while $\beta > 1$ strengthens probability weighting. **DIAL-OPD thus provides a continuous dial between relative and absolute disagreement for token selection.**

4.2 THE ROLE OF β IN DIAL-OPD

We examine how β affects token selection when (i) student and teacher probabilities are both low, or (ii) one is low and the other is moderate or high.

Proposition 1 (Controllable Suppression of Low-Low Disagreement) *For $p, q \in (0, 1]$ and $\beta \geq 0$, scaling both probabilities by $\varepsilon \in (0, 1]$ changes the selection score according to*

$$s_\beta(\varepsilon p, \varepsilon q) = \varepsilon^\beta s_\beta(p, q). \tag{6}$$

For $\varepsilon < 1$ and $p \neq q$, the score remains unchanged at $\beta = 0$, while larger $\beta > 0$ increasingly suppresses low-low disagreement. We next examine when only one model assigns low probability.

Proposition 2 (Sensitivity to One-Sided Low Probabilities) *For fixed $a \in (0, 1]$ and $\beta \geq 0$,*

$$\lim_{b \to 0^+} s_\beta(a, b) = \lim_{b \to 0^+} s_\beta(b, a) = \begin{cases} +\infty, & 0 \leq \beta < 1, \\ a, & \beta = 1, \\ 0, & \beta > 1. \end{cases} \tag{7}$$

Proofs are given in Appendices E.1 and E.2. Together, these propositions reveal a trade-off: increasing β strengthens low-low suppression, but $\beta > 1$ also drives scores to zero when only one probability vanishes. Only $\beta = 1$ preserves a finite, nonzero limit in this case.

4.3 DIAL-OPD

Let $\mathcal{V}_i$ denote the candidate token positions in student-generated response i, with $|\mathcal{V}_i| > 0$. At retention ratio $\rho \in (0, 1]$, DIAL-OPD selects K_i tokens maximizing their total score:

$$K_i = \max\{1, \mathrm{round}(\rho|\mathcal{V}_i|)\}, \quad \mathcal{S}_i^\star \in \operatorname*{arg\,max}_{\substack{\mathcal{S} \subseteq \mathcal{V}_i \\ |\mathcal{S}| = K_i}} \sum_{t \in \mathcal{S}} s_{\beta,i,t}, \quad m_{i,t} = \mathbb{I}\{t \in \mathcal{S}_i^\star\}. \tag{8}$$

Using the binary selection mask $m_{i,t}$, DIAL-OPD applies the original signed OPD reward (Section 2) only to selected tokens. For a training batch $\mathcal{B}$, it minimizes the masked objective:

$$\mathcal{L}_{\mathrm{DIAL}}(\theta) = -\frac{1}{M_\mathcal{B}} \sum_{i \in \mathcal{B}} \sum_{t \in \mathcal{V}_i} m_{i,t}\, \mathrm{sg}[r_{i,t}^{\mathrm{OPD}}] \log \pi_\theta(y_{i,t} \mid c_{i,t}), \qquad M_\mathcal{B} = \sum_{i \in \mathcal{B}} \sum_{t \in \mathcal{V}_i} m_{i,t}. \tag{9}$$

Here, sg denotes stop-gradient, with the mask fixed during differentiation. Unselected tokens remain in context but are excluded from the loss. DIAL-OPD recomputes scores and masks at each training step using fresh rollouts from the current student policy.

5 Experimental Setup

Models and training. We train Qwen3 students (0.6B/1.7B) with 4B/8B teachers (Qwen Team, 2025) on DeepScaleR (Luo et al., 2025) for 800 steps (details in Appendix B).

Evaluation. We evaluate on GSM8K (Cobbe et al., 2021), MATH-500 (Lightman et al., 2023), AMC23, AIME24, AIME25 (Dekoninck et al., 2026), Minerva Math (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024).

Baselines. **(i)** *Reference baselines* include full-token Vanilla OPD, uniform Random selection, PowerOPD with bounded power-transformed rewards (Zhao et al., 2026a), and ExOPD with extrapolated teacher–reference rewards (Yang et al., 2026a). **(ii)** *Token-selection baselines* include TIP, combining student entropy and teacher–student divergence (Xu et al., 2026); SA-OPD, using input dependence and distillation divergence (Jiang et al., 2026); TA-OPD, combining disagreement and teacher support (Wang et al., 2026); and Prefix OPD, retaining response prefixes (Zhang et al., 2026). We also include Capability-Contrast OPD (CC-OPD), adapting Direct-OPD's policy contrast (Feng et al., 2026) through absolute teacher–reference probability gaps. Token-selection methods are compared at matched retention ratios; implementation details appear in Appendix B.2.

6 Experiments

6.1 The Effect of β on Learning

DIAL-OPD uses probability scale to downweight low-low tokens, with β controlling the weighting strength. Increasing β suppresses low-low disagreement more strongly, but may also reduce the

Table 1: Results with a Qwen3-4B teacher on seven benchmarks. **Bold** and underlined mark the best and second-best accuracy per student and retention ratio, excluding full-token references.

