BCOD: RELAXING SUPPRESSIVE TOKEN FEEDBACK IN ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) provides dense teacher supervision on student-generated reasoning trajectories, but its sampled-token reverse-KL update can suppress student-sampled alternatives that the teacher assigns lower probability. We propose **B**alanced-**C**lip **O**n-policy **D**istillation (**BCOD**), which preserves positive teacher–behavior gaps and scales non-positive gaps by λ. At the on-policy point, its expected token update descends a convex asymmetric f-divergence with the same teacher-matching optimum as reverse KL. The operator needs only one rollout. In a matched ablation isolating the operator, attenuation improves Avg@32/Pass@32 by $2.1/2.1$ over OPD; in the full evaluated system, which inherits SCOPE's correctness routing, BCOD reaches 54.6 Avg@32 and 74.5 Pass@32, versus 49.3 and 69.1 for standard OPD (five datasets; SkyWork-OR1-7B as teacher, R1-Distill-Qwen-1.5B as student). On fixed sets of 32 rollouts per prompt, BCOD also retains more observed base-model solutions. These results support attenuating suppressive token feedback in the evaluated OPD setting.

1 INTRODUCTION

On-policy distillation (OPD) is a practical way to add dense token-level teacher supervision to reasoning alignment, complementing reinforcement learning with verifiable rewards (RLVR) when outcome rewards are sparse (Guo et al., 2025; Shao et al., 2024; Lu & Lab, 2025). In OPD, the student samples its own rollouts and is guided by a teacher on the same token sequence, commonly through a reverse-KL signal (Lu & Lab, 2025; Agarwal et al., 2024). At the same time, recent work shows that post-training can improve single-sample accuracy while reducing large-k coverage, making Pass@k a critical diversity-sensitive metric for reasoning models (Zhu et al., 2026; Li et al., 2026a; Zheng et al., 2026).

We focus on the sign of the sampled-token reverse-KL coefficient. For a sampled token a_t, let $A_t = \log \pi_T(a_t) - \log \pi_{\text{old}}(a_t)$ be the teacher–behavior log-probability gap. Positive gaps form a *boost branch*; non-positive gaps form a *suppression branch* that can reduce the probability of student-sampled alternatives. This offers a plausible token-level explanation for coverage loss under OPD.

In the primary evaluation suite, OPD increases Avg@32 from 47.6 to 49.3 while decreasing Pass@32 from 75.4 to 69.1. With routing disabled, attenuating the suppression branch improves both metrics by 2.1 points in the matched ablation (Table 2).

BCOD leaves boost updates unchanged and multiplies suppression updates by $\lambda \in (0, 1]$: $R_t = A_t$ when $A_t > 0$ and $R_t = \lambda A_t$ otherwise. A negative gap indicates teacher–behavior disagreement, but does not establish that a sampled continuation is incorrect; this matters when multiple reasoning paths can solve a problem. We show that the expected on-policy update descends an asymmetric divergence with the teacher-matching optimum. The rescaling uses only sampled-token log-probabilities, so it is defined with $N{=}1$ rollout. The evaluated training system also reuses SCOPE's correctness routing (Zheng et al., 2026).

Our contributions are:

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.

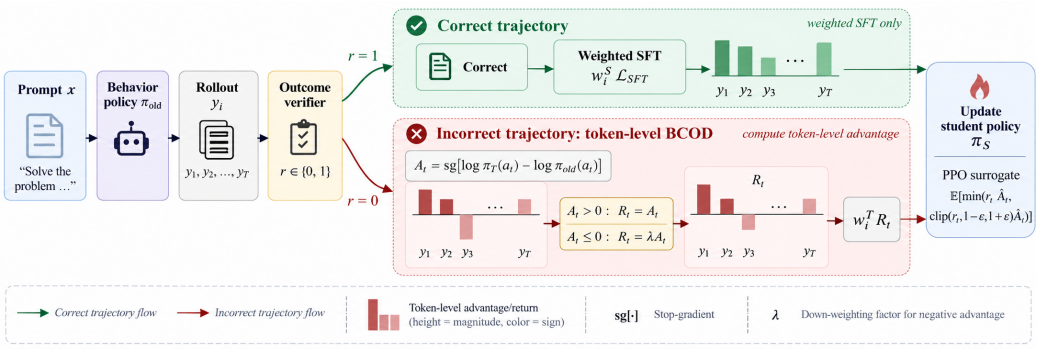


Figure 1: BCOD training graph. A verifier routes correct trajectories to weighted supervised fine-tuning (SFT), while incorrect trajectories use a detached teacher–behavior log-probability gap. BCOD leaves positive gaps unchanged and scales non-positive gaps by λ before applying the trajectory weight w_i^T and the PPO surrogate. The two losses are combined only in the student-policy update.

- **Diagnosing suppressive feedback in OPD.** We distinguish boost and suppression branches in the sampled-token reverse-KL signal and test whether attenuating the latter improves accuracy and coverage.
- **Introducing BCOD and its divergence interpretation.** Its expected local update descends a convex asymmetric f-divergence with reverse KL's teacher-matching optimum. The token-local operator is defined for $N{=}1$ rollout.
- **Evaluating accuracy and coverage.** BCOD improves Avg@32 and Pass@32 over OPD by 5.3 and 5.4 points in the primary setup and reduces the macro fraction of observed forgotten problems from 7.7% to 2.8% on fixed rollout sets.

2 Method

BCOD modifies the token-level reverse-KL coefficient on the distillation path while retaining the correctness-based routing of the underlying training pipeline. Figure 1 summarizes the data flow used in the reported configurations.

2.1 Reverse-KL token signal and its sign

Let x denote a prompt and $y = (a_1, \ldots, a_T)$ a response sampled from the behavior policy π_{old}. We write π_T for the teacher and π_θ for the student policy being optimized. For the sampled token a_t, define the detached reverse-KL token signal

$$A_t \;=\; \text{stopgrad}[\log \pi_T(a_t \mid x, a_{<t}) \;-\; \log \pi_{\text{old}}(a_t \mid x, a_{<t})]\,, \tag{1}$$

so that gradients do not flow through either the teacher or the behavior-policy log-probabilities. At the on-policy point $\pi_\theta = \pi_{\text{old}}$, the corresponding local gradient-ascent contribution can be written as

$$g_{\text{OPD}} \;=\; \sum_{t:A_t>0} A_t \nabla_\theta \log \pi_\theta(a_t \mid x, a_{<t}) \;+\; \sum_{t:A_t\le 0} A_t \nabla_\theta \log \pi_\theta(a_t \mid x, a_{<t}). \tag{2}$$

We call the first term the boost branch and the second term the suppression branch. The coverage-loss analysis in Section 1 motivates modifying only the latter.

