Asymmetric On-Policy Distillation: Bridging Exploitation and Imitation at the Token Level

Anonymous ACL submission

Abstract

On-policy distillation (OPD) trains a student on its own trajectories with token-level teacher feedback and often outperforms off-policy distillation and reinforcement learning with verifiable rewards. However, we find that its standard advantage-weighted policy gradient suffers from three structural weaknesses, including high-variance updates, vanishing gradients in zero-advantage regions, and exploration bottlenecks for models with limited capability. We therefore propose asymmetric on-policy distillation (AOPD), which replaces ineffective policy gradient with localized divergence minimization in non-positive advantage regions while preserving positive reinforcement learning. Experiments on mathematical reasoning benchmarks show that AOPD consistently outperforms standard OPD, with average gains of 4.09 / 8.34 under strong / weak initialization, respectively. AOPD also maintains higher policy entropy during training and better capability retention during sequential tool-use adaptation. Our source code is available at repository.

1 Introduction

The strong capabilities of large language models (LLMs) (Yang et al., 2025; Team et al., 2026b; DeepSeek-AI, 2026) come at a steep computational cost. To make LLMs more accessible, knowledge distillation (Hinton et al., 2014) offers a principled path to compress the capabilities into smaller models. However, this paradigm carries a well-known vulnerability in that the student encounters exposure bias (Bengio et al., 2015) after training exclusively on the off-policy trajectories.

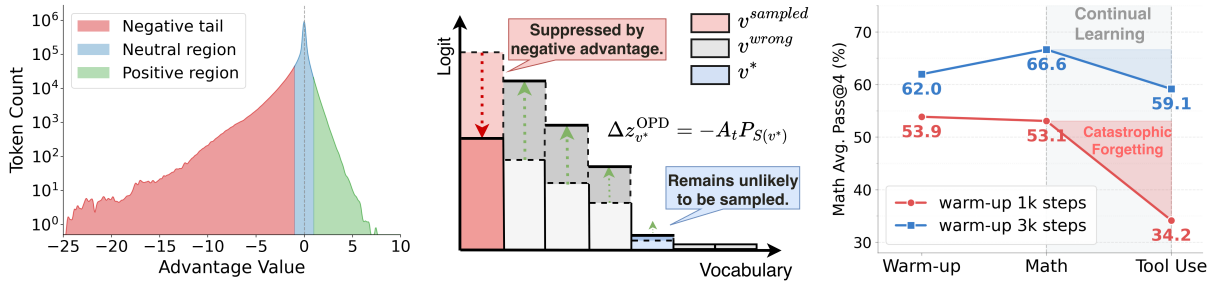
On-policy distillation (OPD) (Lu and Lab, 2025) resolves this mismatch by optimizing the student on its own trajectories through a reinforcement learning objective. Using token-level rewards derived from teacher–student log-probability differences, it achieves strong training efficiency and has become a key technique in post-training (Team et al., 2026a; GLM-5-Team et al., 2026).

However, we observe that this advantage formulation exhibits three limitations: (1) In the negative region, the advantage distribution exhibits a heavy tail with large variance, as shown in Figure 1a. As a result, negative tokens can induce highly variable updates, thereby destabilizing the optimization process. (2) A large fraction of sampled tokens concentrates near zero advantage. The policy gradient consequently produces an almost vanishing update. (3) OPD relies on the exploration mechanism of reinforcement learning for optimization. However, for models with limited initial capability, this process can easily degenerate into blind exploration.

To address the above limitations, we bridge the principles of reinforcement learning and supervised learning at the token level to propose asymmetric on-policy distillation (AOPD). Specifically, when a token receives a negative advantage, AOPD does not rely on self-exploration. Instead, it directly matches the teacher model's target distribution conditioned on the on-policy trajectory prefix via Kullback-Leibler (KL) divergence. This asymmetric design allows the student to retain the benefit of exploitation in reinforcement learning while receiving stronger correction at optimization bottlenecks.

We evaluate AOPD on competition-level mathematical reasoning benchmarks. Across model scales and warm-up settings, AOPD delivers stronger and more robust performance than existing baselines. The localized guidance also leads to better capability retention in continual learning. Our main contributions are summarized as follows:

- We analyze on-policy distillation in zero and negative advantage regions and show that the standard advantage-based update provides limited learning signal for effective correction.
- We propose asymmetric on-policy distillation, a token-level method that adaptively shifts



(a) Advantage distribution. (b) Exploration black hole. (c) OPD under insufficient warm-up.

Figure 1: Observation and analysis of on-policy distillation.

from policy gradient to localized distribution matching when the advantage signal carries limited information.

- Experiments on mathematical reasoning benchmarks show that AOPD consistently improves over baselines, especially under weak initialization, and better preserves prior capabilities during sequential tool-use adaptation.

2 Related Work

The foundation of knowledge distillation was established by Hinton et al. (2014), who showed that a compact student can learn from a larger teacher by matching its softened output logits. Classical off-policy distillation has achieved considerable success (Hsieh et al., 2023; Tian et al., 2025; Xu et al., 2024) in strong-to-weak distillation. For sequence generation, Kim and Rush (2016) extended this idea to Sequence-Level Knowledge Distillation (SeqKD), aligning the student with the teacher's sequence-level distribution. However, all these classical methods are trained off-policy on teacher-generated trajectories and therefore remain vulnerable to exposure bias (Bengio et al., 2015).

To mitigate this mismatch, recent work has shifted toward on-policy distillation, where the student learns on its own rollouts with teacher feedback. GKD (Agarwal et al., 2024) formalizes this as minimizing generalized f-divergences on student trajectories. Gu et al. (2024) and Lu and Lab (2025) formulate on-policy distillation as a reinforcement learning objective with dense token-level rewards derived from the teacher. Li et al. (2026) identify teacher-student compatibility and novel capability transfer as two key conditions for effective OPD. Beyond teacher alignment, ExOPD (Yang et al., 2026) extrapolates reward scaling to surpass the teacher's capabilities. SCOPE (Zheng et al., 2026) further uses perplexity-calibrated adaptive weighting to apply different supervision to correct and incorrect trajectories. Some work further explores multi-teacher extensions of OPD. MiMo-V2 (Team et al., 2026a) introduces Multi-Teacher On-Policy Distillation (MOPD), while DeepSeek-V4 (DeepSeek-AI, 2026) adopts a similar pipeline via full-vocabulary reverse KL divergence. CoPD (Gu et al., 2026) further argues that post-hoc MOPD may fail to fully internalize teacher capabilities, and proposes bidirectional OPD to co-evolve experts. Some work removes the reliance on external teachers by distilling from the model's own distributions. OPSD (Zhao et al., 2026) and its long-context extension OPSDL (Zhang et al., 2026b) leverage ground-truth feedback on self-generated traces, while SDFT (Shenfeld et al., 2026) mitigates catastrophic forgetting under the same paradigm.