Keep	Method	GSM8K			MATH500			AMC23			AIME24			AIME25			Minerva			Olympiad			Mean
		Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg
Student: Qwen3-0.6B																							
	Vanilla OPD	58.5	79.7	506	35.9	52.6	1687	17.3	50.0	5059	1.0	6.7	7266	1.2	10.0	7369	9.2	22.8	1436	12.8	24.5	5412	19.4
100%	PowerOPD	68.5	84.9	428	43.5	60.0	1511	23.6	62.5	4338	1.7	3.3	6535	1.0	10.0	6816	10.9	22.8	1229	17.5	30.3	4531	23.8
	ExOPD	54.8	78.8	533	34.4	52.6	1876	13.1	45.0	5512	1.7	6.7	7416	0.6	10.0	7704	8.2	20.2	1702	12.2	24.5	5565	17.9
	Random selection	38.8	70.7	1880	12.0	32.0	3196	5.6	32.5	6893	0.0	0.0	7787	0.4	6.7	7555	4.0	15.4	2996	4.8	14.4	7057	9.4
	TIP	54.7	76.4	610	29.9	48.0	2058	10.8	**55.0**	5663	0.8	3.3	7611	0.2	3.3	7801	6.8	19.1	1856	10.9	22.8	5856	16.3
	SA-OPD	56.6	80.0	533	35.5	52.4	1731	15.6	52.5	5142	1.2	**6.7**	6979	1.0	**10.0**	7163	**9.3**	22.8	1429	13.8	26.6	5296	19.0
	TA-OPD	7.6	25.8	3160	5.0	17.4	3386	2.2	30.0	6711	0.0	0.0	7097	0.2	3.3	7209	1.7	9.9	3218	2.1	7.4	6876	2.7
20%	Prefix OPD	45.6	75.5	1087	23.8	47.8	2307	11.4	47.5	5736	0.4	3.3	7041	0.8	6.7	7290	5.8	19.5	2133	9.4	23.6	5773	13.9
	CC-OPD	50.2	66.9	2615	21.7	35.0	3523	6.4	27.5	7934	0.0	0.0	8192	1.0	3.3	8192	4.4	14.0	3775	6.2	13.8	7883	12.8
	DIAL-OPD (β=1)	**67.0**	81.3	1072	34.2	47.2	2609	15.9	45.0	6554	0.4	**6.7**	7985	**1.5**	**10.0**	8016	7.6	16.9	2617	13.1	24.2	6658	19.9
	DIAL-OPD (β=1.2)	63.4	**82.6**	553	**39.7**	**54.6**	1793	**17.3**	52.5	4910	**1.9**	3.3	6874	0.8	**10.0**	7239	**9.3**	**23.9**	1522	**13.9**	**27.4**	4907	**20.9**
	Random selection	53.8	80.0	610	33.7	52.0	1997	14.1	52.5	5545	0.6	3.3	7376	0.6	6.7	7381	8.0	22.1	1730	11.8	24.5	5724	17.5
	TIP	56.6	79.2	576	35.1	51.4	1829	16.9	**65.0**	5321	1.0	3.3	7405	0.8	6.7	7666	8.6	21.7	1622	12.3	24.8	5520	18.8
	SA-OPD	55.8	79.8	500	37.0	53.0	1749	14.7	55.0	5142	1.0	3.3	6997	0.6	10.0	7447	8.5	21.0	1491	12.6	24.5	5491	18.6
	TA-OPD	11.9	38.1	2574	7.8	24.6	2782	3.0	27.5	5864	0.2	3.3	6250	0.2	3.3	6177	2.6	15.1	2744	2.7	9.2	6042	4.1
40%	Prefix OPD	50.3	78.0	754	32.2	50.2	2099	13.8	52.5	5715	0.4	3.3	7379	0.4	3.3	7596	8.7	21.3	1863	10.2	21.8	5964	16.6
	CC-OPD	67.5	82.7	1120	39.9	53.6	2312	22.7	55.0	6381	1.7	10.0	7916	**2.3**	**13.3**	7931	8.9	22.1	2302	16.1	27.0	6225	22.7
	DIAL-OPD (β=1)	**69.1**	**84.7**	578	**44.9**	**61.6**	1714	**25.5**	50.0	4774	**2.5**	**13.3**	7001	1.2	6.7	7179	10.7	22.1	1484	**18.8**	**33.5**	4705	**24.7**
	DIAL-OPD (β=1.2)	67.2	83.5	439	44.0	59.0	1521	25.2	62.5	4520	2.1	10.0	6761	0.2	3.3	6884	**10.9**	**25.7**	1216	17.1	29.4	4604	23.8
Student: Qwen3-1.7B																							
	Vanilla OPD	74.2	92.6	348	58.0	73.4	1150	33.8	72.5	3471	7.7	16.7	6150	7.3	26.7	5791	17.2	30.9	1018	26.6	41.2	3887	32.1
100%	PowerOPD	81.5	92.6	340	59.6	74.2	1208	35.5	75.0	3453	7.9	20.0	5946	7.5	26.7	5668	18.4	32.0	992	28.5	43.3	3768	34.1
	ExOPD	65.0	90.6	327	47.3	68.6	1265	25.9	65.0	3653	6.2	23.3	5951	4.8	23.3	5435	14.1	30.1	1047	22.3	39.0	3843	26.5
	Random selection	36.4	78.7	1173	15.8	40.6	2126	7.3	47.5	5092	1.0	13.3	5719	2.5	13.3	5482	4.9	23.2	2285	6.3	19.1	4847	10.6
	TIP	64.3	89.8	346	47.4	70.0	1167	25.6	**75.0**	3561	6.0	20.0	5818	3.8	16.7	5516	13.0	28.3	987	19.4	35.5	3725	25.7
	SA-OPD	68.9	91.0	367	52.9	70.4	1155	30.2	**75.0**	3529	5.6	16.7	5592	4.8	**26.7**	5545	15.7	30.5	931	23.8	38.6	3617	28.8
	TA-OPD	42.6	83.0	1472	35.1	67.6	1930	18.0	62.5	4506	5.4	20.0	5738	4.0	16.7	5472	10.4	27.6	1854	16.9	34.9	4433	18.9
20%	Prefix OPD	54.3	86.6	459	36.6	65.6	1218	21.6	62.5	3101	4.6	20.0	5331	3.3	16.7	4745	10.9	29.0	1170	15.9	34.0	3100	21.0
	CC-OPD	79.6	92.8	662	48.4	62.2	1949	28.9	62.5	5556	4.4	20.0	7717	4.8	16.7	7665	15.7	31.6	2030	20.3	32.2	5736	28.9
	DIAL-OPD (β=1)	**82.3**	91.8	404	**57.8**	70.2	1445	**37.8**	**75.0**	4276	**8.8**	**26.7**	6287	4.8	16.7	6170	**17.9**	**32.7**	1216	**29.5**	**43.2**	4222	**34.1**
	DIAL-OPD (β=1.2)	81.7	**93.5**	342	57.7	**73.0**	1288	35.9	72.5	3640	7.5	20.0	6040	**6.9**	23.3	5730	17.0	30.9	1045	26.5	42.6	3911	33.3
	Random selection	63.8	90.8	349	50.1	72.2	1076	28.3	70.0	3229	6.2	**26.7**	5177	5.6	13.3	4830	15.4	30.5	947	23.0	40.4	3382	27.5
	TIP	70.8	91.4	333	50.7	70.4	1151	29.2	72.5	3264	7.9	23.3	5789	6.5	20.0	5571	14.2	28.3	933	23.7	38.7	3765	29.0
	SA-OPD	68.9	91.6	343	54.3	72.2	1104	30.5	72.5	3605	7.7	20.0	5885	**7.7**	16.7	5210	16.3	32.0	913	25.1	41.2	3634	30.1
	TA-OPD	9.2	31.2	2890	6.7	23.6	3040	2.7	27.5	6197	0.8	10.0	6496	0.2	3.3	6174	2.8	15.8	2957	2.7	9.9	6384	3.6
40%	Prefix OPD	57.7	88.5	435	40.0	67.6	1364	23.4	65.0	3994	5.4	20.0	6371	4.2	16.7	5823	12.4	27.2	1256	19.5	36.5	3980	23.2
	CC-OPD	81.6	92.8	376	59.2	73.2	1337	34.2	70.0	4265	8.1	23.3	6608	7.5	**26.7**	6483	18.2	29.4	1077	28.2	42.1	4339	33.9
	DIAL-OPD (β=1)	**83.2**	92.8	344	**61.5**	73.0	1223	**38.8**	**75.0**	3457	9.8	**26.7**	5992	5.2	16.7	5951	**19.4**	30.9	992	**29.7**	**44.5**	3789	**35.4**
	DIAL-OPD (β=1.2)	82.5	**93.0**	330	61.0	**74.4**	1199	36.6	**75.0**	3679	**10.0**	**26.7**	6191	7.3	**26.7**	5934	19.1	**33.1**	977	27.8	40.9	3887	34.9

priority of useful disagreements supported by one model. We therefore examine how this trade-off affects learning dynamics and final accuracy across different student model sizes.

Experimental setup. We fix the retention ratio at 40% and sweep β over $\{0, 0.25, 0.5, 0.75, 1, 1.2, 1.4, 1.6, 1.8, 2\}$. We use Qwen3-4B as the teacher and Qwen3-0.6B and Qwen3-1.7B as students, following the training settings in Section 5. Figure 6 reports MATH-500 learning curves, with insets summarizing final accuracy.

Results. Figure 6 shows (i) shorter initial stagnation as β increases from 0 to 1 and (ii) non-monotonic final accuracy. For the 0.6B student, accuracy reaches 35.2%, 44.6%, and a peak of 44.8% at $\beta = 0, 1, 1.2$, respectively. For the 1.7B student, accuracy peaks at 61.8% ($\beta = 1$), versus 57.0% ($\beta = 0$) and 60.0% ($\beta = 1.2$). Appendix D compares $\beta = 0$ and $\beta = 1$ across all seven benchmarks. *Probability weighting improves learning speed and final accuracy over reward-only selection, but excessive weighting reduces performance: both students peak near $\beta = 1$.*

6.2 Main Results

Better performance with less supervision. Retaining only 40% of response tokens, DIAL-OPD outperforms Vanilla OPD and both full-token variants in seven-benchmark mean accuracy across all four model pairs. Against Vanilla OPD, gains reach **5.25** percentage points in mean accuracy

Table 2: Results with a Qwen3-8B teacher on seven benchmarks. **Bold** and underlined mark the best and second-best accuracy per student and retention ratio, excluding full-token references.