2.2 Direction-aware rescaling objective

BCOD preserves the boost branch and multiplies the suppression branch by λ:

$$R_t \;=\; A_t\Big(\mathbf{1}\{A_t > 0\} + \lambda\, \mathbf{1}\{A_t \le 0\}\Big), \qquad \lambda \in [0, 1]. \tag{3}$$

The value $\lambda = 0$ is included as a boundary case; the practical BCOD configurations use $\lambda > 0$.

2.3 SELECTIVE TRUST AS AN ASYMMETRIC DIVERGENCE

The gap sign distinguishes two local updates. For a positive gap, the teacher assigns a sampled action higher probability than the behavior policy, supporting an increase in that action's probability. For a negative gap, the teacher assigns it lower probability; this preference alone does not establish that the sampled token is incorrect. In multi-solution reasoning, the teacher may prefer one derivation even when another is valid. BCOD therefore retains positive-gap updates and attenuates negative-gap updates.

This selective-trust principle has a divergence interpretation. At a fixed prefix $s = (x, a_{<t})$, write $p_\theta(a) = \pi_\theta(a \mid s)$, $q(a) = \pi_T(a \mid s)$, and $z = p_\theta(a)/q(a)$. For $\lambda \in (0, 1]$, define

$$f_\lambda(z) = \begin{cases} z \log z - z + 1, & z \le 1, \\ \lambda\big(z \log z - z + 1\big), & z > 1, \end{cases} \qquad D_{f_\lambda}(p_\theta \| q) = \sum_a q(a) f_\lambda\left(\frac{p_\theta(a)}{q(a)}\right). \tag{4}$$

Proposition (local divergence interpretation). Assume $p_\theta(a) > 0$ and $q(a) > 0$. The generator f_λ in Eq. (4) is strictly convex for $\lambda > 0$, and D_{f_λ} is non-negative with the unique distribution-level minimum $p_\theta = q$. At the on-policy point $p_\theta = \pi_{\text{old}}$, its negative gradient is

$$-\nabla_\theta D_{f_\lambda}(p_\theta \| q) = \mathbb{E}_{a \sim p_\theta}\left[A(a)\Big(\mathbf{1}\{A(a) > 0\} + \lambda \mathbf{1}\{A(a) \le 0\}\Big)\nabla_\theta \log p_\theta(a)\right], \tag{5}$$

where $A(a) = \log q(a) - \log p_\theta(a)$. Thus, the sampled coefficient multiplied by $\nabla_\theta \log p_\theta(a)$ is an unbiased estimator of the negative gradient of this conditional divergence at the on-policy point.

To see this, note that $f'_\lambda(z) = \log z$ for $z \le 1$ and $f'_\lambda(z) = \lambda \log z$ for $z > 1$; both one-sided derivatives equal zero at $z = 1$. Moreover, $f''_\lambda(z) = 1/z$ on the first branch and λ/z on the second, establishing strict convexity for $\lambda > 0$. Differentiating Eq. (4) gives $\nabla_\theta D_{f_\lambda} = \mathbb{E}_{a \sim p_\theta}[f'_\lambda(p_\theta(a)/q(a))\nabla_\theta \log p_\theta(a)]$, which yields Eq. (5). Non-negativity and the unique minimum follow from convexity, $f_\lambda(1) = 0$, and $\sum_a q(a)p_\theta(a)/q(a) = 1$.

When $\lambda = 1$, Eq. (4) is $D_{\text{KL}}(p_\theta \| q)$. For $0 < \lambda < 1$, the teacher-matching optimum remains unchanged, while curvature decreases from $1/z$ to λ/z only where $p_\theta(a) > q(a)$. At $\lambda = 0$, this branch has zero curvature and supplies no local restoring update.

The proposition characterizes the unclipped conditional update at the on-policy point. The implemented optimizer additionally uses behavior-policy detachment, PPO ratio clipping, changing rollout prefixes, and correctness routing. The experiments report sequence-level metrics for this complete training system.

2.4 PPO SURROGATE AND COMPLETE OBJECTIVE

For the PPO actor ratio $r_t(\theta) = \pi_\theta(a_t \mid x, a_{<t})/\pi_{\text{old}}(a_t \mid x, a_{<t})$, define the ratio-clipped value $\bar{r}_t = \min\{\max\{r_t, 1 - \epsilon_\ell\}, 1 + \epsilon_h\}$. For any scalar token coefficient u_t, the implemented dual-clip PPO token loss is

$$\ell_t(\theta; u_t) = \begin{cases} \min\{-cu_t,\ \max(-u_t r_t, -u_t \bar{r}_t)\}, & u_t < 0, \\ \max\{-u_t r_t, -u_t \bar{r}_t\}, & u_t \ge 0, \end{cases} \qquad c > 1. \tag{6}$$

At $r_t = 1$, the local ascent direction induced by the minimized loss is $R_t \nabla_\theta \log \pi_\theta(a_t \mid x, a_{<t})$ before trajectory weighting. The unweighted sampled-trajectory contribution is

$$g_{\text{BCOD}} = \sum_{t: A_t > 0} A_t \nabla_\theta \log \pi_\theta(a_t \mid x, a_{<t}) + \lambda \sum_{t: A_t \le 0} A_t \nabla_\theta \log \pi_\theta(a_t \mid x, a_{<t}). \tag{7}$$

For the complete system, let w_i^T and w_i^S denote the inherited teacher- and student-trajectory weights for rollout i, respectively. Let M_{it} be the response-token mask and $C_i = \mathbf{1}\{i \text{ is correct}\}$. With the token-mean aggregation used in our experiments, the objective is

$$\mathcal{L}_{\text{BCOD}} = \frac{\sum_{i,t} M_{it}\, \ell_{it}(\theta; w_i^T R_{it})}{\sum_{i,t} M_{it}} + \frac{\sum_{i,t} M_{it} C_i w_i^S \big(-\log \pi_\theta(a_{it} \mid x_i, a_{i,<t})\big)}{\sum_{i,t} M_{it} C_i}. \tag{8}$$

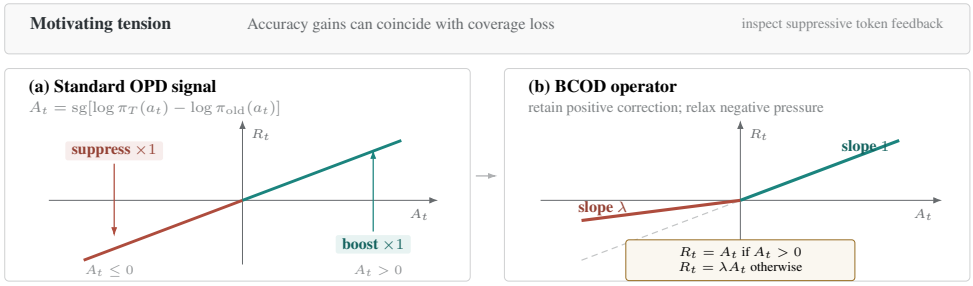


Figure 2: Schematic of the OPD sampled-token coefficient and BCOD rescaling. The negative-gap branch has unit weight under OPD and weight λ under BCOD; the positive-gap branch is unchanged. The observed accuracy–coverage tradeoff motivating this operator is reported in Table 1.

