WHEN TEACHERS CONFLICT WITH VERIFIERS: SUPPORT-PRESERVING ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) removes state mismatch by querying the teacher on student trajectories, yet it still treats the teacher distribution as valid at every visited state. In verifiable reasoning, an outcome verifier can accept a trajectory even when the teacher assigns one of its sampled tokens less probability than the behavior policy. We call this a *teacher–verifier conflict*: OPD's reverse-KL update suppresses behavior from a correct trajectory, and EOPD's forward-KL term can repeat the same error. We introduce Support-Preserving On-Policy Distillation (SP-OPD). On correct trajectories, SP-OPD requires the target probability of each sampled token to remain at least its behavior-policy probability. KL projection produces the closest feasible target in closed form. It changes the target only at conflict positions and preserves the teacher's pairwise probability ratios among alternatives. The projected target corrects both EOPD loss terms, while sequence-balanced reduction aligns the sampled-token loss with trajectory-level verification. Conflicts cover 41.57% of sampled tokens on correct trajectories from a trained EOPD policy and persist throughout training. Across 0.6B, 1.7B, and 4B Qwen3-Base students, SP-OPD improves both Avg@8 and Pass@8 at every scale. At 4B it reaches 43.92% Avg@8 and 64.91% Pass@8, gains of 1.26 and 4.20 points over EOPD.

1 INTRODUCTION

On-policy distillation (OPD) queries the teacher on states visited by the current student, removing the state-distribution mismatch of fixed distillation data (Agarwal et al., 2024; Lu et al., 2025). This solves where supervision is applied, but leaves a second question unresolved: does the teacher still provide the right update at that state? In verifiable reasoning, the answer can be no because dense teacher probabilities and a trajectory-level outcome verifier provide conflicting evidence about a sampled token.

Let the behavior policy p_b generate token y_t at state s_t, and let p_T denote the teacher distribution. We call token position t on a correct trajectory a *teacher–verifier conflict* when $p_T(y_t \mid s_t) < p_b(y_t \mid s_t)$. OPD's clipped reverse Kullback–Leibler (KL) term uses

$$A_t^T = \log p_T(y_t \mid s_t) - \log p_b(y_t \mid s_t) \tag{1}$$

as its detached per-token advantage. When $A_t^T < 0$ on a correct trajectory, gradient descent reduces the probability of a token from a verified solution. The state is on-policy, but the target points against verified success. This is a target-validity failure. Conflicts cover 41.57% of sampled tokens on correct trajectories from a trained EOPD policy and persist throughout 4B training.

Outcome verification supplies asymmetric information. Acceptance certifies the complete sampled trajectory, whereas rejection does not locate the token that caused failure. Existing controls use entropy, reward, or advantage sign to decide where and how strongly to distill (Jin et al., 2026; Akhondzadeh et al., 2026; Xu et al., 2026); the active target remains teacher-defined. Even masking a negative advantage closes only the reverse-KL path, leaving EOPD's forward-KL target free to suppress the same sampled token.

This asymmetry determines the design of Support-Preserving On-Policy Distillation (SP-OPD). For every sampled token on a correct trajectory, SP-OPD requires the target probability to remain

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.

at least its behavior-policy probability. KL projection selects the closest feasible target, raising a conflicting token exactly to the boundary while preserving the teacher's pairwise probability ratios among alternatives. The projected target closes both loss paths. Because the verifier labels whole trajectories, SP-OPD also averages the clipped surrogate within each response before averaging across responses. The target encodes what the verifier establishes; the reduction preserves the granularity at which it is established.

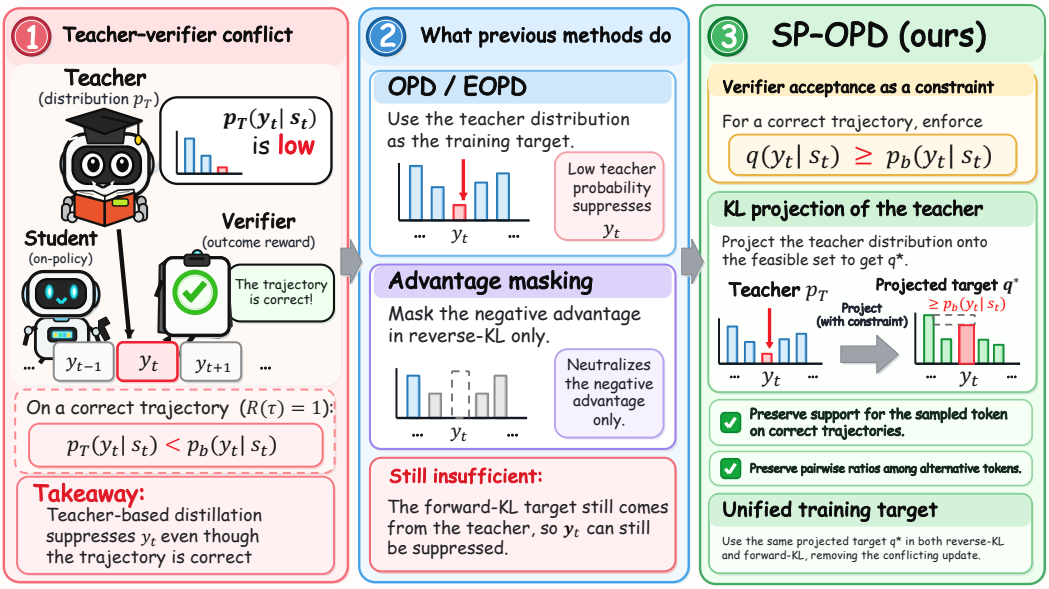


Figure 1: Teacher–verifier conflict and its resolution. (1) The teacher assigns a sampled token from a correct trajectory less probability than the behavior policy, so OPD suppresses it. (2) Advantage masking neutralizes reverse KL but leaves EOPD's conflicting forward-KL target intact. (3) SP-OPD enforces $q(y_t \mid s_t) \geq p_b(y_t \mid s_t)$ through KL projection and uses the resulting target in both loss terms.

The experiments test every link in this mechanistic account. Conflicts recur under a trained EOPD policy and throughout SP-OPD training. Gradient analysis identifies the forward-KL path left open by scalar masking, and component ablations support the resulting predictions. Across three student scales, SP-OPD improves both macro metrics, with larger gains in Pass@8—the metric of valid-solution coverage.