3 Preliminaries

3.1 Formulation of OPD

OPD optimizes the reverse KL divergence with reinforcement learning based on the K1 estimator. We formalize the training procedure below. Given an input prompt x, let y denote a response sampled from the student policy. At step t, let $c_t \triangleq (x, y_{<t})$ denote the current context, and let $P_T(\cdot \mid c_t)$ and $P_S(\cdot \mid c_t)$ denote the teacher and student conditional distributions. The advantage of token $y_t \sim P_S(\cdot \mid c_t)$ is defined as:

$$A_t = \mathrm{sg}[\log P_T(y_t \mid c_t) - \log P_S(y_t \mid c_t)], \quad (1)$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator. The corresponding training loss is

$$\mathcal{L}_{\mathrm{OPD}} = -\mathbb{E}\left[\frac{1}{|y|}\sum_{t=1}^{|y|} A_t \log P_S(y_t \mid c_t)\right]. \quad (2)$$

Motivated by Zhu et al. (2025) and Tang et al. (2025), which highlight the distinct roles of positive and negative samples in reinforcement learning, we decompose OPD into positive and negative reinforcement at the token level. Let S^+ and S^- denote the sets of token positions with positive and negative advantages, respectively, and let y_t^+ and y_t^- denote the corresponding sampled tokens. The OPD loss can be written as

$$\mathcal{L}_{\text{OPD}} = \underbrace{-\mathbb{E}\left[\frac{1}{|y|}\sum_{t\in S^+} A_t \log P_S(y_t^+ \mid c_t)\right]}_{\mathcal{L}_{\text{Pos}}} \underbrace{-\mathbb{E}\left[\frac{1}{|y|}\sum_{t\in S^-} A_t \log P_S(y_t^- \mid c_t)\right]}_{\mathcal{L}_{\text{Neg}}}. \quad (3)$$

Here, $\mathcal{L}_{\text{Pos}}$ corresponds to positive reinforcement, which reinforces teacher-favored sampled tokens. In contrast, $\mathcal{L}_{\text{Neg}}$ corresponds to negative reinforcement, which suppresses teacher-disfavored sampled tokens and promotes exploration.

3.2 Limitations of OPD

Despite its intuitive formulation, standard OPD remains subject to several challenges in practice. Through theoretical analysis and experimental validation, we identify three characteristic limitations.

Heavy Tails in Negative Advantages. Figure 1a shows that the negative side of the advantage distribution exhibits a substantially broader tail. The extreme variance originates from the logarithmic nature of the advantage formulation. The difference between the log probability of the teacher and the student amplifies exponentially as the probability approaches zero. Because the update magnitude is directly scaled by this unbounded scalar, the mechanism artificially inflates gradient variance and renders the overall optimization fragile.

Vanishing Updates at Zero Advantages. A second limitation arises in the neutral regime. As shown in Figure 1a, the vast majority of sampled tokens cluster around an advantage of zero. In this regime, the teacher remains able to offer fine-grained preferences over alternative reasoning paths. However, OPD only receives a zero signal, producing an almost vanishing update.

Exploration Black Hole. A fundamental optimization problem arises in negative reinforcement, where the update of OPD is restricted in both direction and magnitude. To understand this behavior, we examine the OPD update at the logit level. Let z_v denote the student logit for token v under context c_t, so that $P_S(v \mid c_t) = \text{softmax}(z)_v$. By the softmax Jacobian, the update for any token v is given by:

$$\Delta z_v^{\text{OPD}} \propto A_t\big(\mathbb{I}(v = y_t) - P_S(v \mid c_t)\big). \quad (4)$$

When $A_t < 0$, the sampled token is penalized, and for every unsampled token $v \neq y_t$, the update becomes

$$\Delta z_v^{\text{OPD}} \propto -A_t P_S(v \mid c_t) > 0. \quad (5)$$

Hence, the released probability mass is redistributed strictly according to the current prior of the student. As illustrated in Figure 1b, for an essential token v^* with negligible prior probability $P_S(v^* \mid c_t) \approx 0$, the corrective update is correspondingly negligible. OPD suppresses the sampled mistake without effectively promoting the correct alternative, trapping the model in an exploration black hole. This pathology is more severe for models with weaker foundational capabilities, as essential tokens are more likely to reside in the low-probability tail of the student prior. Figure 1c illustrates this effect by comparing OPD models with different steps of SFT warm-up. The model with fewer warm-up steps makes exploration black holes more likely and thereby undermines both reasoning improvement and ability retention.

These observations suggest that tokens in non-positive regions should not be optimized in the same way as positive reinforcement. The learning signal in these regions is insufficient and therefore calls for an asymmetric training rule.

4 Asymmetric On-Policy Distillation

To systematically resolve the structural vulnerability of standard on-policy distillation, we propose asymmetric on-policy distillation (AOPD). As illustrated in Figure 2, AOPD dynamically modulates the learning paradigm at the token level. It shifts between reinforcement learning for exploitation and soft-label supervised learning for imitation based on the student's local capability.

4.1 The AOPD Framework

We now present a formal definition of AOPD. We compute the token-level advantage A_t using Eq. 1

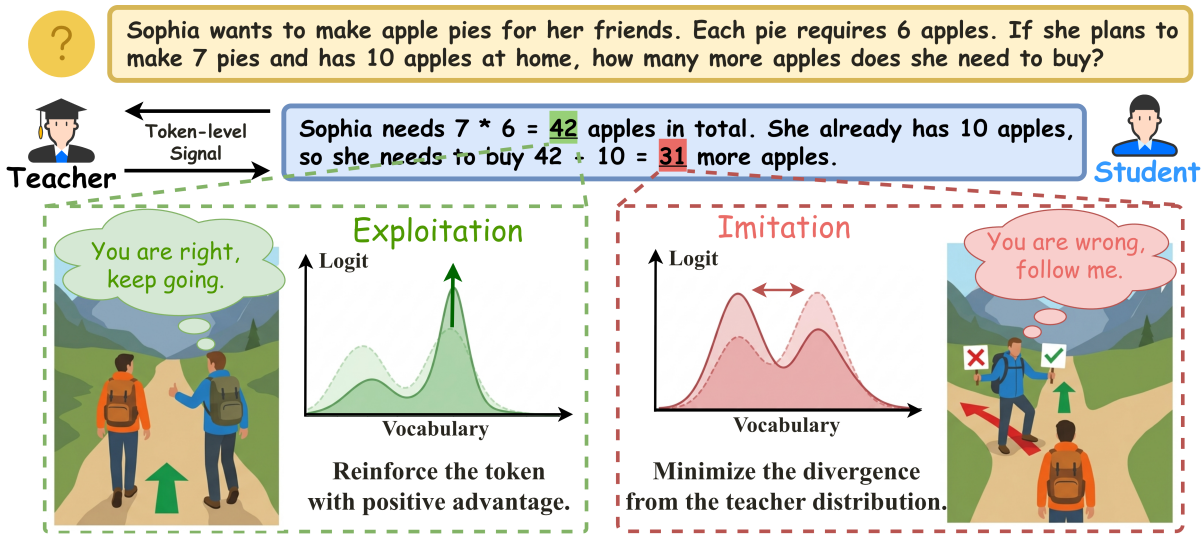


Figure 2: Overview of AOPD. Left (exploitation): When the student's reasoning aligns with the teacher, AOPD reinforces successful exploration. Right (imitation): When the student encounters an exploration bottleneck, AOPD actively switches to directed teacher guidance by minimizing the distributional divergence.