Keep	Method	GSM8K			MATH500			AMC23			AIME24			AIME25			Minerva			Olympiad			Mean
		Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg
Student: Qwen3-0.6B																							
	Vanilla OPD	58.4	81.1	483	39.2	56.8	1573	16.2	57.5	4488	1.7	6.7	6651	1.0	6.7	7033	9.1	21.0	1327	14.1	26.9	4864	20.0
100%	PowerOPD	67.2	83.7	404	43.8	61.2	1513	23.8	57.5	4249	2.3	10.0	6417	1.5	10.0	6577	10.6	21.7	1163	17.0	30.9	4415	23.7
	ExOPD	57.0	79.8	442	38.1	57.0	1568	17.7	52.5	4496	2.3	13.3	6715	0.8	6.7	7153	8.7	22.8	1303	13.5	26.1	5076	19.7
	Random selection	38.2	70.9	1314	9.9	27.0	2879	6.2	45.0	6345	0.0	0.0	7430	1.0	10.0	7201	3.5	15.8	2420	3.7	11.3	6611	9.0
	TIP	56.8	81.0	515	34.4	51.8	1706	13.1	42.5	5191	1.5	3.3	7147	0.8	6.7	7231	8.5	21.7	1526	12.8	26.3	5333	18.3
	SA-OPD	59.2	79.8	463	40.0	57.0	1596	19.5	55.0	4603	1.5	3.3	6502	1.7	**13.3**	7003	9.4	22.8	1301	13.7	25.7	4911	20.7
	TA-OPD	8.6	28.1	3109	6.0	19.8	3286	2.8	30.0	6722	0.4	3.3	6910	0.4	3.3	7076	1.8	10.7	3148	2.9	9.9	6795	3.3
20%	Prefix OPD	49.7	77.0	722	30.6	51.2	1889	13.4	50.0	5274	0.8	3.3	6911	1.2	**13.3**	7144	8.2	22.1	1597	10.2	23.3	5256	16.3
	CC-OPD	54.3	73.1	2035	26.3	39.4	3219	9.7	32.5	7672	0.0	0.0	8179	1.0	3.3	8188	5.1	15.8	3464	7.1	14.8	7654	14.8
	DIAL-OPD (β=1)	63.8	83.2	465	**42.3**	59.6	1633	21.2	**62.5**	4372	**1.7**	**10.0**	6566	**2.1**	10.0	6583	**9.7**	**23.2**	1265	**16.0**	**29.1**	4494	**22.4**
	DIAL-OPD (β=1.2)	**63.9**	**84.1**	469	40.5	**60.0**	1701	**21.6**	57.5	4556	0.8	3.3	6544	1.0	10.0	6786	8.7	**23.2**	1351	13.6	25.7	4567	21.4
	Random selection	55.9	79.7	627	36.4	54.0	1715	15.8	47.5	4963	1.5	3.3	6809	1.5	10.0	7066	8.9	22.8	1449	13.7	26.6	4988	19.1
	TIP	59.2	81.6	468	39.6	57.8	1587	18.4	57.5	4665	1.9	10.0	7000	0.8	10.0	7065	8.6	22.8	1344	14.5	26.4	5132	20.4
	SA-OPD	58.8	80.7	459	39.7	56.4	1529	19.2	60.0	4517	2.3	10.0	6681	0.2	3.3	7151	9.4	19.9	1261	15.6	26.7	4739	20.7
	TA-OPD	17.3	48.7	2615	12.1	34.0	2881	7.5	47.5	6181	0.8	6.7	6850	0.2	3.3	6944	3.8	17.6	2715	5.0	16.2	6151	6.7
40%	Prefix OPD	54.5	79.8	627	37.0	56.6	1868	16.4	57.5	5483	0.6	3.3	7401	0.6	6.7	7555	9.2	**23.9**	1561	12.1	24.0	5597	18.6
	CC-OPD	**69.2**	**85.1**	942	40.8	53.8	2163	23.6	52.5	6103	2.9	6.7	7683	1.7	10.0	7972	9.4	21.3	2089	15.7	27.4	6170	23.3
	DIAL-OPD (β=1)	66.6	84.6	436	**45.6**	**61.4**	1482	**27.3**	62.5	3872	**3.1**	**13.3**	6432	1.2	6.7	6589	**10.6**	22.1	1197	17.7	29.7	4392	**24.6**
	DIAL-OPD (β=1.2)	65.6	82.7	438	43.2	60.4	1496	26.7	**65.0**	4165	1.5	3.3	6162	**1.9**	**13.3**	6526	10.4	22.8	1216	**18.1**	**31.0**	4376	23.9
Student: Qwen3-1.7B																							
	Vanilla OPD	78.7	92.5	348	59.5	72.8	1201	34.7	77.5	3457	7.5	20.0	6135	4.2	13.3	5501	17.9	30.5	983	27.7	42.1	3718	32.9
100%	PowerOPD	80.1	92.4	327	58.3	71.2	1160	35.6	72.5	3343	9.2	30.0	5593	6.7	26.7	5118	16.8	31.2	925	27.6	40.5	3531	33.5
	ExOPD	80.3	92.3	337	58.7	73.0	1191	35.2	70.0	3413	9.2	26.7	5640	6.2	20.0	5323	18.0	28.7	977	27.7	41.8	3721	33.6
	Random selection	66.8	91.0	541	49.0	70.0	1290	26.4	65.0	3323	6.0	16.7	5586	5.2	**23.3**	5176	15.5	30.9	1164	22.3	37.7	3572	27.3
	TIP	74.9	91.8	354	54.6	71.4	1167	28.4	**75.0**	3794	**10.0**	**26.7**	5756	6.2	16.7	5284	15.7	29.8	942	24.1	38.7	3785	30.6
	SA-OPD	75.9	91.4	331	58.6	73.8	1093	31.9	**75.0**	3240	6.2	20.0	5838	6.7	**23.3**	5009	17.1	**32.0**	905	25.7	40.7	3540	31.7
	TA-OPD	80.5	**93.3**	337	58.4	**74.4**	1066	35.5	**75.0**	2977	9.0	23.3	5310	6.5	20.0	4810	17.9	**32.0**	895	26.1	40.8	3163	33.4
20%	Prefix OPD	76.1	91.2	390	52.2	66.0	1181	31.4	67.5	3355	5.8	23.3	5872	5.6	**23.3**	5303	17.1	31.6	1017	21.8	34.3	3528	30.0
	CC-OPD	76.2	91.8	567	49.8	67.2	1770	29.5	60.0	5146	5.0	20.0	7371	5.4	16.7	7150	14.4	30.9	1645	21.8	34.9	5313	28.9
	DIAL-OPD (β=1)	**80.6**	92.9	333	**60.2**	**74.4**	1199	**35.8**	70.0	3404	7.9	16.7	5791	5.4	16.7	5339	**18.3**	31.2	967	**28.2**	**42.3**	3570	**33.8**
	DIAL-OPD (β=1.2)	77.6	91.5	336	59.1	73.6	1213	34.5	72.5	3478	7.5	**26.7**	5793	**6.9**	16.7	5678	17.3	31.6	992	26.4	41.8	3775	32.8
	Random selection	75.5	91.6	362	55.5	70.8	1141	30.8	67.5	3025	**9.2**	**26.7**	5670	4.4	**26.7**	5339	16.5	30.1	983	27.0	41.8	3518	31.3
	TIP	75.7	91.1	346	58.1	**74.2**	1127	35.0	75.0	3422	7.9	**26.7**	5727	4.6	16.7	5441	17.6	**32.7**	961	26.0	41.8	3674	32.1
	SA-OPD	76.7	91.1	330	59.0	72.6	1153	33.0	80.0	3225	7.1	20.0	5637	5.4	20.0	4997	17.9	30.9	887	26.5	41.2	3584	32.2
	TA-OPD	79.1	92.3	323	57.9	71.6	1113	34.2	72.5	3123	8.8	16.7	5622	**6.9**	16.7	4898	**18.2**	32.0	867	27.5	41.1	3460	33.2
40%	Prefix OPD	76.9	92.2	359	54.3	67.4	1199	32.0	75.0	3696	7.9	20.0	6420	4.8	23.3	5363	16.8	30.1	1011	22.1	34.9	3839	30.7
	CC-OPD	79.9	92.5	346	59.2	73.2	1245	35.8	72.5	3800	7.9	23.3	6323	5.6	23.3	6121	17.6	30.5	982	27.4	40.4	4047	33.3
	DIAL-OPD (β=1)	**82.4**	**93.5**	323	**59.7**	73.8	1173	**38.0**	**82.5**	3098	8.8	16.7	5764	6.5	23.3	5425	18.0	31.6	913	**28.9**	**43.5**	3533	**34.6**
	DIAL-OPD (β=1.2)	82.1	92.9	331	**59.7**	72.4	1156	36.4	77.5	3134	**9.2**	20.0	5823	6.0	**26.7**	5306	17.1	28.7	898	27.8	42.4	3536	34.1