Our contributions are:

- We expose teacher–verifier conflict as a target-validity failure in verifier-assisted OPD and show that it remains prevalent after EOPD training.
- We derive a closed-form KL projection that removes the conflict from both EOPD loss terms while retaining the teacher's pairwise probability ratios among alternative tokens.
- We show that the failure survives EOPD training, that repairing both target direction and trajectory weighting produces the full gain, and that the resulting method improves both response accuracy and solution coverage at all three student scales.

2 Related Work

On-policy sampling solves state mismatch, not target validity. Classical KD matches teacher distributions; sequence-level and reasoning distillation imitate fixed teacher traces (Hinton et al., 2015; Kim & Rush, 2016; Ho et al., 2023; Mukherjee et al., 2023). On-policy methods remove the resulting state mismatch by querying the teacher on evolving student trajectories (Gu et al., 2024; Ko et al., 2024; Agarwal et al., 2024). OPD estimates reverse KL from teacher–student log-probability differences on these states (Lu et al., 2025); EOPD adds entropy-selected, top-k forward KL (Jin et al.,

2026). These advances improve where supervision occurs and how it is estimated, while retaining the teacher distribution as the token-level destination. SP-OPD addresses the remaining failure: the state is on-policy, but the destination conflicts with verified success.

Adaptive supervision leaves target validity unresolved. Recent objectives improve OPD by changing its estimator, selecting informative positions or reasoning steps, constraining update magnitude, or transferring hidden representations (Li et al., 2026; Fu et al., 2026; Xie et al., 2026; Yang et al., 2026b; Jiang & Ferraro, 2026; Sun et al., 2026; Ko et al., 2026; Xing et al., 2026; Wu et al., 2026). Complementary approaches rescale the implicit reward or gate supervision by advantage sign, verifier reward, or training progress (Yang et al., 2026c; Xu et al., 2026; Akhondzadeh et al., 2026; Qu et al., 2026; Yang et al., 2026a). Both lines govern the strength and location of teacher supervision. Teacher–verifier conflict instead requires changing the target of that supervision. SP-OPD makes the smallest KL change that satisfies the verifier-implied probability constraint.

Verification supplies asymmetric evidence. Sampling and self-training exploit multiple reasoning trajectories (Wei et al., 2022; Wang et al., 2023; Zelikman et al., 2022), and scaling such trajectories improves mathematical reasoning (Lewkowycz et al., 2022). Outcome and process verifiers add an independent correctness signal (Cobbe et al., 2021; Lightman et al., 2023). RLVR directly optimizes this signal (Shao et al., 2024; Yu et al., 2025); reward-aware distillation uses it to filter or reweight teacher updates (Akhondzadeh et al., 2026; Qu et al., 2026; Yang et al., 2026a). The key distinction is that acceptance certifies a trajectory's outcome, while rejection does not reveal the failed token (Qian et al., 2026). SP-OPD converts this asymmetry into a one-sided target constraint and retains dense teacher supervision wherever the verifier supplies no token-level preference.

3 Support-Preserving On-Policy Distillation

3.1 Why an on-policy teacher can still be the wrong target

At iteration u, the behavior policy $p_b = p_{\theta_{u-1}}$ samples $\tau = (y_1, \ldots, y_L)$ for prompt x, visiting states $s_t = (x, y_{<t})$. The teacher supplies $p_T(\cdot \mid s_t)$, while a deterministic outcome verifier returns $R(\tau) \in \{0, 1\}$ for the completed trajectory; $R = 1$ denotes a trajectory with a verified correct outcome, called *correct* below. The behavior policy stays fixed during the iteration and is refreshed before the next trajectories are sampled.

Write $a_t = p_T(y_t \mid s_t)$ and $b_t = p_b(y_t \mid s_t)$. A correct trajectory contains a teacher–verifier conflict wherever $a_t < b_t$: OPD's per-token advantage is negative, so distillation suppresses the sampled token despite the positive outcome certificate. On-policy sampling cannot remove this contradiction because it concerns the target at the sampled state, not the state distribution.

Acceptance establishes the sequence-level fact needed for a one-sided constraint: the target should not make a sampled token from that correct trajectory less likely than under the policy that produced it. The weakest such constraint is $q(y_t \mid s_t) \geq b_t$. Rejection supplies no token-local diagnosis, so incorrect trajectories retain the teacher distribution. Because OPD acts at each visited state, the constraint applies separately at every position.

3.2 Minimal correction by KL projection

At each token position on a correct trajectory, the verifier constraint defines the feasible set

$$\mathcal{Q}_t^+ = \{q \in \Delta^{|\mathcal{V}|-1} : q(y_t) \geq b_t\}. \tag{2}$$

Let $\mathcal{Q}_t = \mathcal{Q}_t^+$ when $R = 1$ and $\mathcal{Q}_t = \Delta^{|\mathcal{V}|-1}$ when $R = 0$. The constraint leaves the remaining probability mass unspecified. To retain the teacher's information, we choose its closest feasible distribution in KL divergence (Csiszar, 1975):

$$q_t^* = \arg\min_{q \in \mathcal{Q}_t} D_{\mathrm{KL}}(q \,\|\, p_T(\cdot \mid s_t)). \tag{3}$$

Proposition 1 (Closed-form KL projection)**.** *If the teacher distribution satisfies the constraint, $q_t^* = p_T$. Otherwise, assuming $p_T(v \mid s_t) > 0$ for every token v, the unique solution is*

$$q_t^*(y_t) = b_t, \qquad q_t^*(v) = \frac{1 - b_t}{1 - a_t} p_T(v \mid s_t), \quad v \neq y_t. \tag{4}$$

Consequently, the projection preserves every pairwise probability ratio among alternative tokens. The same q_t^* also minimizes $D_{\mathrm{KL}}(p_T(\cdot \mid s_t)\|q)$ over $\mathcal{Q}_t$.

For fixed $z = q(y_t)$, KL minimization over the alternatives gives the proportional rescaling in Equation 4 and reduces the objective to $D_{\mathrm{KL}}(\mathrm{Bern}(z)\|\,\mathrm{Bern}(a_t))$. Minimizing it over $z \geq b_t$ yields $z = \max(a_t, b_t)$; Appendix A gives the full derivation.