as in standard OPD. As discussed in Section 3.2, the main optimization difficulties of OPD are concentrated in negative and zero-advantage regions. We therefore replace the update in non-positive regions with forward KL minimization computed on the teacher-defined support. For each position t, we define the teacher support as

$$S_t = \mathrm{TopK}(P_T(\cdot \mid c_t), K)\,, \tag{6}$$

where S_t contains the K highest-probability tokens under the teacher distribution. On this support, the forward KL guidance loss is instantiated as

$$\mathcal{L}_t^{\mathrm{FKL}} = \frac{1}{K} \sum_{v \in S_t} P_T(v \mid c_t) \Big(\log P_T(v \mid c_t) - \log P_S(v \mid c_t) \Big). \tag{7}$$

Accordingly, the AOPD objective is defined as

$$\mathcal{L}_{\mathrm{AOPD}} = \mathbb{E}\left[\frac{1}{|y|} \sum_{t \in S^{\leq 0}} \mathcal{L}_t^{\mathrm{FKL}} \right] + \mathcal{L}_{\mathrm{Pos}}, \tag{8}$$

where $S^{\leq 0}$ denotes the non-positive token set.

AOPD uses zero advantage as the default intervention boundary. To define the intervention location in a more general form, we further introduce a threshold parameter τ. Considering the logarithmic nature of the standard advantage formulation and its unbounded variance, we use the bounded probability difference to determine whether to trigger teacher intervention. Specifically, we define the token-level mask as:

$$G_t = \mathbb{I}(P_T(y_t \mid c_t) - P_S(y_t \mid c_t) \leq \tau)\,, \tag{9}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The generalized AOPD objective is then defined as:

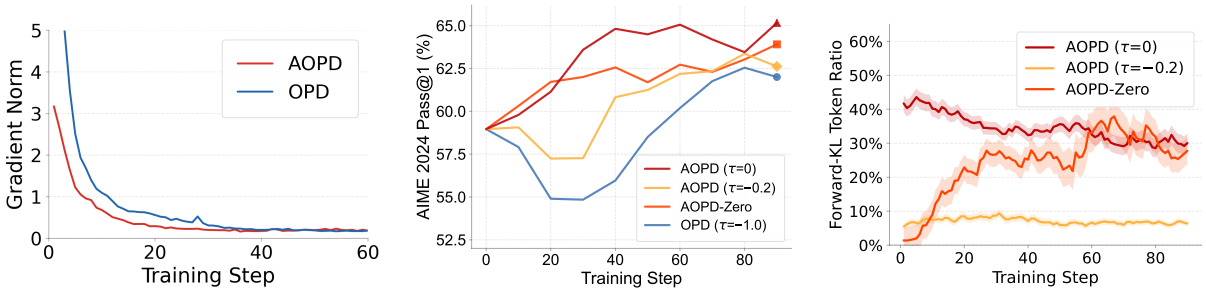
$$\mathcal{L}_{\mathrm{AOPD}} = \mathbb{E}\left[\frac{1}{|y|} \sum_{t=1}^{|y|} \Big(G_t \mathcal{L}_t^{\mathrm{FKL}} + (1 - G_t) \mathcal{L}_t^{\mathrm{OPD}} \Big) \right]. \tag{10}$$

Setting $\tau = -1$ disables intervention for all tokens and reduces AOPD to OPD, while setting $\tau = 1$ applies global supervised distribution matching and recovers a GKD objective. In our method, we set $\tau = 0$, so that intervention is applied exactly to the non-positive regime. The effects of alternative intervention locations and different values of τ are further analyzed in Section 4.3 and Appendix D.3.

4.2 Truncated Forward-KL Guidance

In principle, the intervention objective should minimize the divergence between the teacher and the student over the full vocabulary. The divergence family can be written in the general form of Jensen–Shannon divergence (JSD) as

$$\mathrm{JSD}_\beta(P_T \parallel P_S) = \beta D_{\mathrm{KL}}(P_T \parallel P_M) + (1 - \beta) D_{\mathrm{KL}}(P_S \parallel P_M), \tag{11}$$



(a) Gradient norm during training.

(b) AIME 2024 Pass@1 training curves under different intervention strategies.

(c) Ratio of tokens receiving forward KL guidance during training.

Figure 3: Training dynamics under different divergence-guidance strategies.

where $P_M = \beta P_T + (1-\beta)P_S$ and $\beta \in [0,1]$. Here, $\beta = 1$ corresponds to forward KL, while $\beta = 0$ corresponds to reverse KL.

However, evaluating such a divergence at every intervened position is computationally prohibitive in large-vocabulary LLM training. We therefore instantiate the guidance term on the teacher-selected top-K support in Eq. 6. Under top-K truncation, the learning objective becomes a correction defined on a teacher-selected domain. In this setting, the teacher specifies both the candidate tokens and the reference distribution within that domain. Therefore, forward KL is the natural choice because it preserves the same teacher-conditioned measure in both support construction and loss weighting. In contrast, reverse KL would reweight the update according to the student's current distribution on a domain selected by the teacher. The detailed gradient analysis is provided in Appendix B.1.

Empirical evidence in Section 5.4.1 further favors placing more emphasis on the teacher distribution. We also find that reverse KL induces anomalous training behaviors in Appendix B.2. Combined with the support-consistency argument above, these results motivate our choice of $\beta = 1$.

4.3 A Unified Solution to the Bottlenecks

We now explain why the single confidence-based rule in AOPD is sufficient to unify the solutions for the three optimization bottlenecks identified in Section 3.2. By setting $\tau = 0$, the core mechanism is to switch from policy gradient to direct distribution matching exactly where $A_t \leq 0$.

(1) Eliminating high-variance negative advantages. As analyzed in Section 3.2, standard OPD suffers from high-variance gradient updates when $A_t < 0$. AOPD eliminates this instability by halting the policy gradient update in these regions and replacing it with divergence guidance. For the forward KL case, the logit-level correction is instead governed by the bounded distribution gap:

$$\Delta z_v^{\text{AOPD}} \propto P_T(v \mid c_t) - P_S(v \mid c_t). \qquad (12)$$

Constrained by the probability simplex, this signal naturally caps the maximum update magnitude. Figure 3a confirms this effect, showing that AOPD maintains a substantially smaller gradient norm than OPD in early stage of training.

(2) Recovering gradients in zero-advantage regions. In zero-advantage regions where the original gradient nearly vanishes, AOPD still aligns the student toward alternative reasoning paths. Consequently, the student can still receive meaningful corrections on the teacher-defined support. Figure 3b directly supports this by evaluating AOPD-Zero, a variant that applies KL-divergence intervention only to tokens whose advantages are zero. Compared with standard OPD, this targeted replacement already yields clear gains, confirming that neutral regions contain useful teacher signals that are discarded by the policy gradient method.