(4B→0.6B, $\beta = 1$) and **13.33** points on AIME25 Pass@16 (8B→1.7B, $\beta = 1.2$), doubling it from 13.33% to 26.67%. Against the strongest full-token variant, gains reach **1.24** points in mean accuracy (4B→1.7B) and **10.00** points on AMC23 Pass@16 (8B→1.7B), both at $\beta = 1$ and over PowerOPD. *These results challenge supervision density as a guiding principle for effective distillation: substantially less supervision can outperform standard and improved full-token OPD.*

Substantially outperforming existing token-selection methods. DIAL-OPD achieves the highest seven-benchmark mean accuracy across all eight combinations of model pair and retention ratio. At 20% retention, it exceeds the strongest existing selection method by up to **5.23** percentage points in mean accuracy (4B→1.7B, over CC-OPD) and **10.33** points on GSM8K Avg@4 (4B→0.6B, over SA-OPD), both at $\beta = 1$. *Together, these comparisons demonstrate that DIAL-OPD identifies more effective supervision than existing token-selection methods at the same token budget.*

Effective supervision allocation can outweigh teacher scaling. DIAL-OPD with a 4B teacher outperforms full-token OPD with an 8B teacher at both student scales. With $\beta = 1$ and 40% retention, it achieves mean accuracies of 24.68% and 35.36% for the 0.6B and 1.7B students, respectively, exceeding not only Vanilla OPD but also the strongest full-token variant under the larger teacher by **0.95** and **1.76** percentage points. *These findings suggest that allocating supervision effectively can yield greater gains than simply increasing teacher size.*

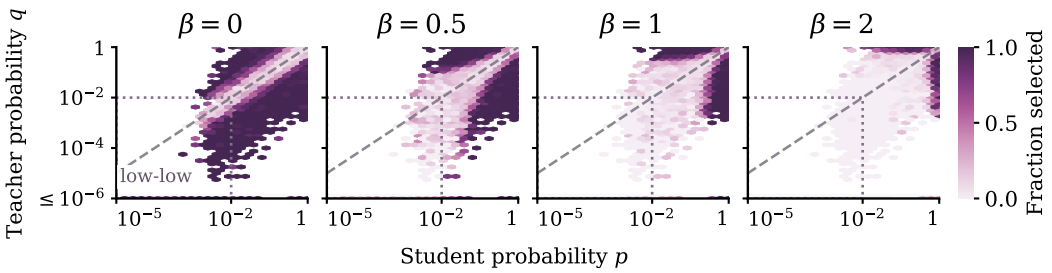

Figure 7: Token selection under varying β at 20% retention. Hexagon colors indicate selection rates; dotted lines delimit the low-low region, and teacher probabilities $q \leq 10^{-6}$ are pooled for visualization.

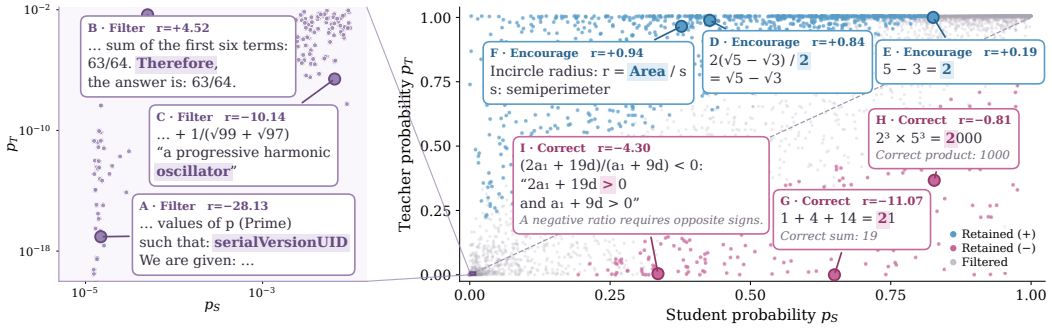

Figure 8: Token selections and signed OPD rewards in context ($\beta = 1$, 20% retention). Blue/pink mark retained tokens with positive/negative rewards; gray marks filtered tokens. Left: low-low zoom on log axes.

6.3 Understanding DIAL-OPD's Selection Behavior

Moderate β achieves the best balance in token selection. We examine how β, the dial between log-probability and probability spaces, shapes token selection. We compare $\beta = 0, 0.5, 1, 2$ on the same 128 responses to DeepScaleR prompts at 20% retention. Figure 7 reveals two stages: (i) increasing β from 0 to 1 shifts selection away from the low-low region, reducing its share of selected tokens from 8.26% to 0.29%; (ii) increasing β to 2 suppresses this share to 0.02%, but also reduces selection in regions where one model assigns substantial probability while the other does not. Together with the performance trends in Section 6.1, these observations suggest that *moderate probability weighting improves performance by suppressing low-low disagreement, whereas excessive weighting may degrade performance by discarding useful signals supported by one model.*

Reward magnitude alone is insufficient to assess learning value. Figure 8 shows retained and filtered tokens in context ($\beta = 1$, 20% retention). DIAL-OPD filters the unrelated identifier `serialVersionUID` (-28.13) and `Therefore` ($+4.52$), a redundant transition to an already-obtained answer. Retained supervision instead encourages `Area` in a correct incircle formula ($+0.94$) and penalizes an incorrect arithmetic digit (-0.81). *Tokens that do not advance reasoning can receive larger reward magnitudes than those providing useful supervision.* This contrast supports using both relative disagreement and absolute probability scale for selection.