The behavior policy fixes the boundary, so projection needs no margin or mixture coefficient. It raises only the sampled token at a conflict position and preserves the teacher's pairwise probability ratios among alternatives: $q_t^*(v)/q_t^*(w) = p_T(v \mid s_t)/p_T(w \mid s_t)$ for $v, w \neq y_t$. At a conflict, its KL cost is exactly $D_{\mathrm{KL}}(\mathrm{Bern}(b_t)\|\,\mathrm{Bern}(a_t))$.

3.3 Why both EOPD loss terms need the projected target

Let $\rho_t = p_\theta(y_t \mid s_t)/p_b(y_t \mid s_t)$ and define the detached per-token advantage determined by the projected target,

$$A_t^q = \log q_t^*(y_t) - \log b_t. \tag{5}$$

EOPD transmits the same teacher preference through two loss paths. Its clipped reverse-KL surrogate uses the sampled-token log-probability difference, while forward KL matches a normalized vocabulary distribution. Correcting only the scalar advantage leaves the second path open. SP-OPD derives both terms from one projected target. For trajectory i, we substitute A_{it}^q into EOPD's PPO-style clipped reverse-KL surrogate (Schulman et al., 2017),

$$\ell_{it}^{\mathrm{clip}} = \max[-\rho_{it}A_{it}^q, -\,\mathrm{clip}(\rho_{it}, 1-\epsilon, 1+\epsilon)A_{it}^q]\,. \tag{6}$$

At every conflict position $q_t^*(y_t) = b_t$, so $A_t^q = 0$; at every non-conflict position the original EOPD clipped reverse-KL term is recovered exactly.

Proposition 2 (Nonnegative advantage on correct trajectories)**.** *For every sampled token on a correct trajectory, $A_t^q \geq 0$; sampled tokens on incorrect trajectories retain A_t^T. At a teacher–verifier conflict the corresponding advantage is exactly zero.*

Equivalently, $A_t^q = \max(A_t^T, 0)$ on a correct trajectory. Projection extends this scalar rule to a normalized vocabulary distribution and therefore also specifies the forward-KL target.

These constraints also imply a likelihood bound for every observed correct trajectory:

$$\prod_{t=1}^{L} q_t^*(y_t) \;\geq\; \prod_{t=1}^{L} p_b(y_t \mid s_t) = p_b(\tau \mid x). \tag{7}$$

Thus the projected targets assign the correct trajectory at least its behavior-policy likelihood.

EOPD activates forward KL when $H(p_T(\cdot \mid s_t)) \geq h_0$ and renormalizes the teacher over its top-k tokens. SP-OPD instead renormalizes q_t^* over a set S_t^k of the same size. This is EOPD's set unless a conflicting sampled token lies outside it, in which case that token replaces the least-probable member. Projection thus supplies a verifier-consistent distribution even when it sets $A_t^q = 0$. Writing $\widetilde{q}_{it}(v) = q_{it}^*(v)/\sum_{w \in S_{it}^k} q_{it}^*(w)$, the token-level forward-KL term is

$$\ell_{it}^{\mathrm{fwd}} = \mathbf{1}\{H(p_T(\cdot \mid s_{it})) \geq h_0\} \sum_{v \in S_{it}^k} \widetilde{q}_{it}(v) \log \frac{\widetilde{q}_{it}(v)}{p_\theta(v \mid s_{it})}. \tag{8}$$

The entropy threshold remains a function of the teacher distribution: projection changes the target but not which positions EOPD selects.

The projected target must also reverse the local forward-KL direction. Let $g_{it}(y_t)$ denote the student's logit for the sampled token. At the start of a training iteration, $p_\theta = p_b$ on the sampled trajectory.

Proposition 3 (Forward-KL gradient at the behavior policy)**.** *At a conflict position where EOPD activates forward KL, evaluated at the behavior policy $p_\theta = p_b$,*

$$\frac{\partial \ell_{it}^{\mathrm{fwd}}}{\partial g_{it}(y_t)} = b_{it} - \widetilde{q}_{it}(y_t) \leq 0. \tag{9}$$

The gradient-descent direction of SP-OPD*'s forward-KL term at the behavior policy therefore weakly increases the sampled token's logit. Write $\widetilde{p}_T$ for EOPD's forward-KL target obtained by renormalizing over its top-k token set. Its corresponding derivative is $b_{it} - \widetilde{p}_T(y_t)$ when y_t is retained and $b_{it} > 0$ when it lies outside the top-k token set, so EOPD's forward-KL term can oppose the same verifier constraint after the negative per-token advantage has been masked.*

The result follows because SP-OPD's forward-KL target has $\widetilde{q}_{it}(y_t) = b_{it} / \sum_{v \in S^k_{it}} q^*_{it}(v) \geq b_{it}$; Appendix A.1 gives the gradient derivation. Projection therefore turns the verifier-implied probability constraint into the actual local direction of forward KL. Scalar advantage masking leaves that direction unchanged.

At a conflict position	EOPD	Advantage masking	SP-OPD
Per-token advantage	$A^T_t < 0$	0	0
Clipped reverse-KL effect	suppress	neutral	neutral
Forward-KL target	$\widetilde{p}_T$	$\widetilde{p}_T$	$\widetilde{q}$
Forward-KL effect	may suppress	may suppress	weakly increases
Enforces $q(y_t) \geq b_t$	no	no	**yes**

Table 1: Loss behavior at a teacher–verifier conflict. Scalar masking repairs the clipped reverse-KL term but retains EOPD's forward-KL target; KL projection makes both terms consistent with the verifier constraint.

3.4 Why trajectory evidence requires sequence-balanced reduction

Projection aligns target content with verifier evidence; loss reduction must also align with its granularity. Token-level reduction weights each verifier label in proportion to response length. SP-OPD averages the clipped surrogate within each response and then averages responses equally. Let $m_{it} = 1$ for a non-padding token and zero otherwise, and let $\mathcal{L}^{\mathrm{EOPD}}_{\mathrm{fwd}}(q^*)$ denote EOPD's forward-KL term, activated by its entropy threshold, with q^* as its target and its original reduction. With EOPD's forward-KL coefficient α, the objective is

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^{B} \frac{\sum_t m_{it} \ell^{\mathrm{clip}}_{it}}{\sum_t m_{it}} + \alpha \mathcal{L}^{\mathrm{EOPD}}_{\mathrm{fwd}}(q^*). \tag{10}$$

We use EOPD's $\alpha = 1$. The first term gives every trajectory equal weight without introducing another coefficient. Its advantages inherit the verifier label, whereas forward KL remains a statewise distribution-matching term and retains EOPD's reduction. We otherwise follow the sampling, teacher-evaluation, and optimization settings described by EOPD.