(3) Escaping exploration bottlenecks through teacher guidance. When trapped in an exploration black hole, OPD merely suppresses the sampled mistake without promoting the correct, unsampled token. AOPD resolves this by explicitly injecting the missing directional signal. As shown in Eq. 12, any token favored by the teacher but underestimated by the student immediately receives a positive correction proportional to their probability gap. By directly matching the teacher's target distribution on the on-policy prefix, AOPD repairs optimization bottlenecks instantly rather than waiting for the student to discover the correct token. This is further evidenced in Figure 3b by the $\tau = -0.2$

Table 1: Main results across multiple math reasoning benchmarks. Baseline capabilities before on-policy distillation are given by the SFT Checkpoint. We compare AOPD against SeqKD, OPD, GKD and ExOPD. **Bold** indicates the best result within each configuration block, while underlined values denote the second best.

Method	AIME 2024		AIME 2025		HMMT 2025 (Feb)		Average	
	Pass@1	*Pass@4*	*Pass@1*	*Pass@4*	*Pass@1*	*Pass@4*	*Pass@1*	*Pass@4*
Qwen3-4B-Base								
– *3k SFT Warm-up*								
SFT Checkpoint	52.29	71.90	46.25	64.82	27.08	38.97	41.87	58.56
+ SeqKD	42.50	65.23	40.00	60.09	24.58	38.01	35.69	54.44
+ OPD	57.08	**76.69**	51.88	68.12	**32.92**	**47.65**	47.29	**64.15**
+ ExOPD	59.17	73.99	51.25	66.59	32.08	47.30	47.50	62.63
+ GKD	58.13	76.61	50.42	69.12	29.79	45.39	46.11	63.71
+ **AOPD**	**61.04**	75.08	**54.37**	**69.73**	32.08	47.40	**49.16**	64.07
Qwen3-8B-Base								
– *1k SFT Warm-up*								
SFT Checkpoint	45.21	68.29	39.79	58.62	24.58	34.68	36.53	53.86
+ SeqKD	40.21	68.08	30.63	50.00	20.62	36.78	26.11	42.87
+ OPD	43.13	67.38	38.12	56.84	23.13	34.95	34.79	53.06
+ ExOPD	45.83	72.31	36.04	53.86	24.37	38.02	35.41	54.73
+ GKD	**54.79**	**77.92**	42.08	63.69	25.83	40.21	40.90	60.61
+ **AOPD**	53.75	77.39	**45.21**	**66.86**	**30.42**	**46.73**	**43.13**	**63.66**
– *3k SFT Warm-up*								
SFT Checkpoint	58.96	74.95	47.71	66.73	31.04	44.19	45.90	61.96
+ SeqKD	51.04	76.56	40.00	56.98	25.42	43.80	38.82	59.11
+ OPD	61.46	78.61	52.29	70.44	34.17	50.86	49.31	66.64
+ ExOPD	61.67	77.55	48.75	65.83	34.17	50.22	48.20	64.53
+ GKD	66.25	**80.65**	48.54	**70.56**	34.17	49.67	49.65	66.96
+ **AOPD**	**66.87**	79.58	**55.00**	70.28	**38.33**	**52.42**	**53.40**	**67.43**

variant, which applies forward KL guidance only when the teacher–student gap is sufficiently negative. Figure 3c shows that this variant intervenes on fewer than 10% of all generated tokens. However, this sparse intervention already yields clear gains over OPD, confirming that targeted teacher distribution guidance is highly efficient at rescuing optimization from exploration black holes.

5 Experiments

5.1 Experimental Setup

Training Data and Benchmarks. Our experiments focus on two abilities: mathematical reasoning and tool use. For mathematical reasoning, we first use the OpenThoughts dataset (Guha et al., 2026) for warm-up training, and then conduct distillation on DeepMath (He et al., 2025). We evaluate mathematical reasoning on AIME 2024/2025 and HMMT 2025 (Feb). For tool use, we further train the model on ToolAlpaca (Tang et al., 2023) in the continual learning stage, and evaluate it on the test set with the tool-call success rate. This setting allows us to examine both reasoning distillation performance and the ability to acquire new skills while retaining previously learned reasoning capability.

Models. Following convention, we evaluate our method under two settings: (1) distilling from Qwen3-32B to Qwen3-8B-Base, and (2) distilling from Qwen3-8B to Qwen3-4B-Base (Yang et al., 2025). These challenging scenarios validate our method across different parameter scales.

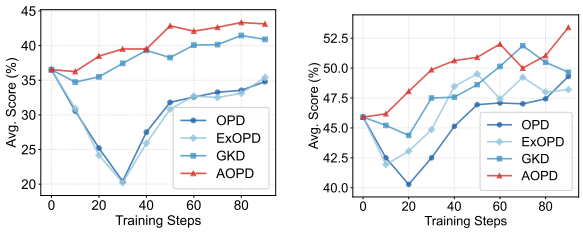
Baselines. We compare AOPD against four baselines: (i) **SeqKD** (Kim and Rush, 2016), which continues supervised fine-tuning on the trajectories generated by the teacher model; (ii) **OPD** (Lu and Lab, 2025), standard on-policy distillation using K1-estimated reverse KL divergence as the per-token advantage; (iii) **GKD** (Agarwal et al., 2024), which we implement as forward KL divergence on the teacher's top-K support over student-generated trajectories; and (iv) **ExOPD** (Yang et al., 2026), an extrapolation-based variant for OPD.

Training Pipeline. We first warm up the base models to equip them with basic abilities. Specifically, we warm up base models for 3,000 steps, alongside an experimental setting of 1,000 steps to yield variants with distinct foundational capabilities for robustness assessment. At distillation stage, models are first trained on mathematical tasks, followed by tool-call tasks for continual learning. Dur-

ing the subsequent distillation phase, we uniformly set the teacher support size K to 32 in AOPD and GKD for a fair comparison. Regarding ExOPD, we use its best-reported hyperparameters in the paper. More details about implementation and training hyperparameters are reported in Appendix C.

5.2 Main Results on Reasoning

Table 1 shows that AOPD consistently achieves the best overall reasoning performance across model scales and warm-up settings. For Qwen3-4B-Base, it already attains the best average Pass@1, demonstrating its effectiveness in smaller-scale distillation. For Qwen3-8B-Base with the more challenging 1k-step warm-up, AOPD still delivers the best overall performance, while standard OPD falls below the warm-up baseline, highlighting AOPD's robustness. Under the 3k-step warm-up, AOPD further achieves the best overall results, reaching 53.40 average Pass@1 and 67.43 average Pass@4. Extended results are provided in Appendix D.1.



(a) Weak initialization. (b) Strong initialization.

Figure 4: Training dynamics of Qwen3-8B-Base under different initializations.

Figure 4 further illustrates the training dynamics of different methods in terms of average math score. Across both settings, OPD and ExOPD exhibit a noticeable early performance drop. This issue is especially severe under the weaker 1k-step initialization, where recovery from poor states is difficult. GKD shows a more stable trajectory, but its final performance still falls short of AOPD. In contrast, AOPD improves steadily from the start under both initialization settings and consistently reaches the highest performance ceiling. Its asymmetric design not only avoids the early degradation but also translates training stability into stronger final reasoning performance. Detailed training dynamics are further shown in Appendix D.2.