Here $w_i^T = 0$ for correct trajectories and $w_i^S = 0$ for incorrect trajectories. The PPO term is aggregated over the full response mask, so its denominator includes correct-response tokens even though their effective coefficient $w_i^T R_{it}$ is zero. The SFT term is aggregated only over correct-response tokens and is defined to be zero for a batch that contains none. Thus, incorrect trajectories receive the weighted BCOD distillation loss, whereas correct trajectories receive weighted SFT. Equations (1)–(8) describe the optimization graph used for all reported BCOD results; helper losses not connected to this graph are omitted.

Figure 2 contrasts the OPD and BCOD token coefficients. Equation (3) gives the exact operator, including the $\lambda = 0$ boundary case.

2.5 Boundary cases and trajectory-level weighting

Setting $\lambda = 0$ removes the suppression branch. In the evaluated routing-enabled configuration, this boost-only boundary case collapses during training into a degenerate early-EOS regime. We therefore use a positive coefficient to attenuate, rather than eliminate, suppressive feedback.

SCOPE (Zheng et al., 2026) routes trajectories by correctness and assigns group-relative weights. BCOD modifies the token coefficient only within the distillation path; it adds no routing rule.

Because the rescaling uses only the sampled token's teacher and behavior-policy log-probabilities, it is defined with $N{=}1$ rollout. Correctness routing still applies; with one rollout per prompt, the active inherited sequence weight is 1. Section 3.8 evaluates generation cost.

3 Experiments

3.1 Setup

Models and data. Following SCOPE (Zheng et al., 2026), the primary setup distills DeepSeek-R1-Distill-Qwen-1.5B (Guo et al., 2025) from SkyWork-OR1-7B (He et al., 2025), trained on DeepMath-103k (He et al., 2026). A second setup distills Qwen3-1.7B-Instruct from Qwen3-8B-Instruct (Yang et al., 2025) on 25% of the same data. We use the *instruct* variant because the *base* variant often repeats until the generation cap during training rollouts. The instruct model's <|im_end|> token terminates these rollouts naturally; this training issue differs from SCOPE's evaluation-time repetition-penalty fix.

Metrics. We report Avg@32 and Pass@32 over five benchmarks (5ds): AIME 2024, AIME 2025, AMC 2023, MATH500 (Hendrycks et al., 2021), and Minerva (Lewkowycz et al., 2022). For each prompt, we sample 32 independent rollouts. Avg@32 is the average correctness over the 32 rollouts, while Pass@32 denotes whether at least one rollout is correct. Dataset scores are macro-averaged over prompts. All evaluations use rollout$=32$, temperature 0.6, top-p 0.95, top-k 20, and max tokens 16384. DeepSeek uses repetition penalty 1.0, while Qwen3-Instruct uses 1.08.

Baselines. We compare against the base student, GRPO (Shao et al., 2024), local SeqKD, standard OPD (reverse-KL, $\lambda = 1.0$) (Lu & Lab, 2025), a SCOPE reproduction at $N{=}4$ (Zheng et al.,

Model	AIME24		AIME25		AMC23		MATH500		Minerva		**Average**	
	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32	A@32	P@32
Teacher: SkyWork-OR1-7B → Student: R1-Distill-Qwen-1.5B												
Base	26.9	76.4	27.3	56.4	67.5	94.0	84.2	97.6	32.0	52.6	47.6	75.4
w/ GRPO	31.5	69.8	34.4	59.8	74.3	96.5	85.8	97.1	34.2	52.9	52.0	75.2
w/ SeqKD	28.0	70.0	25.0	45.0	68.0	94.0	85.0	97.0	31.0	54.0	47.4	72.0
w/ OPD	29.5	62.0	33.3	44.4	65.5	92.0	86.9	97.1	31.4	49.8	49.3	69.1
w/ SCOPE-N4	34.7	71.9	34.6	44.4	70.9	94.6	88.1	97.0	34.0	50.4	52.4	71.7
w/ EOPD	27.5	68.3	25.2	44.0	61.1	85.1	85.3	97.0	29.1	49.2	45.6	68.7
w/ TrOPD	32.5	65.0	36.3	47.4	68.5	94.0	89.9	98.0	34.4	52.8	52.3	71.4
w/ **BCOD**	**37.7**	**72.1**	**36.2**	**54.7**	**74.6**	**94.9**	**89.4**	**97.6**	**34.9**	**53.1**	**54.6**	**74.5**
Teacher: Qwen3-8B-Instruct → Student: Qwen3-1.7B-Instruct (25% data)												
Base	2.9	27.9	1.3	6.7	17.4	68.8	39.1	83.5	6.0	30.6	13.3	43.5
w/ GRPO	33.5	62.0	31.0	42.0	69.0	87.0	87.5	96.0	36.0	52.0	51.4	67.8
w/ SeqKD	35.0	67.0	32.5	45.0	72.0	89.0	88.5	97.0	38.0	55.0	53.2	70.6
w/ OPD	37.0	68.5	35.0	46.7	74.5	90.0	90.0	98.0	40.7	58.1	55.4	72.3
w/ SCOPE-N4	38.8	71.0	34.8	51.0	74.1	90.0	90.3	98.0	40.3	57.1	55.6	73.4
w/ EOPD	35.5	67.4	31.0	45.4	72.0	88.8	89.4	97.4	40.1	57.5	53.6	71.3
w/ TrOPD	36.8	69.2	34.3	49.0	73.2	90.4	89.9	97.8	40.3	56.1	54.9	72.5
w/ **BCOD**	**39.2**	**72.3**	**36.5**	**52.0**	**74.7**	**92.1**	**90.2**	**97.6**	**41.0**	**56.8**	**56.3**	**74.2**

Table 1: Results for both teacher–student configurations (5-dataset average, %). Avg@32 (A@32) and Pass@32 (P@32). Bold marks the primary BCOD configuration ($\lambda = 0.3$) in each block; measured rows share local prompts, decoding, and scorer.