On-policy update. Each iteration (i) samples one trajectory per prompt from the current student, (ii) obtains a verifier label and teacher distributions on the visited prefixes, (iii) projects correct-trajectory targets at conflict positions, and (iv) updates the student with the sequence-balanced clipped surrogate and projected forward-KL target. The updated student generates the next batch, so SP-OPD remains on-policy.

4 Experiments

The experiments test the proposed teacher–verifier conflict mechanism. The explanation makes three linked predictions: conflict remains prevalent after EOPD training, correcting both loss paths outperforms scalar masking, and matching the reduction to trajectory-level evidence improves solution coverage. We connect final accuracy to these mechanism-level predictions rather than treating benchmark gains as a standalone outcome.

Setting. We distill a frozen Qwen3-8B teacher in non-thinking mode into Qwen3-0.6B-Base, Qwen3-1.7B-Base, and Qwen3-4B-Base students (Yang et al., 2025). The 0.6B and 1.7B students train on the 7,500-example MATH-lighteval split (7,493 after length filtering) for three epochs and 174 training iterations. The 4B student trains on the 14,116-example English DAPO-Math split (14,108 after filtering) for two epochs and 220 training iterations (Yu et al., 2025). Every iteration

samples one fresh trajectory per prompt and evaluates the teacher once at the visited states. Batch size is 128 with four optimizer minibatches; Appendix Table 6 gives the full configuration.

Comparisons. The baselines test target validity against four competing explanations. KD provides a fixed-teacher imitation reference, GRPO provides a verifier-reward control, and OPD provides an on-policy sampling control (Hinton et al., 2015; Shao et al., 2024; Lu et al., 2025). EOPD provides the direct reference because SP-OPD corrects its two teacher-driven loss paths (Jin et al., 2026). ExOPD and the additional controls test whether reward rescaling, exploration, divergence weighting, or broader forward-KL placement can deliver the same accuracy–coverage balance (Yang et al., 2026c).

Evaluation. The benchmark suite contains MATH-500, AMC23, OlympiadBench, AIME24, and AIME25, with 500, 40, 675, 30, and 30 problems, respectively (Hendrycks et al., 2021; He et al., 2024; Mathematical Association of America, 2025). We sample eight responses per problem at temperature 1 and top-$p = 0.8$, with top-k disabled and a maximum response length of 8,192 tokens. Avg@8 is mean response accuracy; Pass@8 is the fraction of problems solved by at least one response. We score answers with the Qwen2.5-Math parser and symbolic equivalence checker (Yang et al., 2024). Macro Avg@8 and macro Pass@8 are the corresponding unweighted means across benchmarks.

4.1 Does correcting target conflict improve reasoning?

		Avg@8↑ / Pass@8↑					
Size	Method	MATH-500	AMC23	OlympiadBench	AIME24	AIME25	Macro
8B	Teacher	83.65/94.00	69.69/92.50	49.54/68.74	26.67/43.33	19.17/43.33	49.74/68.38
0.6B	Initial	38.38/70.40	17.81/52.50	12.48/35.70	0.83/6.67	0.42/3.33	13.98/33.72
	KD	47.80/69.60	23.43/52.50	18.15/36.59	2.19/6.67	0.83/6.67	18.48/34.41
	GRPO	**51.83**/74.40	25.25/55.00	18.91/34.96	**4.58**/10.00	0.83/**10.00**	20.28/36.87
	OPD	50.09/73.20	24.69/57.50	**20.15**/37.19	2.50/10.00	1.25/6.67	19.74/36.91
	ExOPD	50.55/**76.20**	29.06/57.50	19.43/**40.59**	3.75/**16.67**	0.83/3.33	20.72/38.86
	EOPD	50.63/74.60	**29.38/62.50**	19.78/40.30	2.92/10.00	**2.50/10.00**	21.04/39.48
	SP-OPD	51.60/75.60	29.06/**62.50**	19.94/**40.59**	3.33/13.33	**2.50/10.00**	**21.29/40.41**
1.7B	Initial	54.33/82.20	29.69/62.50	21.44/46.52	4.17/10.00	2.92/20.00	22.51/44.24
	KD	63.80/84.20	38.11/70.00	27.66/48.40	10.06/**20.00**	3.36/**16.67**	28.60/47.85
	GRPO	**68.83**/84.60	**40.62**/70.00	29.89/50.81	9.17/**20.00**	4.58/**16.67**	30.62/48.42
	OPD	67.76/84.80	39.06/70.00	30.09/51.56	8.33/**20.00**	**6.25/16.67**	30.30/48.61
	ExOPD	66.45/85.20	39.38/**75.00**	29.76/51.85	10.00/**20.00**	5.00/**16.67**	30.12/49.74
	EOPD	68.50/**87.80**	40.00/65.00	30.89/52.00	8.33/16.67	5.42/13.33	30.63/46.96
	SP-OPD	68.13/85.60	40.31/**75.00**	**31.30/53.93**	**10.83/20.00**	5.42/**16.67**	**31.20/50.24**
4B	Initial	49.05/87.80	29.69/77.50	26.48/56.74	8.75/23.33	8.33/23.33	24.46/53.74
	KD	74.73/92.20	48.13/**85.00**	37.33/62.52	12.50/30.00	12.08/23.33	36.95/58.61
	GRPO	**80.47**/91.00	58.44/75.00	43.37/60.74	14.85/30.00	12.66/26.67	41.96/56.68
	OPD	78.81/90.80	57.33/80.00	42.10/58.80	18.33/26.67	12.08/30.00	41.73/57.25
	ExOPD	76.08/92.00	54.37/**85.00**	41.06/61.19	**20.00/36.67**	12.92/36.67	40.88/62.30
	EOPD	79.45/91.60	56.25/77.50	43.44/**64.44**	18.33/**36.67**	15.83/33.33	42.66/60.71
	SP-OPD	79.23/**92.80**	**60.63/85.00**	**43.52**/63.41	17.92/**36.67**	**18.33/46.67**	**43.92/64.91**

Table 2: Five-benchmark comparison across student scales (%). Each cell reports Avg@8 / Pass@8. Bold and underlined values mark the best and second-best adaptation scores, respectively, within each student scale and benchmark; the initial students and frozen teacher provide references.