7 Conclusion

We show that more supervision does not necessarily improve OPD, motivating selective supervision. Yet prioritizing large teacher–student disagreements can overemphasize harmful low-low tokens. DIAL-OPD incorporates absolute probability scale through the logarithmic mean to address this blind spot. Selection analyses suggest that moderate probability weighting suppresses low-low tokens while preserving useful disagreements, whereas excessive weighting can discard valuable supervision. Across four teacher–student pairs and seven benchmarks, it outperforms full-token OPD with fewer tokens and existing selection methods at matched retention ratios. Effective supervision allocation can even outweigh teacher scaling, underscoring the importance of token selection.

AI USE STATEMENT

We used generative AI tools to assist with method implementation and to polish the manuscript. The authors manually reviewed the AI-assisted code to verify implementation correctness and checked the manuscript text for accuracy. The authors take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies token selection for language-model distillation using existing model checkpoints and mathematical reasoning datasets. Its evaluation concerns benchmark accuracy rather than deployment safety. Distillation may transfer errors and biases from the teacher, and probability-based token selection does not by itself establish the correctness or safety of the resulting model. Applications beyond the evaluated setting require separate assessment of reliability, bias, and potential misuse, as well as adherence to the terms governing the underlying models and data.

REPRODUCIBILITY STATEMENT

Section 4 specifies the selection score, per-response retention rule, and masked training objective. Section 5 identifies the training data, model sizes, benchmarks, and baselines. Appendix B provides checkpoint identities, training and evaluation settings, and details of the supervision-density experiments, baseline adaptations, and low-low token analysis. Appendix C derives the sampled-token OPD update and states the scope of its unbiasedness claim, while Appendix E provides proofs of both propositions under their stated assumptions. Appendix D reports the probability-weighting ablation across all four teacher–student pairs.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes, 2024. URL https://arxiv.org/abs/2306.13649.

Rajendra Bhatia. The logarithmic mean. *Resonance*, 13(6):583–594, 2008. doi: 10.1007/s12045-008-0063-4. URL https://repository.ias.ac.in/2562/.

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.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv:2501.12948, 2025. URL https://arxiv.org/abs/2501.12948.

DeepSeek-AI. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026. URL https://arxiv.org/abs/2609.19969.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms, 2026. URL https://arxiv.org/abs/2605.00674.

Shiyuan Feng, Huan ang Gao, Haohan Chi, Hanlin Wu, Zhilong Zhang, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Weak-to-strong generalization via direct on-policy distillation, 2026. URL https://arxiv.org/abs/2607.05394.

Zixuan Fu, Bingxiang He, Yuxin Zuo, Haohuan Huang, Jinqian Zhang, Ruhang Xiao, Cheng Qian, Qinyu Luo, Huan ang Gao, Yudong Wang, Zhiyuan Liu, Ning Ding, and Chaojun Xiao. Rethinking on-policy distillation of large language models ii: One training example, 2026. URL https://arxiv.org/abs/2609.04172.

GLM-5 Team. GLM-5: From vibe coding to agentic engineering. arXiv:2602.15763, February 2026. URL https://arxiv.org/abs/2602.15763.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models, 2026. URL https://arxiv.org/abs/2306.08543.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng 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, 2024. URL https://arxiv.org/abs/2402.14008.

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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020. URL https://arxiv.org/abs/1904.09751.

ZhiYan Hou, Xinyu Tang, Hongyan An, Jianjin Zhang, Weizhen Wang, Yunyun Han, Gengsheng Li, Xiangzhao Hao, Haiyun Guo, Wenbin Hu, Jinqiao Wang, and Yafeng Deng. Dash: Divergence-adaptive supervision horizons for on-policy self-distillation of reasoning models, 2026. URL https://arxiv.org/abs/2608.06243.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation, 2026. URL https://arxiv.org/abs/2601.20802.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation, 2026. URL https://arxiv.org/abs/2601.07155.

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, 2026. URL https://arxiv.org/abs/2605.06387.

Yinuo Jiang, Yongjie Ye, Zhou Tao, Xiang Zhuang, Qiang Zhang, Huajun Chen, and Tiankai Li. When teachers mislead: Spurious-signal-aware on-policy distillation, 2026. URL https://arxiv.org/abs/2608.03632.

Yuxuan Jiang and Francis Ferraro. Bridging reasoning trajectories in on-policy distillation via near-future guidance, 2026. URL https://arxiv.org/abs/2606.00305.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models, 2026. URL https://arxiv.org/abs/2603.07079.

Kimi Team. Kimi K3: Open frontier intelligence. arXiv:2607.24653, July 2026. URL https://arxiv.org/abs/2607.24653.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation, 2026. URL https://arxiv.org/abs/2603.11137.

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

Enhan Li, Junhao He, and Hongyang Du. Crop: Task relevance via counterfactuals for selective on-policy distillation, 2026a. URL https://arxiv.org/abs/2608.13387.

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

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Zhishuai Liu, Xingzi Xu, Mehmet Saygin Seyfioglu, Pan Xu, and Karim Bouyarmane. Extremely sparse supervision incentivizes reasoning ability, 2026. URL https://arxiv.org/abs/2609.04565.

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.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2, 2025. Notion Blog.

Qwen Team. Qwen3 technical report. arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

Qitai Tan, Zefang Zong, Mo Li, Yipeng Shi, Yang Li, and Peng Chen. Atod: Annealed turn-aware on-policy distillation for multi-turn agentic tasks, 2026. URL https://arxiv.org/abs/2606.27814.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation, 2026. URL https://arxiv.org/abs/2605.26844.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Xiaomi LLM-Core Team. MiMo-V2-Flash technical report, 2026. URL https://arxiv.org/abs/2601.02780.

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

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

Yongkang Yang, Zhezheng Hao, Hong Zhang, Yi Liu, Xiankun Lin, Wence Ji, Fanjunduo Wei, Jiarui Yu, Qiang Lin, Xiaoyun Liang, and Hande Dong. Matching supervision to the student's learning capacity: A unified framework for on-policy self-distillation, 2026b. URL https://arxiv.org/abs/2608.08176.

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

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D. Lyng, Sanjit Singh Batra, and Robert E. Tillman. Fast and effective on-policy distillation from reasoning prefixes, 2026. URL https://arxiv.org/abs/2602.15260.

Anhao Zhao, Junlong Tong, Yingqi Fan, Ping Nie, Wenjie Li, and Xiaoyu Shen. Poweropd: Stabilizing on-policy distillation with bounded power transformation, 2026a. URL https://arxiv.org/abs/2606.17199.

Qingfei Zhao, Huan Song, Shuyu Tian, Jiawei Shao, and Xuelong Li. Prefix-guided on-policy distillation: Mining golden trajectories from rollouts, 2026b. URL https://arxiv.org/abs/2606.21994.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models, 2026c. URL https://arxiv.org/abs/2601.18734.

Zilin Zhu, Chengxing Xie, Xin Lv, and slime Contributors. slime: An llm post-training framework for rl scaling. https://github.com/THUDM/slime, 2025. GitHub repository. Corresponding author: Xin Lv.