2026), and BCOD variants. We additionally report a local EOPD reproduction in the primary setup. Optimization and evaluation-selection details are provided in Appendix A.

Training. BCOD/OPD/SCOPE use rollout temperature 0.6 and learning rate 5×10^{-5}. First configuration train on all of DeepMath-103k; the second configuration uses 25% of DeepMath. Additional training and reporting details are deferred to Appendix A.

3.2 Main results

Table 1 reports per-benchmark results for both teacher–student setups. We first examine the measured primary (DeepSeek) results.

Relative to the base student, standard OPD raises Avg@32 from 47.6 to 49.3 but lowers Pass@32 from 75.4 to 69.1. BCOD reaches $54.6/74.5$, improving both metrics over OPD and preserving more of the base student's Pass@32. Figure 3 shows the primary BCOD configuration above OPD at every evaluated k (bootstrap over the fixed 32 rollouts per prompt). The local EOPD reproduction reaches $45.6/68.7$. In the Qwen3-instruct setup, BCOD reaches $56.3/74.2$ versus $55.4/72.3$ for OPD. Table 2 examines the two BCOD components under a matched ablation protocol.

3.3 Observed solution-support transitions

Scalar Pass@k does not reveal whether post-training gains come from newly covered problems or whether they conceal losses on problems already covered by the base student. We therefore align all 857 prompts by dataset, question, and answer, and define a model as *observed-solved* on a problem if at least one answer in its fixed set of 32 rollouts is correct. For base indicator B_i and post-training indicator M_i, we report retained ($B_i{=}1, M_i{=}1$), newly solved ($B_i{=}0, M_i{=}1$), and forgotten ($B_i{=}1, M_i{=}0$) problems.

Figure 4 shows the observed pattern in the original fixed rollout set. Relative to OPD, BCOD retains more observed-solved problems (74.67% vs. 69.79%) and forgets fewer (2.82% vs. 7.69%). Newly solved support is also higher for BCOD (1.52% vs. 0.48%), although the absolute count is small. The same direction holds under micro aggregation: BCOD forgets 15 of 857 problems, compared with 26 for OPD. Prefix checks at $K \in \{8, 16, 32\}$ preserve the direction of the forgotten-support contrast, and the matched step-200 runs show 15 forgotten problems for BCOD, compared with 25 for OPD.

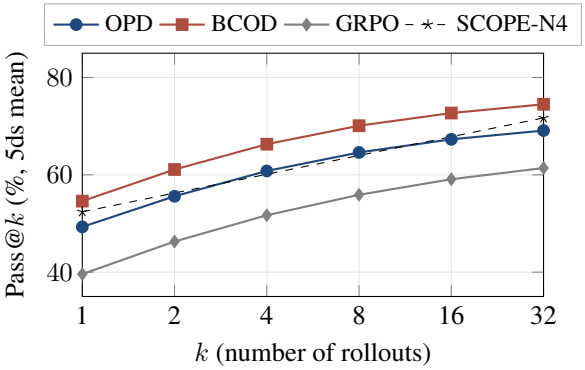


Figure 3: Pass@k scaling on the 5-dataset average.

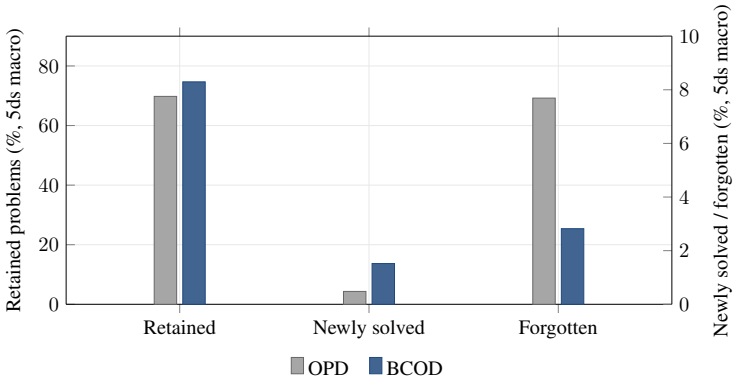


Figure 4: Observed problem-level support transitions from the matched local base student. Bars are unconditional percentages macro-averaged over five datasets; jointly unsolved problems are omitted. A problem is observed-solved if any answer in its fixed 32-rollout set is correct; the left axis (0–90) reads retained support and the right axis (0–10) reads newly solved and forgotten support, so the two axes have different scales. In micro counts, OPD retains/learns/forgets $683/7/26$ problems versus $694/13/15$ for BCOD, which has no more forgotten problems on any dataset.

An independent evaluation with a newly generated set of 32 rollouts per prompt preserves the positive BCOD–OPD ordering in Avg@32 and Pass@32, while the forgotten-support difference is smaller. The thresholded transition counts are therefore sensitive to the particular rollout set.

As an operator-only check at a matched optimization horizon with routing disabled, changing OPD to the clip-only ablation reduces the number of forgotten problems from 25 to 20 and the macro percentage from 6.40% to 5.08%. This routing-off contrast isolates attenuation under the matched ablation protocol; the full-system comparison also changes routing. The complete breakdown is reported in Appendix C.

3.4 Component ablation

Table 2 varies suppress-half attenuation and dual-path routing independently under a shared data and optimization protocol. Values are Avg@32 / Pass@32.

Adding attenuation changes OPD to Clip-only by $+2.1/+2.1$ with routing disabled and Routing-only to Full BCOD by $+2.2/+2.0$ with routing enabled. Adding routing yields $+2.2/+0.9$ without attenuation and $+2.3/+0.8$ with it. Both components coincide with higher Avg@32 and Pass@32 under this protocol.

As additional logged diagnostics, entropy is higher for Clip-only than for OPD (0.433 vs. 0.327), while the correct ratio is higher for Full BCOD than for Clip-only (0.442 vs. 0.333).

Method	suppress attenuation	dual-path routing	Avg@32 / Pass@32
OPD	off	off	49.0 / 69.6
Clip-only	on	off	51.1 / 71.7
Routing-only	off	on	51.2 / 70.5
Full BCOD (λ=0.5)	on	on	53.4 / 72.5
Full BCOD (primary, λ=0.3)	on	on	**54.6 / 74.5**

Table 2: Matched 2×2 component ablation. The four matched variants are compared at step 200 under a shared data and optimization protocol; attenuation-enabled rows use $\lambda = 0.5$. The bottom row is the primary $\lambda = 0.3$ run from Table 1 (also at step 200), shown for reference; its contrast with the $\lambda = 0.5$ row is descriptive rather than an isolated λ effect (Section 3.7).