The cross-scale results are consistent with the intended likelihood-preserving effect. Relative to EOPD, SP-OPD improves Avg@8 by 0.25, 0.57, and 1.26 points at 0.6B, 1.7B, and 4B, while Pass@8 rises by 0.93, 3.28, and 4.20 points. At every scale, the larger Pass@8 gain is consistent with preserving at least one valid solution path for more problems. At 4B, SP-OPD reaches 43.92/64.91, the strongest macro pair in Table 2, and recovers more than half of EOPD's remaining Pass@8 gap to the 8B teacher.

4.2 CAN GENERIC OBJECTIVE CHANGES EXPLAIN THE GAIN?

If SP-OPD merely adds useful regularization or exploration, an objective that changes those quantities without resolving teacher–verifier conflict should match its accuracy–coverage balance. The controls below vary sampled-token weighting, divergence weighting, and the amount or location of forward KL while leaving the teacher-defined target intact.

Method	Avg@8 ↑	Pass@8 ↑
OPD (Lu et al., 2025)	30.35	47.87
GKD ($\lambda = 1$) (Agarwal et al., 2024)	29.95	47.29
OPD + entropy bonus (Schulman et al., 2017)	30.18	49.42
OPD + advantage shaping (Cheng et al., 2025)	29.76	<u>50.88</u>
OPD + forward KL at all positions	30.18	**51.13**
OPD + random-position forward KL	29.39	46.41
BD-KD (Amara et al., 2024)	29.99	48.99
AKL (Wu et al., 2024)	29.08	46.30
ToDi (Jung et al., 2025)	30.09	49.62
ExOPD (Yang et al., 2026c)	30.21	49.22
EOPD	<u>30.56</u>	45.70
SP-OPD	**31.17**	49.32

Table 3: Alternative explanations for SP-OPD's gain on the 1.7B student (four-benchmark macro %). The controls change exploration, weighting, divergence, or forward-KL placement without enforcing the verifier-implied target constraint. Bold and underlined values mark the best and second-best results. Appendix Tables 8– 9 give benchmark-level values and the GKD mixture-weight sweep.

SP-OPD is the only method that improves both metrics over EOPD, adding 0.61 Avg@8 and 3.62 Pass@8. Forward KL at every position attains 1.81 more Pass@8 than SP-OPD but loses 0.99 Avg@8, revealing the reliability cost of indiscriminate broadening. Conflict-specific target correction instead improves the two objectives together. The next comparison identifies which part of that correction produces the balance.

4.3 WHY MUST FORWARD KL USE THE PROJECTED TARGET?

The gradient analysis makes a falsifiable distinction. If conflict acted only through the clipped reverse-KL term, zeroing its negative advantage would be sufficient. If EOPD's forward-KL target carries the same contradiction, forward KL must use the projected target. Advantage masking tests the first explanation; KL projection tests the second.

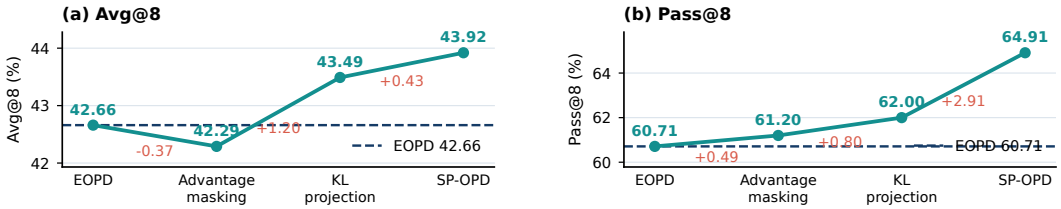


Figure 2: Stepwise construction of SP-OPD on the 4B student. Advantage masking repairs only the sampled reverse-KL signal; projection also repairs the forward-KL target; sequence balancing then matches the clipped surrogate to trajectory-level verification. Dashed lines mark EOPD.

Scalar masking falls below EOPD in Avg@8, showing that removing one negative advantage is insufficient. KL projection adds 1.20 Avg@8 and 0.80 Pass@8, as predicted when forward KL is the remaining conflict path. Because projection changes only conflict positions and recovers the teacher elsewhere, the comparison supports target validity as the relevant factor under fixed supervision volume.

Projection leaves one granularity mismatch. Correct trajectories are systematically shorter, so token-level reduction overweights unsuccessful trajectories. Table 4 quantifies this imbalance directly.

Verifier outcome	Responses (%)	Tokens (%)	Relative mean length
Correct	37.19	27.31	0.73×
Incorrect	62.81	72.69	1.16×

Table 4: Length imbalance in the 4B SP-OPD training trajectories. Relative mean length is normalized by the mean over all trajectories.

Although correct trajectories constitute 37.19% of responses, they contain only 27.31% of sampled tokens; an incorrect trajectory consequently receives 1.58× the average weight of a correct one. Sequence-balanced reduction restores one vote per trajectory and adds the final 0.43/2.91 Avg@8/Pass@8 points.

Configuration	MATH-500	AMC23	OlympiadBench	AIME24	AIME25	Macro
EOPD	79.45/91.60	56.25/77.50	43.44/64.44	18.33/36.67	15.83/33.33	42.66/60.71
+ Advantage masking	79.30/92.60	55.62/90.00	43.20/63.41	15.00/26.67	18.33/33.33	42.29/61.20
+ KL projection	79.40/93.40	59.06/80.00	43.59/63.26	17.92/30.00	17.50/43.33	43.49/62.00
+ Sequence-balanced reduction (SP-OPD)	79.23/92.80	60.63/85.00	43.52/63.41	17.92/36.67	18.33/46.67	**43.92/64.91**
+ Incorrect-trajectory constraint	79.72/92.60	60.62/87.50	43.67/64.44	16.67/30.00	18.75/36.67	43.89/62.24

Table 5: Benchmark-level cumulative 4B construction ablation (%). Each cell reports Avg@8 / Pass@8; each row repairs the next mismatch predicted by the teacher–verifier conflict analysis.