5.3 Continual Learning

We evaluate continual learning on ToolAlpaca to test whether models can acquire a new ability while

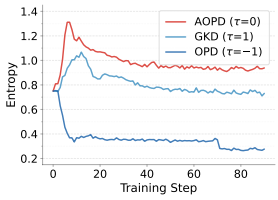


Figure 5: Policy entropy during training.

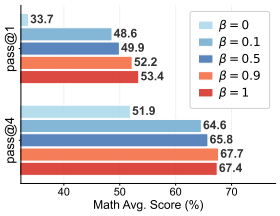


Figure 6: Effect of the divergence weighting.

retaining mathematical reasoning performance. Table 2 shows that AOPD achieves the best balance between adaptation and retention, attaining a competitive 75.0% success rate while preserving the strongest math performance. In contrast, OPD, ExOPD, and GKD all suffer math degradation after continual learning, with OPD showing the largest drop. These results suggest that localized teacher guidance enables adaptation to new abilities without overwriting prior reasoning ability.

Table 2: Performance after continual learning.

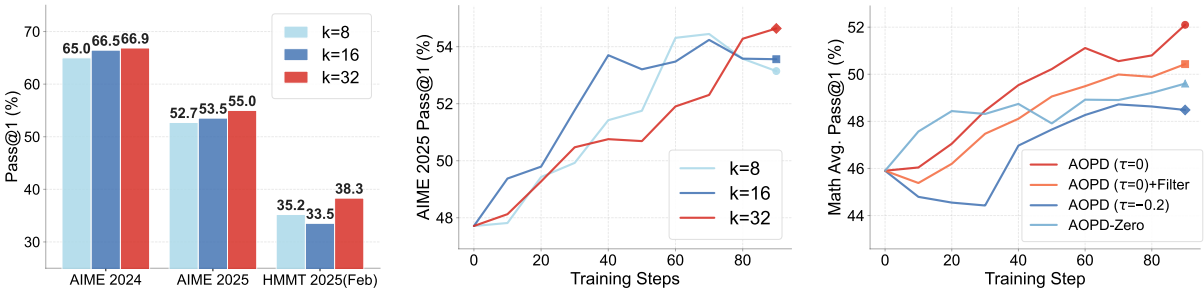
Method	Tool Use	Math Avg. Pass@4 Before / After	Drop (↓)
OPD	65.0	66.64 / 59.15	-7.49
ExOPD	67.0	64.53 / 63.04	-1.49
GKD	**76.0**	66.96 / 64.54	-2.42
AOPD	75.0	**67.43 / 68.04**	**+0.61**

We further examine the training dynamics of AOPD to explain its capability retention. As earlier shown in Figure 3c, the fraction of tokens receiving divergence guidance is limited and gradually decreases from about 40% to 30% during training. Most tokens are still optimized by policy gradient. Figure 5 further shows that OPD suffers severe entropy collapse, while GKD keeps the policy near its initial entropy. In contrast, AOPD maintains higher entropy throughout training, suggesting that it preserves a broader policy space while applying correction only at difficult positions. This helps explain why AOPD better acquires new tool-use ability while retaining prior reasoning competence. In other words, AOPD does not rely on globally reshaping the student into the teacher, but on selectively repairing the difficult parts of the trajectory where explicit guidance is most needed.

5.4 Ablation Study

5.4.1 Divergence Weighting in AOPD

The parameter β controls the interpolation between student-centric smoothing ($\beta \to 0$) and teacher-



(a) Scores of different top-K. (b) Training dynamics of different top-K. (c) Effect of intervention locations.

Figure 7: Ablation studies on top-K and intervention location.

centric guidance ($\beta \to 1$) during interventions. Figure 6 shows that AOPD attains its highest Pass@1 score at $\beta = 1$ and the highest Pass@4 at $\beta = 0.9$. More generally, the results show a clear trend that performance increases as the divergence becomes more biased toward the teacher distribution. This empirical trend further validates the support-consistency principle discussed in Section 4.2.

5.4.2 Effect of the Teacher Support Size

We further study the effect of the teacher support size K in AOPD. As K increases from 8 to 16 and 32 in Figure 7a, the final Pass@1 scores on AIME 2024/2025 and HMMT 2025 (Feb) improve, suggesting that a larger top-K preserves a richer portion of the teacher distribution and thus provides more complete signals. At the same time, we observe that smaller K often yields larger gains at the early stage of training, as shown in Figure 7b. This suggests that restricting supervision to a few high-probability teacher tokens produces denser and more concentrated learning signals, which is more beneficial for rapid initial improvement. In contrast, larger K brings less pronounced early gains. However, it retains a broader candidate distribution and therefore provides more informative signals in the later stage of training, leading to a higher final performance ceiling. While larger K could in principle further improve final performance, increasing K beyond 32 only adds low-probability tokens with negligible training signal and degrades training efficiency. We therefore set $K = 32$ as the default throughout our experiments.

5.4.3 Effect of Intervention Location

The effectiveness of AOPD depends not only on whether KL guidance is used, but also on where the intervention is triggered. We consider three strategies: a token-level threshold τ, which selects low-confidence positions for intervention; an accuracy filter, which applies KL guidance only to negative advantage regions from trajectories with incorrect final answers; and AOPD-Zero, which focuses exclusively on neutral regions with $A_t \approx 0$.

Figure 7c shows that standard AOPD with $\tau = 0$ achieves the strongest overall performance. However, the more conservative setting $\tau = -0.2$ improves more slowly, indicating that shrinking the intervention range weakens the benefit of teacher guidance. Extended studies on τ are presented in Appendix D.3. The accuracy-filtered setting remains competitive but consistently trails standard AOPD under the same threshold, suggesting that informative correction signals also arise in correct trajectories. Meanwhile, AOPD-Zero outperforms $\tau = -0.2$, which confirms that neutral regions are a major source of optimization stagnation in OPD. These results highlight that the benefit of AOPD is not confined to correcting obvious errors, but also arises in correct trajectories or neutral positions. Overall, the best results are achieved only by standard AOPD, and any reduction in the set of intervened positions leads to a drop in performance.

6 Conclusion

In this work, we identify three limitations of OPD and propose asymmetric on-policy distillation (AOPD). AOPD preserves policy gradient updates where they are reliable and applies direct teacher distribution alignment where stronger correction is needed. Evaluations across multiple benchmarks show that AOPD consistently improves mathematical reasoning and remains stable under different model scales and initialization steps, where standard OPD often struggles. More broadly, our results suggest that combining exploitation and imitation at the token level can preserve policy entropy and better retain previously acquired capabilities. We further discuss the limitations in Appendix E.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2024. On-Policy distillation of language models: Learning from self-generated mistakes. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28.

Jack Chen, Fazhong Liu, Naruto Liu, Yuhan Luo, Erqu Qin, Harry Zheng, Tian Dong, Haojin Zhu, Yan Meng, and Xiao Wang. 2025a. Step-wise adaptive integration of supervised fine-tuning and reinforcement learning for task-specific llms. *Preprint*, arXiv:2505.13026.