Control	coefficient allocation	Avg@32 / Pass@32
Fixed-α Uniform (measured)	all tokens $\times 0.4033$	51.6/72.9
Branch-shuffled, L1-matched	permuted attenuation mask	50.9/71.4
Reciprocal, L1-matched	boost attenuated, suppress retained	49.5/69.5
Full BCOD (measured)	boost $\times 1$, suppress $\times \lambda$	54.6/74.5

Table 3: Directionality controls in the primary setup. The branch-shuffled and reciprocal controls retain routing and match BCOD's per-batch token-coefficient L1, not its parameter-gradient norm.

3.5 Directionality controls

Measured fixed-scale control. The fixed-α Uniform-scale control in Table 3 applies the same multiplier to every active OPD token. We calibrate this multiplier once on the first training batch as

$$\alpha = \frac{\sum_{t \in \mathcal{D}_1} |w_t R_t^{\mathrm{BCOD}}|}{\sum_{t \in \mathcal{D}_1} |w_t A_t|} = 0.4033, \tag{9}$$

where $\mathcal{D}_1$ is the set of active response tokens after masking and routing, w_t is the inherited routing weight, and R_t^{BCOD} is the coefficient in Eq. (3). We then keep α fixed throughout training. This matches BCOD's token-coefficient L1 on the calibration batch, not parameter-gradient norm. The control obtains $51.6/72.9$, whereas Full BCOD obtains $54.6/74.5$. Full BCOD is higher than all three controls on both metrics in the reported evaluations.

Sign-location controls. We additionally define two single-configuration controls. For *branch-shuffled* attenuation, the binary mask $m_t = \mathbf{1}\{A_t \leq 0\}$ is randomly permuted only among active tokens within each routed incorrect response, fixed across optimizer epochs for that batch, and applied as $A_t[1 - (1 - \lambda)\widetilde{m}_t]$. This keeps the number of attenuated tokens in every response but deliberately breaks the connection between attenuation and the sign of A_t; a detached batch scalar then matches its total coefficient L1 to BCOD. For the *reciprocal* control, the raw coefficient is $A_t[\lambda \mathbf{1}\{A_t > 0\} + \mathbf{1}\{A_t \leq 0\}]$, after which a detached batch scalar again matches BCOD's coefficient L1. In plain terms, the shuffled control shrinks the same number of tokens but chooses their locations at random, while the reciprocal control spends relatively more update mass on suppression than on boost. These interventions test sign alignment and branch allocation, not whether an individual suppressed token is a valid alternative solution.

3.6 Token-signal and generation-behavior analysis

Token-signal composition. To quantify how often the two reverse-KL branches occur, we aggregate token-gap statistics over steps 50–200 of the measured fixed-scale control in the primary setup: 76.53% of response tokens have non-positive teacher–student gaps, and the suppression branch accounts for 83.83% of the coefficient L1 among tokens active on the distillation path. Before correctness routing, correct trajectories contain 36.24% of suppression tokens and 30.56% of raw suppression L1; these trajectories subsequently enter the SFT path rather than the distillation path. Thus, suppressive feedback is not a rare boundary event: it constitutes most of the active sampled-token reverse-KL mass in this configuration.

Stage	Statistic	OPD	BCOD
Training	Logged policy entropy	0.3284	0.5382
Training	Batch correct fraction	0.3210	0.4584
Training	Mean response length (tokens)	7309.6	6632.4
Training	Length-cap ratio (%)	67.24	52.31

Table 4: Training diagnostics in the primary setup; values are means of 151 logged steps (50–200), with the correct fraction on a 0–1 scale. Saved-generation (evaluation) diagnostics are reported in Table 7.

Method	negative-half mechanism	Avg@32	Pass@32
GRPO (pure RL)	(none)	39.6	61.4
AOPD (no warm-up)	forward-KL imitation	N/A	N/A
AOPD (SFT warm-up only)	supervised warm-up	50.5	71.5
AOPD (warmed)	forward-KL after SFT	52.5	72.5
BCOD	λ-relaxed reverse-KL	**54.6**	**74.5**

Table 5: Comparison among GRPO, AOPD, and BCOD in the DeepSeek setting. The GRPO row is the 200-step diagnostic endpoint, distinct from the earlier GRPO checkpoint in Table 1. The prescribed SFT warm-up stabilizes AOPD and improves over warm-up alone; BCOD obtains the highest Avg@32/Pass@32 pair among the methods listed without a separate warm-up stage. Cold-start AOPD collapses before benchmark evaluation.

Training and generation behavior. Table 4 adds training diagnostics to the endpoint metrics: over steps 50–200, BCOD has higher policy entropy and batch correctness than OPD, with shorter responses and fewer length-cap hits. On a shared fresh $K = 32$ evaluation, BCOD responses are also shorter, less often near the cap, more often close the reasoning segment, and less often have low-diversity tails (Appendix A, Table 7).

These behavioral differences accompany the endpoint gains.

3.7 Evaluated λ settings

The evaluated $N{=}1$, full-data BCOD configurations under dual-path routing comprise the primary $\lambda = 0.3$ run ($54.6/74.5$; logged policy entropy 0.545) and the $\lambda = 0.5$ component-ablation run ($53.4/72.5$; entropy 0.408). Because these runs belong to different comparison protocols, their difference is descriptive rather than an isolated λ effect. The $\lambda = 0$ boost-only endpoint collapses mid-training.

3.8 Compute and the $N{=}1$ advantage

Because BCOD's clip is group-free, the operator remains defined at $N{=}1$ (Section 2.5). Under equal data coverage of one full epoch over DeepMath, BCOD at $N{=}1$ generates 103k rollouts ($\sim$691M tokens), one-quarter of SCOPE at $N{=}4$ (412k rollouts, $\sim$2763M tokens) and one-eighth of SCOPE at $N{=}8$ (824k, $\sim$5527M). This compares generation volume rather than end-to-end wall-clock efficiency; per-step phase timings for the three configurations appear in Appendix A (Table 6).

3.9 Comparison with AOPD: robustness without warm-up

AOPD (Jia et al., 2026) is a concurrent asymmetric token-level OPD that, on the negative half ($A_t \leq 0$), *strengthens* the teacher's pull by switching to a forward-KL imitation term on the teacher's top-K support. Its stated problems are optimization inefficiency (high variance, vanishing gradient, exploration black hole), and it reports Pass@1/4/8 rather than our 5ds Pass@32 protocol. We evaluate both cold-start AOPD and its prescribed SFT-warm-up recipe; Table 5 reports the resulting accuracy and coverage, with training-dynamics diagnostics deferred to Appendix A.