The final row explains why the constraint is one-sided. Applying it to incorrect trajectories reaches 43.89/62.24, below SP-OPD by 0.04 Avg@8 and 2.67 Pass@8. Acceptance provides a positive outcome certificate for the sampled sequence; rejection does not identify which token should be suppressed or protected. Treating the two outcomes symmetrically therefore injects a token-level claim that the verifier never made.

4.4 Is target conflict persistent enough to matter?

The mechanism matters only if conflict survives changes in the student and remains present after EOPD training. Both predictions hold.

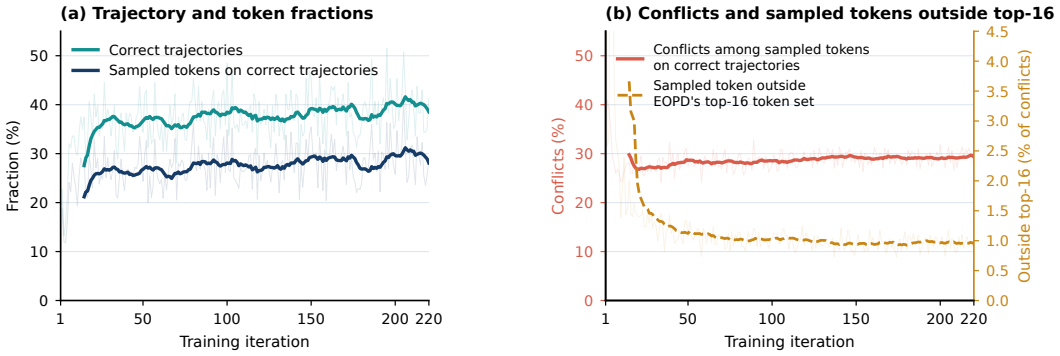


Figure 3: Training dynamics during 4B SP-OPD training. (a) Fraction of trajectories that are correct and fraction of all sampled tokens that occur on correct trajectories. (b) Fraction of sampled tokens on correct trajectories that are conflict positions (left) and fraction of conflict positions whose sampled token lies outside EOPD's top-16 token set (right). Dark curves are 15-iteration moving averages.

During 4B SP-OPD training, conflicts cover 28.73% of sampled tokens on correct trajectories (7.84% of all sampled tokens) and remain present through all 220 iterations. The target correction therefore acts repeatedly rather than only during initialization. Just 1.14% of these positions place the sampled token outside EOPD's top-16 set. Most conflicts therefore arise while the sampled token remains inside EOPD's token set: the failure is its assigned probability, precisely the quantity corrected by projection.

The target statistics connect prevalence to the predicted gradient. In the 1.7B run, projection activates on 8.08% of all sampled tokens, or 23.83% of tokens on correct trajectories, and raises sampled-token

probability by 0.13 on average. Among projected positions where EOPD also activates forward KL, 98.84% receive the conflicting EOPD direction derived in Proposition 3. Projection changes their mean sampled-token target mass from 0.27 to 0.43. Thus the theoretically identified second conflict path is also the typical one at positions where both terms act.

EOPD does not eliminate the failure from its own state distribution. On the same 288 DAPO-Math prompts, conflicts cover 42.44% of 17,120 sampled tokens on correct initial-student trajectories and 41.57% of 137,828 tokens on correct trajectories from the trained EOPD policy. EOPD changes the student while leaving the conflicting target relation prevalent.

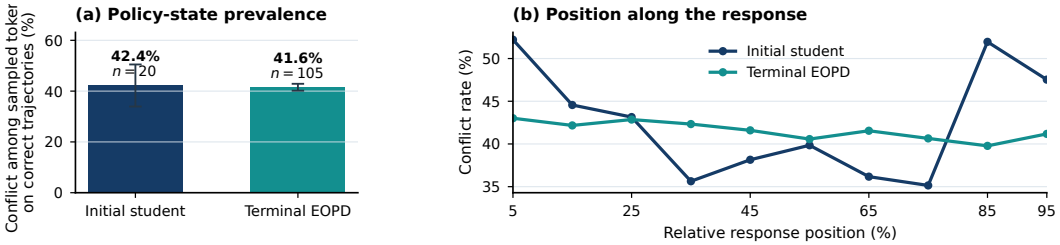


Figure 4: Teacher–verifier conflict on baseline-policy trajectories. (a) Fraction of sampled tokens on correct trajectories that are conflict positions, with trajectory-bootstrap 95% intervals; n is the number of correct, explicitly terminated trajectories. (b) The same fraction by relative response-position decile. Both policies use the same 288 prompts and training decoding.

5 Discussion and Limitations

The results reveal a target-validity failure beyond state coverage. OPD already queries the teacher on student states, yet conflicts remain in a trained EOPD policy and throughout SP-OPD optimization. Only 0.78% of conflicts under EOPD place the sampled token outside its top-16 set. The dominant failure is therefore not missing the token entirely, but assigning it too little probability inside an otherwise informative teacher distribution. This explains why a local projection is effective: it corrects the contradictory coordinate while retaining the teacher's pairwise probability ratios among alternatives.

The ablations locate the effect within this failure. Advantage masking removes the negative reverse-KL scalar but leaves forward KL free to recreate the conflict; using the projected target in forward KL improves both metrics. Sequence balancing restores the weight lost when shorter correct trajectories are averaged tokenwise. Extending the constraint to incorrect trajectories erodes coverage because rejection contains no token-local certificate. These results support a single mechanism: the objective must respect both the content and the granularity of verifier evidence. The Pass@8 gain means that more problems retain at least one valid sampled solution, while the teacher continues to rank alternatives within the feasible set.

SP-OPD applies when outcome verification and teacher token probabilities are available. The likelihood bound holds for each constructed target, while shared model parameters couple optimization across states. Our experiments establish the mechanism across three student scales in mathematical reasoning. Process-level or learned verification can extend the same construction to finer or confidence-weighted constraints, and other verifiable domains and model families offer a broader test bed.

6 Conclusion

On-policy sampling resolves where to query the teacher, but not whether the teacher supplies a valid target. SP-OPD converts verified success into a token-probability constraint and uses the closest feasible target in both EOPD loss terms. Conflict prevalence, gradient direction, and component ablations support the proposed mechanism. Across five benchmarks and three student scales, it improves both response accuracy and valid-solution coverage.

Reproducibility Statement

Appendix A gives the projection proof, and Table 6 provides the complete training and evaluation configuration.