Liang Chen, Xueting Han, Li Shen, Jing Bai, and Kam-Fai Wong. 2025b. Beyond two-stage training: Cooperative SFT and RL for LLM reasoning. *arXiv preprint arXiv:2509.06948*.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. 2026. SRFT: A single-stage method with supervised and reinforcement fine-tuning for reasoning. In *Proceedings of the Fourteenth International Conference on Learning Representations*.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, and 168 others. 2026. Glm-5: from vibe coding to agentic engineering. *Preprint*, arXiv:2602.15763.

Naibin Gu, Chenxu Yang, Qingyi Si, Chuanyu Qin, Dingyu Yao, Peng Fu, Zheng Lin, Weiping Wang, Nan Duan, and Jiaqi Wang. 2026. Co-evolving policy distillation. *Preprint*, arXiv:2604.27083.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. MiniLLM: Knowledge distillation of large language models. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Rea Sprague, Ashima Suvarna, Benjamin Feuer, Leon Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, and 32 others. 2026. Openthoughts: Data recipes for reasoning models. In *The Fourteenth International Conference on Learning Representations*.

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. 2025. DeepMath-103K: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2014. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Zeyu Huang, Tianhao Cheng, Zihan Qiu, Zili Wang, Yinghui Xu, Edoardo M. Ponti, and Ivan Titov. 2025. Blending supervised and reinforcement Fine-Tuning with prefix sampling. *arXiv preprint arXiv:2507.01679*.

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

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

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

Kevin Lu and Thinking Machines Lab. 2025. On-Policy distillation. *Thinking Machines Lab: Connectionism*.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Yanhao Li, Wentao Zhang, and Bin CUI. 2026. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. In *The Fourteenth International Conference on Learning Representations*.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. 2026. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*.

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

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. 2023. ToolAlpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*.

Xinyu Tang, Yuliang Zhan, Zhixun Li, Wayne Xin Zhao, Zhenduo Zhang, Zujie Wen, Zhiqiang Zhang, and Jun Zhou. 2025. Rethinking sample polarity in reinforcement learning with verifiable rewards. *Preprint*, arXiv:2512.21625.

Core Team, Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, Gang Xie, Hailin Zhang, Hanglong Lv, Hanyu Li, Heyu Chen, Hongshen Xu, Houbin Zhang, Huaqiu Liu, and 107 others. 2026a. Mimo-v2-flash technical report. *Preprint*, arXiv:2601.02780.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, and 307 others. 2026b. Kimi k2.5: Visual agentic intelligence. *Preprint*, arXiv:2602.02276.

Yijun Tian, Yikun Han, Xiusi Chen, Wei Wang, and Nitesh V. Chawla. 2025. Beyond answers: Transferring reasoning capabilities to smaller LLMs using multi-teacher knowledge distillation. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining*, pages 251–260.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. 2025. Learning to reason under Off-Policy guidance. In *Advances in Neural Information Processing Systems*, volume 38.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. 2026. Learning beyond teacher: Generalized On-Policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. 2026a. On-Policy RL meets Off-Policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. In *Proceedings of the Fourteenth International Conference on Learning Representations*.

Xinsen Zhang, Zhenkai Ding, Tianjun Pan, Run Yang, Chun Kang, Xue Xiong, and Jingnan Gu. 2026b. Opsdl: On-policy self-distillation for long-context language models. *Preprint*, arXiv:2604.17535.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. 2026. Self-distilled reasoner: On-policy self-distillation for large language models. *Preprint*, arXiv:2601.18734.

Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. 2026. Scope: Signal-calibrated on-policy distillation enhancement with dual-path adaptive weighting. *Preprint*, arXiv:2604.10688.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. 2025. The surprising effectiveness of negative reinforcement in LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

A Related Work about Joint SFT-RL Optimization

Because on-policy distillation follows a reinforcement learning paradigm, it remains vulnerable to exploration bottlenecks. To alleviate this issue, recent work incorporates supervised learning or expert demonstrations into RL. Some methods dynamically interleave SFT and RL, such as ReLIFT (Ma et al., 2026), SASR (Chen et al., 2025a), and BRIDGE (Chen et al., 2025b), while others unify them in a single-stage objective, including SRFT (Fu et al., 2026) and Prefix-RFT (Huang et al., 2025). A related line instead interprets SFT as off-policy guidance within on-policy RL, as in LUFFY (Yan et al., 2025) and CHORD (Zhang et al., 2026a). In this spirit, AOPD introduces localized soft-label supervised learning to improve exploration efficiency in OPD.

B Extended Analysis on Truncated Teacher Guidance

B.1 Gradient Analysis

In this section, we compare the logit-level gradients induced by forward KL and reverse KL under the same teacher-defined support S_t. For simplicity, $P_S(\cdot \mid c_t)$ and $P_T(\cdot \mid c_t)$ denote the normalized student and teacher distributions on S_t. We first consider the forward KL objective

$$\begin{aligned}\mathcal{L}_t^{\mathrm{FKL}} &= D_{\mathrm{KL}}(P_T(\cdot \mid c_t) \,\|\, P_S(\cdot \mid c_t)) \\ &= \sum_{v \in S_t} P_T(v \mid c_t) \log \frac{P_T(v \mid c_t)}{P_S(v \mid c_t)}.\end{aligned} \tag{13}$$

Taking derivatives with respect to the student logits, the term $\sum_{v \in S_t} P_T(v \mid c_t) \log P_T(v \mid c_t)$ can be treated as a constant. We therefore have

$$\begin{aligned}\frac{\partial \mathcal{L}_t^{\mathrm{FKL}}}{\partial z_j} &= \frac{\partial}{\partial z_j}\left(-\sum_{v \in S_t} P_T(v \mid c_t) \log P_S(v \mid c_t)\right) \\ &= P_S(j \mid c_t) - P_T(j \mid c_t), \qquad j \in S_t.\end{aligned} \tag{14}$$

We next consider the reverse KL objective

$$\begin{aligned}\mathcal{L}_t^{\mathrm{RKL}} &= D_{\mathrm{KL}}(P_S(\cdot \mid c_t) \,\|\, P_T(\cdot \mid c_t)) \\ &= \sum_{v \in S_t} P_S(v \mid c_t) \log \frac{P_S(v \mid c_t)}{P_T(v \mid c_t)}.\end{aligned} \tag{15}$$

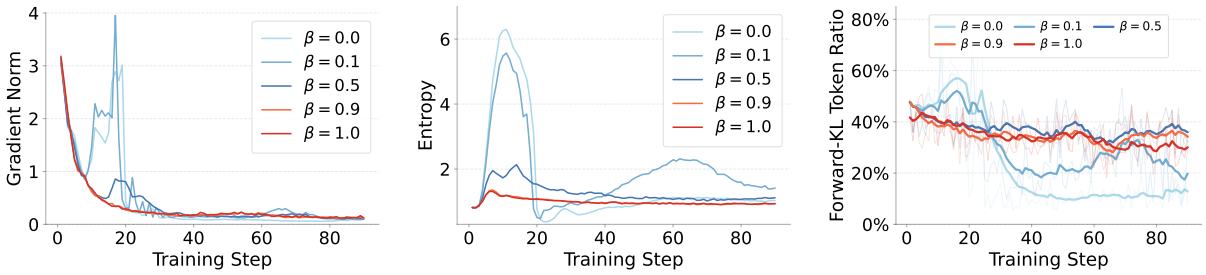
Differentiating with respect to z_j gives

$$\begin{aligned}\frac{\partial \mathcal{L}_t^{\mathrm{RKL}}}{\partial z_j} &= P_S(j \mid c_t)\Big(\log \frac{P_S(j \mid c_t)}{P_T(j \mid c_t)} \\ &\quad - D_{\mathrm{KL}}(P_S(\cdot \mid c_t) \,\|\, P_T(\cdot \mid c_t))\Big), \\ &\quad j \in S_t.\end{aligned} \tag{16}$$