AOPD without SFT warm-up fails to produce an evaluable endpoint: accuracy decreases from 46.5% to 2.9%, entropy decreases from 0.29 to 0.12, and 69% of responses reach the maximum length. This

difference is objective-level rather than a strict mirror image: AOPD replaces the non-positive region with teacher-support forward-KL imitation, whereas BCOD attenuates the original sampled-token reverse-KL suppress update. The warm-up-only and warmed rows report performance before and after the AOPD stage under the same 5ds Avg@32/Pass@32 protocol.

4 Related Work

RLVR, diversity, and sample polarity. RLVR can improve reasoning accuracy while reducing large-k coverage. Prior work distinguishes correct from incorrect trajectories on Pass@k (Zhu et al., 2026), shapes verifier-derived advantages asymmetrically (Tang et al., 2026), or changes the divergence used to regularize RL (Li et al., 2026a); polarity is there defined through correctness or RL advantages, whereas BCOD uses the sign of a token's realized teacher–behavior gap in distillation.

OPD foundations and analyses. OPD trains on student-generated prefixes with dense teacher supervision (Agarwal et al., 2024), one implementation using a sampled-token reverse-KL signal (Lu & Lab, 2025); analyses distinguish prefix sources and KL directions (Zhao et al., 2026) and identify failure settings (Li et al., 2026b). BCOD keeps this setting, changing only the coefficient of non-positive token gaps.

Routing and gating. SCOPE routes correct rollouts to weighted SFT and incorrect rollouts to weighted KL (Zheng et al., 2026); RG-OPD gates trajectories by reward and teacher–student sequence-likelihood agreement (Akhondzadeh et al., 2026), and FiRe-OPD combines trajectory filtering with token reweighting (Li et al., 2026c). The evaluated system inherits SCOPE's routing and weights; BCOD changes only the token coefficient on the distillation path.

Token-level objectives. AOPD replaces the non-positive region with forward-KL imitation over the teacher's top-K support (Jia et al., 2026); EOPD uses teacher-support forward KL at high-entropy tokens (Jin et al., 2026); KL-mixing adapts the divergence (Zhao et al., 2026); and vOPD reduces Monte Carlo variance with a control-variate baseline while preserving the expected reverse-KL gradient (Oh et al., 2026). BCOD keeps the sampled-token loss form but scales its non-positive coefficient by a fixed λ, changing the expected update rather than only its variance.

5 Conclusion

BCOD attenuates non-positive sampled-token reverse-KL gaps while preserving positive gaps. At the on-policy point, its expected update descends a convex asymmetric f-divergence with the same teacher-matching optimum as reverse KL. In the primary five-dataset setup, BCOD reaches 54.6 Avg@32 and 74.5 Pass@32, versus 49.3 and 69.1 for standard OPD; fixed-rollout analyses also show fewer observed losses of base-model solution support. Thus, the evaluated results support suppression-branch attenuation as a useful local modification.

Limitations

The evaluation covers the reported teacher–student pairs, data fractions, and benchmarks. AIME24 and AIME25 contain only 30 and 15 problems, respectively, increasing per-benchmark uncertainty. The Qwen3 transfer study uses instruct-tuned models, 25% of DeepMath, and a different repetition penalty, so it is not a direct replication of the primary setup. Only two positive λ settings were evaluated under different comparison protocols.

The divergence result applies to the conditional on-policy token update; Pass@k and entropy are measured outcomes of the complete training system. The matched ablation isolates attenuation under its protocol, while the full BCOD–OPD comparison also includes routing inherited from SCOPE. The external safety check (Appendix D) evaluates matched Qwen3 step-50 checkpoints rather than the later reasoning-result checkpoints. In that check, BCOD has lower observed unsafe rates than OPD on both benchmarks, while the Base checkpoint has lower rates than both post-trained checkpoints.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing and grammar checking of the manuscript. All experiments, figures, and tables, as well as the reported analyses, were conducted by the authors. We have reviewed all AI-assisted text and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

All teacher and student models and the DeepMath-103k training set are publicly available, and the training and evaluation configurations are reported in Section 3.1 and Appendix A, including checkpoint selection and data budgets for each comparison. The support-transition analysis, including prompt-level assignments, scorer fingerprints, and bootstrap procedures, is detailed in Appendix C; the accompanying artifact contains the analysis implementation. The safety-evaluation protocol, decoding settings, and judge classifiers are specified in Appendix D.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Mohammad Sadegh Akhondzadeh, Vijay Lingam, Atula Tejaswi, Chanakya Ekbote, Sujay Sanghavi, and Aleksandar Bojchevski. Reward-gated on-policy distillation. *arXiv preprint arXiv:2607.04037*, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, et al. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 138306–138322, 2026.

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. *arXiv preprint arXiv:2103.03874*, 2021.

Nan Jia, Haojin Yang, Xing Ma, Jiesong Lian, Shuailiang Zhang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Zequn Sun. Asymmetric on-policy distillation: Bridging exploitation and imitation at the token level. *arXiv preprint arXiv:2605.06387*, 2026.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

Long Li, Zhijian Zhou, Jiaran Hao, Jason Liu, Yanting Miao, Wei Pang, Xiaoyu Tan, Wei Chu, Zhe Wang, Shirui Pan, et al. The choice of divergence: A neglected key to mitigating diversity collapse in reinforcement learning with verifiable reward. In *International Conference on Learning Representations*, volume 2026, pp. 64742–64768, 2026a.

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

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. *arXiv preprint arXiv:2606.02684*, 2026c.

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.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Minjae Oh, Sangjun Song, Gyubin Choi, Yunho Choi, and Yohan Jo. Kl for a kl: On-policy distillation with control variate baseline. *arXiv preprint arXiv:2605.07865*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xinyu Tang, Yuliang Zhan, Zhixun Li, Wayne Xin Zhao, Zhenduo Zhang, Zujie Wen, Zhiqiang Zhang, and Jun Zhou. Rethinking sample polarity in reinforcement learning with verifiable rewards. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2928–2954, 2026.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, et al. Sorry-bench: Systematically evaluating large language model safety refusal. In *International Conference on Learning Representations*, volume 2025, pp. 59937–59973, 2025.

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

Anhao Zhao, Haoran Xin, Yingqi Fan, Junlong Tong, Wenjie Li, and Xiaoyu Shen. Decoupling kl and trajectories: A unified perspective for sft, dagger, offline rl, and opd in llm distillation. *arXiv preprint arXiv:2605.16826*, 2026.

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

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in llm reasoning. *Advances in Neural Information Processing Systems*, 38:126546–126573, 2026.