APPENDIX

A RELATED WORK

On-policy distillation. OPD trains a student on its own generated responses using teacher supervision, reducing the mismatch between training and inference. GKD (Agarwal et al., 2024) studies flexible divergence objectives on student-generated sequences, MiniLLM (Gu et al., 2026) develops reverse-KL optimization, and Lu & Lab (2025) present a practical sampled-token training recipe. Subsequent developments address two complementary questions. **(i)** *Constructing teacher supervision.* Qwen3 (Qwen Team, 2025) transfers capabilities from larger teachers through off-policy initialization followed by on-policy distillation; MiMo-V2-Flash (Xiaomi LLM-Core Team, 2026) integrates domain-specialized teachers; GLM-5 (GLM-5 Team, 2026) recovers capabilities from earlier training checkpoints; and Kimi K3 (Kimi Team, 2026) consolidates experts specialized by domain and reasoning effort. Self-distillation instead constructs a teacher from the model itself: OPSD (Zhao et al., 2026c) provides it with solutions unavailable to the student, while SDPO (Hübotter et al., 2026) conditions it on textual feedback or successful attempts. **(ii)** *Optimizing the distillation objective.* ExOPD (Yang et al., 2026a) separates reward strength from KL regularization, while Direct-OPD (Feng et al., 2026) derives rewards from the teacher's policy change during reinforcement learning. PowerOPD (Zhao et al., 2026a) bounds rewards through power transformations, and Veto (Jang et al., 2026) stabilizes alignment through a geometric target distribution combining teacher and student probabilities. AOPD (Jia et al., 2026) uses policy-gradient reinforcement for positive advantages and local distribution matching otherwise; EOPD (Jin et al., 2026) adds forward-KL supervision where teacher entropy is high. Recent evidence also highlights a bottleneck in supervision utilization: Fu et al. (2026) show that training on a single query recovers much of full-data OPD's improvement while teacher–student alignment remains slow.

Selective supervision for OPD. Selective supervision addresses which parts of generated responses should contribute to learning by selecting, weighting, or restructuring teacher supervision. Two complementary approaches consider individual tokens and the structure of the surrounding trajectory. **(i)** *Assessing token learning value.* TIP (Xu et al., 2026) combines student uncertainty with teacher–student divergence, while REOPOLD (Ko et al., 2026) couples entropy-guided token sampling with reward clipping. SuRe (Shao et al., 2026) uses bounded surprise-based weighting to emphasize tokens assigned low probability by the student, and TA-OPD (Wang et al., 2026) assesses disagreement together with teacher support for the student's likely alternatives. Beyond distributional statistics, SA-OPD (Jiang et al., 2026) filters high-impact supervision weakly dependent on the question, whereas CROP (Li et al., 2026a) measures sensitivity to changed task conditions relative to meaning-preserving rewrites. USD (Yang et al., 2026b) jointly adjusts token weights and the information supplied to a self-teacher according to student learning capacity. TIDE (Yu et al., 2026) applies different corrections to tokens overestimated by the student and alternatives favored by the teacher. **(ii)** *Using trajectory structure.* Prefix OPD (Zhang et al., 2026) trains on truncated reasoning prefixes, and PG-OPD (Zhao et al., 2026b) uses teacher–student agreement on prefixes to decide which rollouts to extend. FiRe-OPD (Li et al., 2026b) filters trajectories by teacher likelihood before weighting their tokens, while TOPD (Jiang & Ferraro, 2026) uses near-future continuations to guide corrections. DASH (Hou et al., 2026) aggregates divergence across successive positions to construct token weights; ATOD (Tan et al., 2026) weights interaction turns while gradually shifting from distillation to reinforcement learning. Evidence from extremely sparse OPD further shows that supervising one or two selected tokens per trajectory can match or exceed dense training (Liu et al., 2026). Building on these findings, we identify how relative disagreement can overemphasize tokens assigned low probability by both models and test their training value through targeted replacement. DIAL-OPD uses the logarithmic mean to connect relative disagreement with absolute probability scale, selecting a fixed fraction of tokens within each response while preserving the original signed OPD reward.

B EXPERIMENTAL DETAILS

Models. Students are initialized from Qwen3-0.6B-Base and Qwen3-1.7B-Base, while teachers use the instruction-tuned Qwen3-4B and Qwen3-8B checkpoints. The equal-size comparison in

Section 3.1 additionally uses Qwen3-4B-Base as the student. We omit the "-Base" suffix in the main text for brevity.

Training details. All methods use AdamW with BF16 mixed precision, learning rate 5×10^{-7}, and batch size 32. Each rollout contains up to 1,024 response tokens. Training rollouts use temperature 1.0, top-p 0.95, and top-k 20. Prompt, padding, and terminal EOS positions are excluded from the training loss for all methods.

Evaluation protocol. For the main comparisons, we sample $k = 4$ responses per question on GSM8K, MATH-500, and OlympiadBench; $k = 8$ on Minerva Math; and $k = 16$ on AMC23, AIME24, and AIME25. Sampling uses temperature 0.6 and top-p 0.95. Tables abbreviate average correctness, success in at least one sampled response, and mean response length as Avg, Pass, and Len, respectively. Mean denotes the unweighted average of Avg across the seven benchmarks.

B.1 Supervision Density Experiment

Training follows the configuration in Section 5. For each response, we exclude prompt, padding, and terminal EOS positions, leaving N eligible tokens. We uniformly sample $K = \max(1, \mathrm{round}(\rho N))$ positions without replacement, where ρ is the retention ratio. Selection masks are refreshed at every training step. Only selected positions contribute to the original signed OPD loss, averaged over all selected tokens in the batch; unselected tokens remain in the context. We evaluate on MATH-500 before training and every 50 steps, sampling one response per question with temperature 0.6, top-p 0.95, and a maximum response length of 4,096 tokens. The same-size experiment repeats the retention-ratio sweep with Qwen3-4B-Base as the student and Qwen3-4B as the teacher.

B.2 Baseline Implementations

We adapt token-selection baselines to matched per-response retention ratios. For SA-OPD, we select tokens by input-dependence ranking instead of its original joint filtering conditions. For TA-OPD, teacher support measures the probability mass that the teacher assigns to the student's top-K candidate tokens (Wang et al., 2026); we retain batch-level normalization and perform selection within each response. Prefix OPD selects the initial fraction of each complete rollout rather than stopping generation early. For CC-OPD, the teacher's pre-RL checkpoint is unavailable, so we use a smaller instruction-tuned model from the same family as the reference and select tokens with the largest absolute teacher–reference probability gaps.

B.3 Low-Low Token Analysis

For the concentration analysis in Section 3.3, we pool selected-token counts and absolute reward sums across responses before computing low-low shares, rather than averaging per-response percentages. The token share divides the selected low-low count by the total selected count; the reward share divides their summed absolute rewards by the absolute reward sum over all selected tokens. Replacement candidates are ranked by absolute reward, while the OPD loss retains the original reward signs. Both replacement strategies use training and evaluation settings matched to the corresponding Reward-Top baseline at each retention ratio.

C Derivation of the Sampled-Token OPD Update

We derive the update in Eq. (2) by differentiating the conditional reverse KL while holding the rollout context distribution fixed. The derivation also establishes the unbiasedness of its Monte Carlo estimator and explains the stop-gradient treatment of the token-level reward.

Fixed-context objective. Let $\bar{\theta}$ denote the student parameters used to collect rollouts, and let θ denote the parameters with respect to which we differentiate. For $x \sim \mathcal{D}$ and $y \sim \pi_{\bar{\theta}}(\cdot \mid x)$, the

context preceding token y_t is $c_t = (x, y_{<t})$. With the teacher policy fixed, define

$$\mathcal{L}_{\mathrm{ctx}}(\theta; \bar{\theta}) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_{\bar{\theta}}(\cdot|x)} \left[\sum_{t=1}^{|y|} D_{\mathrm{KL}}(\pi_\theta(\cdot \mid c_t) \,\|\, \pi_T(\cdot \mid c_t)) \right]. \tag{C.1}$$

Here, $\bar{\theta}$ is held fixed during differentiation: the contexts come from the rollout policy, while the conditional student distributions at those contexts depend on θ. We evaluate the resulting gradient at $\theta = \bar{\theta}$. We assume differentiable, strictly positive student and teacher probabilities over a finite vocabulary, bounded rollout lengths, and the regularity conditions required to interchange differentiation and expectation.