AI Use Statement

We used generative AI tools only in a limited, ancillary manner for minor language polishing and preliminary figure drafts. All references were independently reviewed and verified by the authors. The authors revised and finalized all AI-assisted material and take full responsibility for the final manuscript.

References

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

Mohammad Sadegh Akhondzadeh, Vijay Lingam, Atula Tejaswi, Chanakya Ekbote, Sujay Sanghavi, and Aleksandar Bojchevski. Reward-gated on-policy distillation, 2026. URL https://arxiv.org/abs/2607.04037.

Ibtihel Amara, Nazanin Sepahvand, Brett H. Meyer, Warren J. Gross, and James J. Clark. Bd-kd: Balancing the divergences for online knowledge distillation, 2024. URL https://arxiv.org/abs/2212.12965.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective, 2025. URL https://arxiv.org/abs/2506.14758.

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

I. Csiszar. I-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability*, 3(1):146 – 158, 1975. doi: 10.1214/aop/1176996454. URL https://doi.org/10.1214/aop/1176996454.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes, 2026. URL https://arxiv.org/abs/2603.25562.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers, 2023. URL https://arxiv.org/abs/2212.10071.

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

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

Seongryong Jung, Suwan Yoon, DongGeon Kim, and Hwanhee Lee. Todi: Token-wise distillation via fine-grained divergence control, 2025. URL https://arxiv.org/abs/2505.16297.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation, 2016. URL https://arxiv.org/abs/1606.07947.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models, 2024. URL https://arxiv.org/abs/2402.03898.

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

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

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

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

Kevin Lu et al. On-policy distillation. Thinking Machines Lab technical blog, 2025. URL https://thinkingmachines.ai/blog/on-policy-distillation/.

Mathematical Association of America. American mathematics competitions and american invitational mathematics examination, 2025. URL https://maa.org/student-programs/amc/.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4, 2023. URL https://arxiv.org/abs/2306.02707.

Haolong Qian, Xianliang Yang, Yinuo Ma, Lirong Che, Feng Lu, Ye Guo, Lei Song, Jiang Bian, and Chun Yuan. The quality-utility paradox: Why high-reward data impairs small model mathematical reasoning, 2026. URL https://arxiv.org/abs/2606.16152.

Zikun Qu, Min Zhang, Mingze Kong, Zhiwei Shang, Zhengyu Chen, Yikun Ban, Shuang Qiu, and Zhongxiang Dai. Spot: Sparse probing and outcome calibration for on-policy distillation, 2026. URL https://arxiv.org/abs/2608.04419.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800.

Jianyu Wu, Yizhou Wang, Encheng Su, Chen Tang, and Shixiang Tang. Dapd: Dual-anchored policy distillation, 2026. URL https://arxiv.org/abs/2608.01735.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models, 2024. URL https://arxiv.org/abs/2404.02657.

Yan Xie, Sijie Zhu, Tiansheng Wen, Bo Chen, and Yifei Wang. On the position bias of on-policy distillation, 2026. URL https://arxiv.org/abs/2606.22600.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation, 2026. URL https://arxiv.org/abs/2606.01249.

Haoran Xu, Hongyu Wang, Yifei Gao, Jiaze Li, Xiaofeng Zhang, and Xiaosong Yuan. Sg-opd: Sign-gated on-policy distillation via sign-consistency gating and phased teacher sampling, 2026. URL https://arxiv.org/abs/2606.09304.

An Yang et al. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024. URL https://arxiv.org/abs/2409.12122.

An Yang et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Chen Yang, Haiyuan Wan, Rengrong Xiong, Yize Chen, and Danny H. K. Tsang. Beyond imitation: Filtering on-policy distillation by reasoning progress, 2026a. URL https://arxiv.org/abs/2608.19408.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Haobo Wang, Mingxuan Xia, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, Junbo Zhao, and Gang Chen. Oprd: On-policy representation distillation, 2026b. URL https://arxiv.org/abs/2606.06021.

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

Qiying Yu et al. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning, 2022. URL https://arxiv.org/abs/2203.14465.

A Proof of the KL Projection

Proof of Proposition 1. Write $a = p_T(y_t \mid s_t)$ and $z = q(y_t)$. For fixed z, minimizing $D_{\mathrm{KL}}(q\|p_T)$ over the alternative tokens subject to $\sum_{v\neq y_t} q(v) = 1 - z$ gives

$$q(v) = \frac{1-z}{1-a} p_T(v), \qquad v \neq y_t, \tag{11}$$

by Lagrangian stationarity. Substitution reduces the objective to

$$z \log \frac{z}{a} + (1 - z) \log \frac{1 - z}{1 - a} = D_{\mathrm{KL}}(\mathrm{Bern}(z) \| \mathrm{Bern}(a)) \,. \tag{12}$$

This function is strictly convex and uniquely minimized at $z = a$. If that point is feasible, $q_t^* = p_T$. Otherwise the constrained minimum is the boundary $z = b$. Substituting $z = b$ yields Equation 4; its common multiplicative factor proves that all pairwise probability ratios among alternative tokens are preserved.

For the opposite KL orientation, fixing $z = q(y_t)$ and minimizing $D_{\mathrm{KL}}(p_T \| q)$ over probabilities of the alternative tokens gives the same proportional rescaling. The remaining objective is $D_{\mathrm{KL}}(\mathrm{Bern}(a) \| \mathrm{Bern}(z))$, whose unconstrained minimum is $z = a$. When $a < b$, it is increasing for $z > a$ and therefore reaches its constrained minimum at $z = b$. Thus both KL orientations produce Equation 4. □

Proof of Proposition 2. If the teacher distribution is feasible, $q_t^* = p_T$ and the stated inequality follows from monotonicity of the logarithm and the definition of the feasible set. If the teacher distribution is infeasible, Equation 4 gives $q_t^*(y_t) = b_t$, hence $A_t^q = 0$. In SP-OPD, sampled tokens on incorrect trajectories are unconstrained and retain A_t^T by definition. □

A.1 Forward-KL Gradient at the Behavior Policy

Proof of Proposition 3. For any normalized target r on a token set S, the target-dependent part of the forward-KL loss is $-\sum_{v \in S} r(v) \log p_\theta(v)$. Differentiating through the softmax gives

$$\frac{\partial}{\partial g(y)} \left[- \sum_{v \in S} r(v) \log p_\theta(v) \right] = p_\theta(y) - \mathbf{1}\{y \in S\} r(y). \tag{13}$$