A Evaluation selection and optimization details

Within each teacher–student block, measured rows use the same local benchmark prompts, decoding configuration, and scorer. Data and optimization budgets are matched for the stated comparisons or marked where they differ. The primary Base row evaluates the untrained DeepSeek-R1-Distill-Qwen-1.5B student under the same protocol as the corresponding post-training runs. Full-data runs use DeepMath-103k; the SCOPE-budget comparison at $N = 4$ uses 11.6% of DeepMath; and the Qwen3-instruct transfer block uses 25%. Table 1 therefore supports comparisons under these local protocols, not exact reproduction of numbers obtained under different source-paper protocols.

In the primary block, the reported OPD checkpoint is at step 215 and the primary BCOD checkpoint is at step 200. The component ablation in Table 2 compares all four variants at step 200 and uses $\lambda = 0.5$ when attenuation is enabled; the table additionally lists the primary $\lambda = 0.3$ run for reference. The main-table GRPO row uses an earlier checkpoint ($52.0/75.2$); its step-200 diagnostic endpoint ($39.6/61.4$) appears in Figure 3 and Table 5. Pass@k values for $k < 32$ in Figure 3 are computed by bootstrap-sampling k of the 32 rollouts per prompt (1,000 resamples, seed 42). The batch-64 GRPO curve in Figure 5 is a separate diagnostic configuration.

Figure 5 plots the training-accuracy and policy-entropy trajectories behind these diagnostics, including AOPD's cold-start collapse and BCOD's stable trajectory.

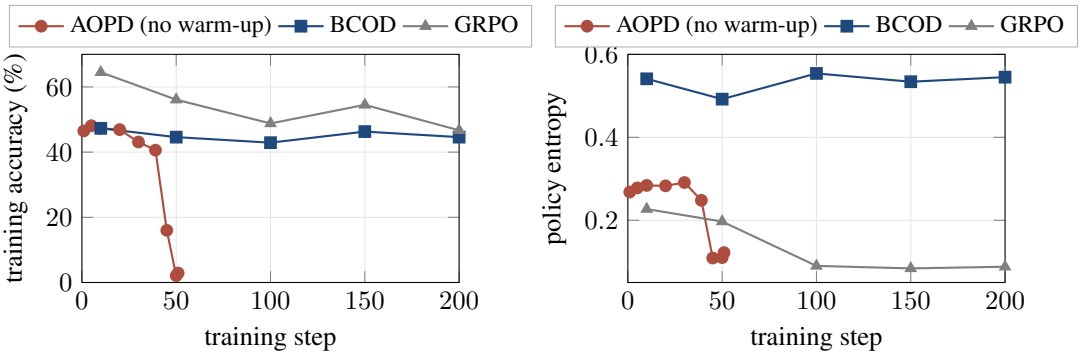


Figure 5: Diagnostic training trajectories in the no-warm-up DeepSeek setting. AOPD's accuracy decreases from 46.5% to 2.9%, and its entropy decreases from 0.29 to 0.12 around step 45; BCOD remains numerically stable over the plotted trajectory. The GRPO curve is a batch-64 diagnostic variant, distinct from the main-table GRPO run.

Table 6 reports per-step phase timings and equal-coverage generation counts for the three rollout configurations referenced in Section 3.8.

Config	seq/step	gen	teacher	update	step time (s)
N=8, batch= 60	480	128	118	90	350
N=4, batch= 60	240	122	65	54	251
N=1, batch= 480	480	124.5	125.4	232.2	530.8

Config	N	rollouts (full data)	gen tokens	rel. cost
BCOD N=1	1	103k	∼691M	1.0×
SCOPE N=4	4	412k	∼2763M	4.0×
SCOPE N=8	8	824k	∼5527M	8.0×

Table 6: Left: per-step timing (seconds) by configuration. Right: generated-rollout and token counts at equal data coverage. BCOD N=1 uses one-quarter as many generated rollouts as SCOPE N=4 and one-eighth as many as SCOPE N=8.

Table 7 reports the saved-generation diagnostics referenced in Section 3.6.

Optimization hyperparameters. The dual-clip PPO parameters in Eq. (6) are set to $\epsilon_\ell = 0.2$, $\epsilon_h = 0.2$, and $c = 3$ for all reported BCOD, OPD, and SCOPE runs, so every method shares the same

Evaluation statistic	OPD	BCOD
Mean re-tokenized length (tokens)	8858.5	7064.9
Near-cap rate (%)	25.77	16.82
Closed </think> rate (%)	74.91	83.87
Low tail 4-gram diversity (%)	3.38	1.58
Immediate repeated tail (%)	0.16	0.10

Table 7: Saved-generation diagnostics in the primary setup, summarizing 27,424 saved fresh responses per model on a shared K=32 evaluation. "Near cap" means a decoded response re-tokenizes to at least 16,368 tokens, not that the engine reported a length stop; the re-tokenized near-cap indicator is thus a proxy for response length near the generation cap rather than an engine finish reason. Low tail diversity means that fewer than half of the 4-grams in the final 512 tokens are unique; immediate repetition requires two identical adjacent terminal blocks of 16, 32, 64, 128, or 256 tokens.

clipping configuration and the reported comparisons isolate only the token coefficient. The trajectory weights w_i^T and w_i^S follow the correctness-routed trajectory weighting defined in the SCOPE training system (Zheng et al., 2026) and are inherited unchanged by BCOD; at N=1 the active sequence weight equals 1 (Section 2.5).

B Local interpretation and limits

BCOD changes the relative weighting of the boost and suppression branches in the policy-gradient term: the boost branch retains coefficient one, whereas the suppression branch receives coefficient λ. In the reported system, the BCOD coefficient is active only on the incorrect-trajectory path; correct trajectories use a λ-independent weighted SFT objective. The divergence proposition applies at a fixed prefix and on-policy point; entropy and Pass@32 are measured outcomes of the complete training system.

Figure 6 plots the logged policy-entropy trajectories underlying these outcomes: BCOD sustains higher training entropy than OPD and the GRPO diagnostic throughout the plotted steps.

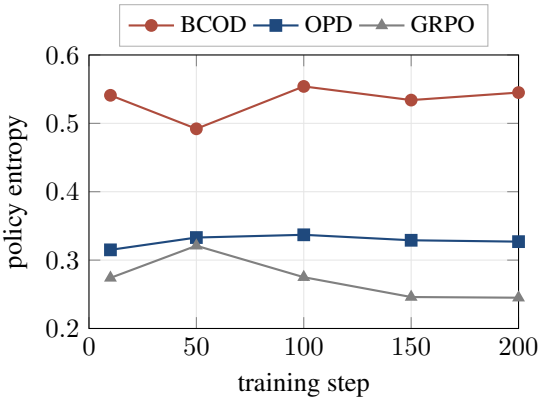


Figure 6: Diagnostic policy-entropy trajectories in the primary setup. The BCOD ($\lambda = 0.3$), OPD, and GRPO curves show logged values through step 200. The main table reports OPD at step 215 and GRPO at an earlier checkpoint; their plotted endpoints therefore differ from the reported checkpoints.