The difference between Eq. 14 and Eq. 16 is important in the intervention regime. Forward KL yields a direct distribution gap correction on the selected support. By contrast, the reverse KL gradient depends on the current student distribution in a more involved way. Its update is explicitly scaled by $P_S(j \mid c_t)$, and the inner term is also determined by a student-dependent log ratio relative to the global reverse KL value. As a result, the recovery signal is less direct when a teacher-preferred token has already been strongly suppressed by the student. This comparison explains why forward KL is better suited to the teacher-defined support used in AOPD.

B.2 Training Dynamics

We further examine the training dynamics under different β values and observe an anomalous pattern when the objective tilts toward reverse KL. Figure 8a shows that as the divergence becomes more biased toward the student distribution, the optimization exhibits more unstable gradient behavior. In particular, the gradient norm exhibits the most severe spikes at $\beta = 0$ and 0.1, whereas the teacher-leaning settings $\beta = 0.9$ and 1 yield the smoothest training dynamics. Figure 8b shows that the entropy under reverse KL ($\beta = 0.0$) undergoes a sharp initial surge, followed by an abrupt collapse to near zero. This trajectory indicates that the student rapidly transits into a converged, low-entropy policy. Concurrently, Figure 8c reveals that the proportion of tokens receiving divergence guidance plummets at the same entropy collapse phase, implying that most generated positions register positive advantages and exit the intervention regime. The ablation results in Section 5.4.1 demonstrate that this apparent convergence coincides with severe degradation in reasoning capability. We attribute this to a reward-hacking phenomenon in on-policy distillation. The student discovers a mode-collapsed strategy that artificially suppresses the reverse KL value while failing to acquire the teacher's reasoning patterns. In contrast, forward KL maintains stable entropy and sustained



(a) Gradient norm during training.

(b) Policy entropy evolution.

(c) Ratios of tokens receiving JS divergence guidance.

Figure 8: Training dynamics under different β values.

intervention throughout training, yielding both superior and more robust performance.

C Implementation Details

Environment. Experiments are implemented on top of VeRL (Sheng et al., 2024) (v0.3.1) for on-policy distillation training and vLLM (Kwon et al., 2023) (v0.8.5) for teacher model deployment. Training is distributed via PyTorch FSDP and coordinated by Ray across 16 A100-80G GPUs, with an additional 8 A100-80G GPUs dedicated to serving the teacher model.

Table 3: Training and evaluation parameters.

Hyperparameter	Value
Training	
Learning rate	1×10^{-5}
Train batch size	512
Max response length	12288
Rollout temperature	1.0
Teacher top-K	32
Evaluation	
Temperature	0.6
top-p	0.95
top-k	20
Max generation length	32768

Parameters. Table 3 summarizes the main training and evaluation parameters used in our experiments. The full set of parameters is available in our code repository.

D Extended Experiments

D.1 Extended Main Results

We further supplement Table 4 with the results of Qwen3-4B-Base under a 1K-step SFT warm-up initialization. We observe that AOPD still attains the best overall performance, which is consistent with the trend observed in the main text.

D.2 Detailed Training Curves

We report the full test-set performance dynamics throughout training for both warm-up settings on Qwen3-8B-Base. Figure 10 traces Pass@1/4/8 every 10 steps across the training trajectory.

Across both initialization settings, AOPD consistently achieves the highest scores on Pass@1/4/8 across all benchmarks. Under weak initialization, this advantage is particularly pronounced. The substantial performance gap indicates that AOPD remains effective even for students with weaker foundational capabilities. Under strong initialization, the margin between AOPD and baselines narrows, though AOPD still reaches the highest performance ceiling. This convergence suggests that methods such as OPD are more dependent on the student's initial capability. A stronger base model produces trajectories where student tokens more frequently receive positive advantages from the teacher, thereby reducing the frequency of exploration black holes and alleviating the structural weaknesses of negative reinforcement.

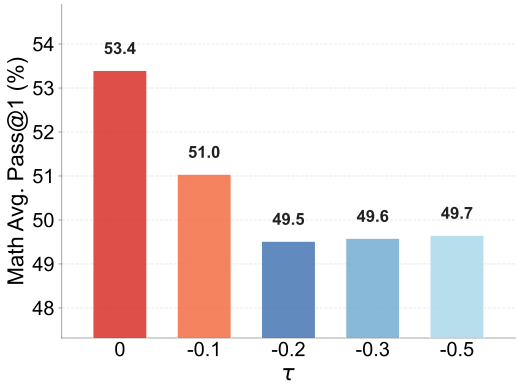
D.3 Extended Ablation Studies on τ

We further conduct an extended study on the intervention threshold τ. As shown in Figure 9a, the mathematical reasoning performance gradually declines as τ decreases, and begins to plateau once τ reaches -0.2. This trend suggests that reducing τ lowers the proportion of positions that receive forward KL supervision, thereby weakening the teacher-guided correction effect in AOPD. As the intervention becomes increasingly sparse, the model correspondingly moves closer to standard OPD, and the performance eventually stabilizes.

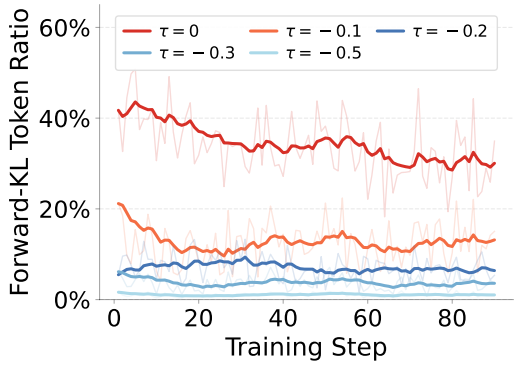
This behavior is further explained by the intervention ratio in Figure 9b. When $\tau \leq -0.2$, the fraction of tokens that trigger the supervision loss drops to below 10% of all generated tokens. As

Table 4: Supplementary distillation results for Qwen3-4B-Base across multiple math reasoning benchmarks. We compare AOPD against SeqKD, OPD, GKD and ExOPD. **Bold** indicates the best result within each configuration block, while underlined values denote the second best.