From reverse KL to a policy-gradient update. For a fixed context c, write $p_\theta(v) = \pi_\theta(v \mid c)$ and $q(v) = \pi_T(v \mid c)$. Differentiating the vocabulary-wide reverse KL gives

$$\begin{aligned} \nabla_\theta D_{\mathrm{KL}}(p_\theta \| q) &= \nabla_\theta \sum_{v\in\mathcal{V}} p_\theta(v) \log \frac{p_\theta(v)}{q(v)} \\ &= \sum_{v\in\mathcal{V}} \left(\log \frac{p_\theta(v)}{q(v)} + 1 \right) \nabla_\theta p_\theta(v) \\ &= \sum_{v\in\mathcal{V}} p_\theta(v) \log \frac{p_\theta(v)}{q(v)} \nabla_\theta \log p_\theta(v). \end{aligned} \tag{C.2}$$

The last equality uses $\nabla_\theta p_\theta(v) = p_\theta(v) \nabla_\theta \log p_\theta(v)$ and the normalization identity

$$\sum_{v\in\mathcal{V}} \nabla_\theta p_\theta(v) = \nabla_\theta \sum_{v\in\mathcal{V}} p_\theta(v) = \nabla_\theta 1 = 0. \tag{C.3}$$

Consequently, the negative gradient is an expectation of the teacher–student log-ratio multiplied by the student's log-probability gradient. Averaging this identity over the fixed rollout context distribution and evaluating at the rollout policy yields

$$\begin{aligned} g_{\mathrm{OPD}}(\bar{\theta}) &:= -\nabla_\theta \mathcal{L}_{\mathrm{ctx}}(\theta; \bar{\theta})\big|_{\theta=\bar{\theta}} \\ &= \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_{\bar{\theta}}(\cdot|x)} \left[\sum_{t=1}^{|y|} \underbrace{\log \frac{\pi_T(y_t \mid c_t)}{\pi_{\bar{\theta}}(y_t \mid c_t)}}_{r_t^{\mathrm{OPD}}} \nabla_\theta \log \pi_\theta(y_t \mid c_t)|_{\theta=\bar{\theta}} \right]. \end{aligned} \tag{C.4}$$

This is the REINFORCE policy-gradient form, with the log-ratio serving as the token-level reward (Williams, 1992; Sutton et al., 1999). Removing the distinction between the differentiation variable and the rollout parameters after evaluation recovers the notation of Eq. (2).

Monte Carlo estimation and stop-gradient implementation. Given B independent prompt–response pairs sampled from $\mathcal{D}$ and $\pi_{\bar{\theta}}$, let $c_{i,t}$ denote the context of token $y_{i,t}$, and let $r_{i,t}^{\mathrm{OPD}}$ be its log-ratio reward evaluated at $\bar{\theta}$. The sampled update is

$$\widehat{g}_B = \frac{1}{B} \sum_{i=1}^{B} \sum_{t=1}^{|y_i|} r_{i,t}^{\mathrm{OPD}} \nabla_\theta \log \pi_\theta(y_{i,t} \mid c_{i,t})|_{\theta=\bar{\theta}}, \qquad \mathbb{E}[\widehat{g}_B] = g_{\mathrm{OPD}}(\bar{\theta}). \tag{C.5}$$

Each sampled token supplies one Monte Carlo sample of the conditional expectation in Eq. (C.2). Tokens within a response need not be independent: conditioned on its prefix, each token is sampled from the student distribution, and unbiasedness follows from conditional expectation and linearity. In implementation, the same update is obtained by minimizing

$$\widehat{\ell}_{\mathrm{OPD}}(\theta; \bar{\theta}) = -\frac{1}{B} \sum_{i=1}^{B} \sum_{t=1}^{|y_i|} \mathrm{sg}\big[r_{i,t}^{\mathrm{OPD}}\big] \log \pi_\theta(y_{i,t} \mid c_{i,t}), \tag{C.6}$$

Table 3: Probability-weighting ablation across four teacher–student pairs and seven benchmarks. Each retention ratio compares reward-only selection ($\beta = 0$) with $\beta = 1$; **bold** marks the higher accuracy within each pair, including ties.

Keep	β	GSM8K			MATH500			AMC23			AIME24			AIME25			Minerva			Olympiad			Mean
		Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg	Pass	Len	Avg
Teacher: Qwen3-4B → Student: Qwen3-0.6B-Base																							
20%	0	57.3	79.2	535	**36.4**	**52.4**	1786	**15.9**	**57.5**	5254	**1.7**	**10.0**	7297	0.8	6.7	7685	**8.5**	**20.6**	1593	12.9	**24.6**	5419	19.1
	1	**67.0**	**81.3**	1072	34.2	47.2	2609	**15.9**	45.0	6554	0.4	6.7	7985	**1.5**	**10.0**	8016	7.6	16.9	2617	**13.1**	24.2	6658	**19.9**
40%	0	56.1	81.3	507	34.1	52.4	1752	15.8	**52.5**	4996	1.9	3.3	7180	0.4	**6.7**	7516	8.0	21.0	1504	12.4	24.5	5495	18.4
	1	**69.1**	**84.7**	578	**44.9**	**61.6**	1714	**25.5**	50.0	4774	**2.5**	**13.3**	7001	**1.2**	**6.7**	7179	**10.7**	**22.1**	1484	**18.8**	**33.5**	4705	**24.7**
Teacher: Qwen3-4B → Student: Qwen3-1.7B-Base																							
20%	0	69.4	90.8	354	53.1	**71.6**	1149	30.0	**75.0**	3476	6.0	20.0	5589	4.2	**16.7**	5478	14.7	29.4	943	24.0	39.2	3568	28.8
	1	**82.3**	**91.8**	404	**57.8**	70.2	1445	**37.8**	**75.0**	4276	**8.8**	**26.7**	6287	**4.8**	**16.7**	6170	**17.9**	**32.7**	1216	**29.5**	**43.2**	4222	**34.1**
40%	0	67.2	91.9	363	51.6	72.0	1124	27.7	**82.5**	3272	8.3	**30.0**	5530	**5.4**	13.3	5303	14.3	29.0	872	24.3	41.1	3393	28.4
	1	**83.2**	**92.8**	344	**61.5**	**73.0**	1223	**38.8**	75.0	3457	**9.8**	26.7	5992	5.2	**16.7**	5951	**19.4**	**30.9**	992	**29.7**	**44.5**	3789	**35.4**
Teacher: Qwen3-8B → Student: Qwen3-0.6B-Base																							
20%	0	58.2	80.4	471	39.2	57.4	1568	15.8	57.5	4527	1.5	3.3	6626	0.2	3.3	7127	8.0	20.6	1402	13.8	25.8	5025	19.5
	1	**63.8**	**83.2**	465	**42.3**	**59.6**	1633	**21.2**	**62.5**	4372	**1.7**	**10.0**	6566	**2.1**	**10.0**	6583	**9.7**	**23.2**	1265	**16.0**	**29.1**	4494	**22.4**
40%	0	58.8	80.5	488	40.9	57.6	1562	16.9	57.5	4697	1.9	**13.3**	6521	**1.2**	**10.0**	7080	9.9	**22.8**	1276	14.4	27.2	4967	20.6
	1	**66.6**	**84.6**	436	**45.6**	**61.4**	1482	**27.3**	**62.5**	3872	**3.1**	**13.3**	6432	**1.2**	6.7	6589	**10.6**	22.1	1197	**17.7**	**29.7**	4392	**24.6**
Teacher: Qwen3-8B → Student: Qwen3-1.7B-Base																							
20%	0	74.8	91.7	345	58.2	73.8	1116	31.4	**72.5**	3309	7.5	**16.7**	6003	**6.5**	**16.7**	5179	16.3	30.1	931	27.0	42.0	3494	31.7
	1	**80.6**	**92.9**	333	**60.2**	**74.4**	1199	**35.8**	70.0	3404	**7.9**	**16.7**	5791	5.4	**16.7**	5339	**18.3**	**31.2**	967	**28.2**	**42.3**	3570	**33.8**
40%	0	80.0	92.4	347	58.1	71.8	1183	33.1	80.0	3262	7.3	**20.0**	5828	5.6	20.0	5390	17.5	**32.7**	991	26.5	40.2	3693	32.6
	1	**82.4**	**93.5**	323	**59.7**	**73.8**	1173	**38.0**	**82.5**	3098	**8.8**	16.7	5764	**6.5**	**23.3**	5425	**18.0**	31.6	913	**28.9**	**43.5**	3533	**34.6**