C Observed support-transition details

Table 8 provides the dataset-level breakdown corresponding to Figure 4. The entries denote retained, newly solved, and forgotten percentages among all prompts in each dataset. Jointly unsolved problems complete the four-way partition and are omitted for clarity. BCOD has a lower forgotten fraction on each dataset and under micro aggregation; thus, the macro result is not driven solely by the two small AIME sets.

Dataset (n)	OPD R/N/F	BCOD R/N/F	Forgotten counts O/B
AIME24 (30)	63.3 / 0.0 / 16.7	70.0 / 3.3 / 10.0	5 / 3
AIME25 (15)	46.7 / 0.0 / 13.3	60.0 / 0.0 / 0.0	2 / 0
AMC23 (40)	92.5 / 0.0 / 2.5	95.0 / 0.0 / 0.0	1 / 0
MATH500 (500)	97.2 / 0.2 / 0.8	97.6 / 0.2 / 0.4	4 / 2
Minerva (272)	49.3 / 2.2 / 5.1	50.7 / 4.0 / 3.7	14 / 10
Five-dataset macro	69.8 / 0.5 / 7.7	74.7 / 1.5 / 2.8	not applicable
857-prompt micro	79.7 / 0.8 / 3.0	81.0 / 1.5 / 1.8	26 / 15

Table 8: Observed support transitions at $K = 32$ (R/N/F: retained/newly solved/forgotten) under the primary evaluation protocol. Percentages describe one fixed rollout set and should not be interpreted as noise-free model-level labels.

Table 9 varies the fixed prefix length while preserving prompt alignment. Because the candidates at $K = 8$ and $K = 16$ are prefixes of the same fixed 32-rollout sets, this analysis measures threshold sensitivity under the primary evaluation protocol. The point estimate for the forgotten contrast favors BCOD at all three values.

K	OPD forgotten	BCOD forgotten	BCOD−OPD [95% CI]
8	7.20	5.33	-1.87 $[-3.81, -0.47]$
16	7.44	4.47	-2.97 $[-7.05, -0.27]$
32	7.69	2.82	-4.87 $[-9.17, -1.39]$

Table 9: Fixed-prefix sensitivity of observed forgotten support (five-dataset macro percentages; 5,000 prompt-paired bootstrap replicates per row). Prefix length changes both the base and post-training observed-support indicators.

Table 10 reports matched-horizon transitions for OPD and the primary BCOD run.

Method	Retained	Newly solved	Forgotten	ΔForgotten vs. OPD [95% CI]
OPD	71.09	0.37	6.40	reference
BCOD	74.67	1.52	2.82	-3.58 $[-7.43, -0.45]$

Table 10: Matched-horizon observed support transitions (five-dataset macro percentages; fixed $K = 32$ rollout sets). Pooled forgotten counts are 25 and 15 in row order. Intervals condition on each trained model and its fixed rollout set

Table 11 examines the token operator independently of inherited routing at a matched optimization horizon using the component-ablation configuration. Both OPD and attenuation-only training disable dual-path routing, making their contrast the relevant operator-only comparison. Attenuation-only training shifts all three point estimates in the favorable direction: more retained and newly solved problems, and fewer forgotten problems. Full BCOD in this ablation has the same micro transition counts as the attenuation-only variant; their macro percentages differ because the changed prompts occur in different datasets. The primary full-system transition result uses $\lambda = 0.3$, whereas this ablation uses $\lambda = 0.5$ when attenuation is enabled.

Method	Retained	Newly solved	Forgotten	Micro R/N/F counts
OPD (routing off)	71.09	0.37	6.40	684 / 6 / 25
Attenuation only (routing off)	72.41	1.19	5.08	689 / 9 / 20
Full BCOD (routing on)	71.81	1.81	5.67	689 / 9 / 20

Table 11: Matched-horizon observed transition analysis for attenuation-enabled rows (five-dataset macro percentages).

The accompanying artifact includes the analysis implementation, prompt-level assignments, scorer fingerprints, macro and micro summaries, and bootstrap intervals.

D EXTERNAL SAFETY-BOUNDARY DETAILS

We evaluate the Qwen3-1.7B Base checkpoint and the matched step-50 OPD and BCOD checkpoints. No checkpoint receives safety-specific training. All three models use the same Qwen3 tokenizer and chat template. The primary safety decoding protocol matches the Qwen3 evaluation setting: temperature 0.6, top-p 0.95, top-k 20, repetition penalty 1.08, and a maximum of $16{,}384$ generated tokens. We score the extracted final response with the official HarmBench behavior-conditioned classifier (Mazeika et al., 2024) and the official SORRY-Bench compliance classifier (Xie et al., 2025). Judge parse coverage is 100%.

Table 12 reports unsafe rates and BCOD–OPD differences. HarmBench aggregates four matched decoding seeds, while SORRY-Bench uses one frozen generation seed.

Benchmark	Base	OPD	BCOD	BCOD$-$OPD
HarmBench standard (4×159)	38.1	41.8	41.0	-0.8
SORRY-Bench base (450)	75.3	80.4	77.6	-2.9

Table 12: Official unsafe rates (%) for the Qwen3 safety-boundary evaluation. Lower is better. HarmBench averages four decoding seeds; SORRY-Bench uses one frozen seed.

BCOD has lower observed unsafe rates than OPD on both benchmarks at the matched step-50 checkpoints. The Base checkpoint has lower rates than either post-trained checkpoint. The later checkpoints used for the main reasoning results are not included in this safety evaluation.

As a decoding-sensitivity check, we repeat the evaluation with repetition penalty 1.0 and a $32{,}768$-token cap, matching the primary DeepSeek safety diagnostic. HarmBench over three new seeds gives Base/OPD/BCOD rates of $33.5/46.1/44.4$, with BCOD$-$OPD -1.7 points. SORRY-Bench gives $73.3/80.7/79.3$, with BCOD$-$OPD -1.3 points. BCOD's observed unsafe rate remains lower than OPD's under this decoding protocol, although the absolute rates vary with decoding and seed.

The primary 16K runs have no extraction failures for Base or BCOD and one of 159 HarmBench responses without an extracted final answer for OPD; all SORRY-Bench responses yield extracted final answers. Under the 32K sensitivity protocol, the largest group-level extraction-failure rate is 1.9%. All failures remain in the denominator and are counted as unsafe. These rates satisfy the pre-specified 5% closure threshold.