Method	AIME 2024		AIME 2025		HMMT 2025 (Feb)		Average	
	Pass@1	*Pass@4*	*Pass@1*	*Pass@4*	*Pass@1*	*Pass@4*	*Pass@1*	*Pass@4*
Qwen3-4B-Base								
– *1k SFT Warm-up*								
SFT Checkpoint	36.25	58.61	32.71	47.25	19.17	28.80	29.38	44.89
+ SeqKD	42.50	65.23	40.00	60.09	24.58	38.01	35.69	54.44
+ OPD	46.25	65.99	42.08	60.75	22.50	35.67	36.94	54.14
+ ExOPD	44.58	64.41	41.46	61.70	20.62	32.32	35.55	52.81
+ GKD	50.21	70.47	**44.17**	63.03	27.50	42.21	40.63	58.57
+ **AOPD**	**51.25**	**72.01**	43.96	**63.49**	**28.33**	**46.02**	**41.18**	**60.51**



(a) Scores of different τ



(b) Ratio of tokens receiving forward KL guidance.

Figure 9: Training dynamics under different τ values.

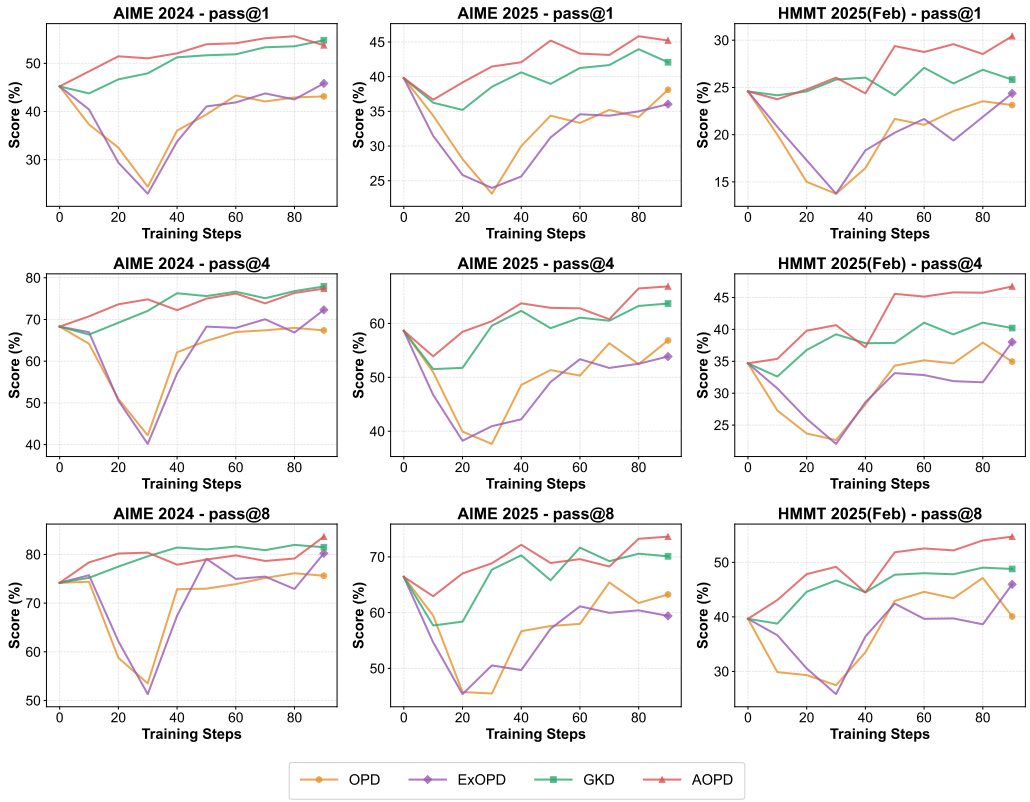
a result, only a very small subset of positions receives explicit teacher guidance, which explains why the performance gap becomes relatively small once τ is reduced beyond -0.2. These results suggest that the full benefit of AOPD is realized only when teacher guidance spans the entire zero-advantage and negative-advantage regions, whereas reducing forward KL supervision in the negative-advantage region correspondingly weakens the effect of AOPD.

E Limitations

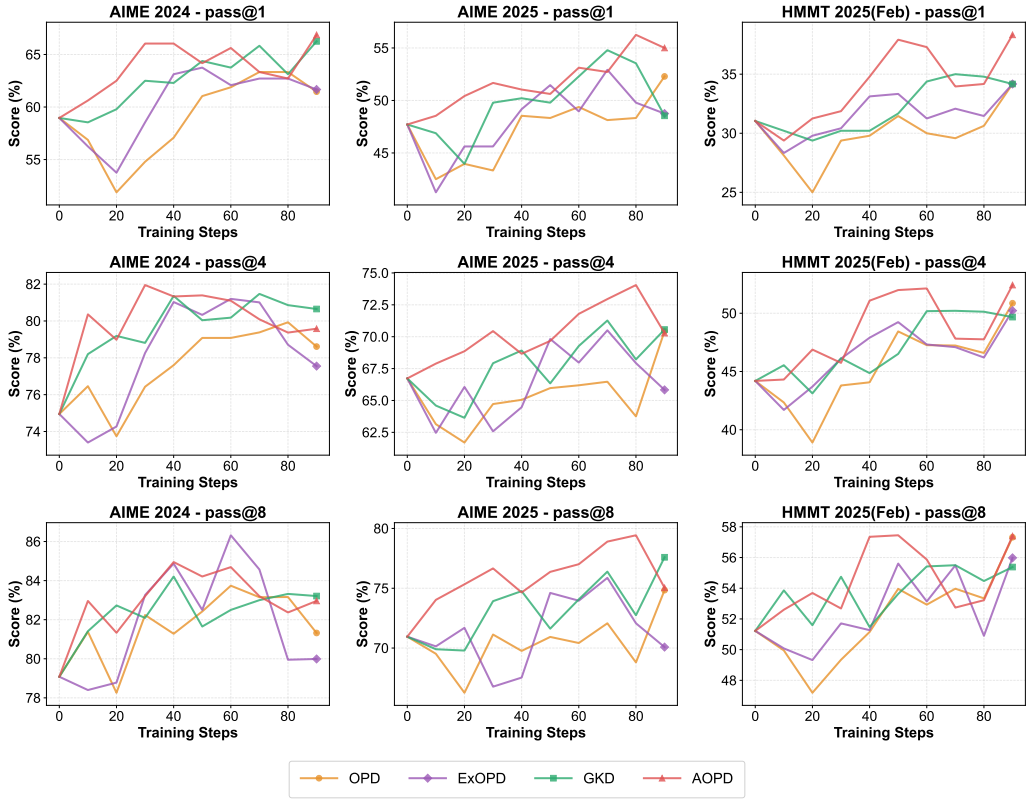
Our evaluation is limited to mathematical reasoning and tool-use settings, and it remains to be explored how AOPD performs on other task types. In addition, the current instantiation of AOPD adopts a fixed intervention design, and future work may investigate more adaptive intervention strategies.

F LLM Usage Statement

During the preparation of this manuscript, we utilized LLM as a general-purpose writing assistant. The use of LLMs was primarily for improving grammar, spelling, and clarity. We reviewed and edited all suggested changes and take full responsibility for the final content of this paper. All research contributions, including problem formulation, methodology design, implementation, experiments, data analysis, and scientific conclusions, were conceived, executed, and verified by the authors. This use conforms to the ACL Ethics Policy regarding authorship criteria.



(a) Weak initialization.



(b) Strong initialization.

Figure 10: Detailed training dynamics of Qwen3-8B-Base under weak and strong initialization.