where sg denotes the stop-gradient operator. By construction,

$$-\nabla_\theta \widehat{\ell}_{\mathrm{OPD}}(\theta;\bar{\theta})\Big|_{\theta=\bar{\theta}} = \widehat{g}_B. \tag{C.7}$$

The stop-gradient operation implements the coefficient obtained from the analytical KL gradient; it does not assume that the original log-ratio is independent of the student parameters. Its explicit parameter dependence has already been accounted for in Eq. (C.2), where the additional term cancels by normalization. Thus, Eq. (C.6) is a surrogate loss that reproduces the required gradient at the rollout policy, rather than a direct differentiation of the sampled log-ratio alone.

Relation to the full on-policy objective. At $\theta = \bar{\theta}$, the fixed-context objective in Eq. (C.1) has the same value as the objective in Eq. (1), but their derivatives need not coincide. Differentiating Eq. (1) through the entire trajectory distribution additionally accounts for how earlier token choices change the contexts encountered later in the response. Equation (2) instead uses the conditional reverse-KL gradients at the sampled contexts, following the immediate-token update used in on-policy distillation (Lu & Lab, 2025). The unbiasedness statement in Eq. (C.5) therefore concerns this fixed-context update direction. Contexts are nevertheless refreshed by new student rollouts during training; holding their distribution fixed during differentiation does not make the training procedure off-policy.

D Probability Weighting Ablation

To isolate the contribution of probability weighting, Table 3 compares DIAL-OPD at $\beta = 0$ and $\beta = 1$ under the same retention ratio across all four teacher–student pairs. The $\beta = 0$ setting corresponds to Reward-Top in Section 3.2. With a Qwen3-4B teacher and Qwen3-1.7B-Base student at 40% retention, $\beta = 1$ improves seven-benchmark mean accuracy by **6.96** percentage points and GSM8K Avg@4 by **15.96** points over $\beta = 0$. These results extend the learning-dynamics analysis in Section 6.1 to the full evaluation suite.

E PROOFS OF DIAL-OPD PROPERTIES

E.1 PROOF OF PROPOSITION 1

Fix $p, q \in (0, 1]$, $\beta \geq 0$, and $\varepsilon \in (0, 1]$. Scaling both probabilities by ε preserves their ratio while scaling their logarithmic mean by the same factor. From the integral definition in Equation 3,

$$\begin{aligned} L(\varepsilon p, \varepsilon q) &= \int_0^1 (\varepsilon p)^{1-u} (\varepsilon q)^u \, du \\ &= \int_0^1 \varepsilon^{1-u+u} p^{1-u} q^u \, du = \varepsilon L(p, q). \end{aligned} \tag{E.8}$$

Meanwhile, $|\log[(\varepsilon q)/(\varepsilon p)]| = |\log(q/p)|$. Substituting these identities into Equation 4 yields

$$\begin{aligned} s_\beta(\varepsilon p, \varepsilon q) &= L(\varepsilon p, \varepsilon q)^\beta \left| \log \frac{\varepsilon q}{\varepsilon p} \right| \\ &= [\varepsilon L(p, q)]^\beta \left| \log \frac{q}{p} \right| = \varepsilon^\beta s_\beta(p, q). \end{aligned} \tag{E.9}$$

This identity also holds when $p = q$, since both scores are zero. For $p \neq q$, the score is positive; hence $\varepsilon < 1$ leaves it unchanged at $\beta = 0$ and strictly reduces it at $\beta > 0$. Larger β produces a smaller scaling factor ε^β, establishing stronger suppression under the same proportional reduction in both probabilities.

E.2 PROOF OF PROPOSITION 2

Fix $a \in (0, 1]$ and $\beta \geq 0$, and let $b \to 0^+$. We prove the limit of the selection score from Equation 4 when one probability vanishes and the other remains fixed. Since $a > 0$, we may restrict attention to $0 < b < a$ without affecting the limit. Both probabilities are strictly positive throughout the derivation; the score need not be defined at $b = 0$.

Symmetry. The logarithmic mean and the absolute log-ratio are symmetric in their arguments. In particular,

$$L(a, b) = \frac{a - b}{\log a - \log b} = \frac{b - a}{\log b - \log a} = L(b, a), \qquad \left| \log \frac{a}{b} \right| = \left| \log \frac{b}{a} \right|. \tag{E.10}$$

It follows that $s_\beta(a, b) = s_\beta(b, a)$ for every positive a, b. Thus, the same limit applies whether the student probability or the teacher probability approaches zero, and it suffices to analyze $s_\beta(a, b)$.

Asymptotic form. For $0 < b < a$, both $a - b$ and $\log(a/b)$ are positive. Substituting the closed form of the logarithmic mean into the score gives the exact identity

$$\begin{aligned} s_\beta(a, b) &= L(a, b)^\beta \left| \log \frac{a}{b} \right| \\ &= \left(\frac{a - b}{\log(a/b)} \right)^\beta \log(a/b) \\ &= (a - b)^\beta [\log(a/b)]^{1-\beta}. \end{aligned} \tag{E.11}$$

As $b \to 0^+$, the factor $(a - b)^\beta$ converges to $a^\beta > 0$, while $\log(a/b) = \log a - \log b \to +\infty$. More precisely,

$$\frac{s_\beta(a, b)}{a^\beta [\log(a/b)]^{1-\beta}} = \left(1 - \frac{b}{a} \right)^\beta \longrightarrow 1, \qquad s_\beta(a, b) \sim a^\beta [\log(a/b)]^{1-\beta}. \tag{E.12}$$

Thus, the limiting behavior is determined by the exponent $1 - \beta$, since the remaining factor approaches a finite, strictly positive constant.

The three regimes. (i) If $0 \leq \beta < 1$, then $1 - \beta > 0$, so $[\log(a/b)]^{1-\beta} \to +\infty$ and hence $s_\beta(a, b) \to +\infty$. (ii) If $\beta = 1$, the logarithmic factor cancels exactly: $s_1(a, b) = a - b \to a$. (iii) If $\beta > 1$, then

$$s_\beta(a, b) = \frac{(a - b)^\beta}{[\log(a/b)]^{\beta-1}} \longrightarrow 0, \tag{E.13}$$

because the numerator converges to a^β while the denominator diverges. Combining these cases with symmetry establishes both limits in Proposition 2. In particular, the zero limit for $\beta > 1$ concerns arbitrarily small positive b; it does not imply that the score equals zero at any finite $b \in (0, a)$.