Before the first optimizer step of a training iteration, $p_\theta(y_t) = p_b(y_t) = b_t$. SP-OPD retains the sampled token in its k-token set and assigns it normalized mass $\widetilde{q}_t(y_t) = b_t / Z_t$, where $Z_t = \sum_{v \in S_t^k} q_t^*(v) \leq 1$. Hence $b_t - \widetilde{q}_t(y_t) \leq 0$. Substituting EOPD's forward-KL target yields the two cases stated in the proposition. □

B Optimization Details

For each training iteration, we first hold a snapshot of the current student fixed as p_b and sample one trajectory per prompt. The deterministic outcome verifier produces the verifier label. We then evaluate the teacher at the same states, construct q_t^* at conflict positions, and update the student. No gradient flows through p_b, the outcome verifier, or the projected target. We use bfloat16 fully sharded data-parallel training on eight H200 GPUs. We evaluate the trained policies after 220 training iterations at 4B and 174 iterations at 0.6B and 1.7B.

SP-OPD averages the clipped reverse-KL surrogate within each response and then across responses. EOPD's forward-KL term retains its original reduction, as shown in Equation 10.

C Training and Evaluation Configuration

Setting	Value
Student / teacher	Qwen3-4B-Base / Qwen3-8B
Training prompts	English DAPO-Math split; 14,116 examples, 14,108 after filtering
Epochs / training iterations	2 / 220
Batch size / optimizer minibatch size	128 / 32
Optimizer minibatches per iteration	4
Trajectories sampled per prompt	1, freshly sampled per training iteration
Maximum training response length	4,096 tokens
Training decoding	temperature 1, top-p 1, top-k disabled
Training prompt template	Qwen3 non-thinking chat template
Optimizer	AdamW, $\beta = (0.9, 0.999)$, weight decay 0.01
Learning rate	3×10^{-6}, cosine decay, no warmup
Maximum gradient norm	1.0
Maximum prompt length / per-GPU token budget	1,024 / 8,192
PPO clipping interval	$[0.8, 1.2]$
Forward-KL top-k / entropy threshold	16 / 0.8
EOPD forward-KL coefficient	1
Numerical probability floor	10^{-8}
Clipped reverse-KL reduction	sequence-balanced (response mean of token means)
Forward-KL reduction	EOPD token reduction
Outcome verifier	DAPO rule-based verifier
Training seed	42
Responses per evaluation problem	8
Maximum evaluation response length	8,192 tokens
Evaluation decoding	temperature 1, top-p 0.8, top-k disabled
Evaluation prompt template	zero-shot Qwen3 template with boxed-answer instruction
Evaluation scorer	Qwen2.5-Math parser and symbolic equivalence checker
Qwen2.5-Math scorer timeout	10 seconds

Table 6: Complete 4B training configuration and evaluation protocol. The 0.6B and 1.7B studies use the corresponding Qwen3-Base student, the MATH-lighteval split, three epochs, and 174 iterations.

D Additional Mechanism Evidence

Teacher–verifier conflict persists across both policy state and response position. Figure 4 compares the initial 4B student and trained EOPD policy on the same 288 DAPO-Math prompts under training decoding, while Table 7 gives exact counts over correct, explicitly terminated trajectories. EOPD reduces the mean probability increase required at each conflict but leaves the contradictory direction prevalent throughout the response.

Policy	Correct terminated trajectories	Sampled tokens	Conflict tokens	Conflict rate (95% interval)	Outside EOPD's top-16 (%)	Mean probability increase
Initial student	20	17,120	7,266	42.44% $[33.94, 50.50]$	8.38	0.183
Trained EOPD	105	137,828	57,294	41.57% $[40.17, 42.88]$	0.78	0.051

Table 7: Teacher–verifier conflict on baseline-policy trajectories. Rates and auxiliary measurements are conditioned on sampled tokens from correct, explicitly terminated trajectories; "outside top-16" and probability increase are further conditioned on conflict positions. Intervals bootstrap trajectories.

E Ablation Definitions

Each 4B ablation changes one implication of the proposed explanation while holding the model pair, data, optimizer, and 220-iteration endpoint fixed.

- **Mask negative per-token advantages on correct trajectories**: sets negative per-token advantages to zero on correct trajectories while leaving EOPD's forward-KL target unchanged.

- **KL projection + token-level reduction**: uses the per-token advantage determined by the projected target and its forward-KL target, while retaining EOPD's reduction for the clipped surrogate.
- SP-OPD: adds sequence-balanced reduction of the clipped surrogate to KL projection.
- **Also constrain sampled tokens on incorrect trajectories**: adds the complementary constraint at every token position on an incorrect trajectory.

The first contrast asks whether correcting a scalar advantage is sufficient; the second asks whether loss aggregation must match trajectory-level evidence; the last asks whether rejection licenses the converse constraint.

Constraint-direction variant. Extending SP-OPD's constraint to sampled tokens on incorrect trajectories yields 43.89 Avg@8 and 62.24 Pass@8, compared with 43.92 and 64.91 for SP-OPD. SP-OPD exceeds the symmetric variant by 0.04 Avg@8 and 2.67 Pass@8 points, supporting the verifier's asymmetric treatment of correct and incorrect trajectories.

F Additional Baselines

The additional baselines contrast target validity with broader changes to exploration, sampled-token weighting, teacher–student divergence, and forward-KL placement. Tables 8– 9 compare these alternatives for the Qwen3-1.7B-Base student on MATH-500, AMC23, AIME24, and AIME25. The divergence controls cover entropy-difference weighting, head–tail adaptation, and vocabulary-discrepancy weighting (Jin et al., 2026; Amara et al., 2024; Wu et al., 2024; Jung et al., 2025).

Method	MATH-500	AMC23	AIME24	AIME25	Macro Avg@8↑
OPD (Lu et al., 2025)	67.76	39.06	8.33	**6.25**	30.35
GKD ($\lambda = 0.5$) (Agarwal et al., 2024)	64.53	35.62	5.83	4.58	27.64
GKD ($\lambda = 0.8$) (Agarwal et al., 2024)	65.28	37.50	6.67	3.75	28.30
GKD ($\lambda = 1$; on-policy forward KL) (Agarwal et al., 2024)	67.20	38.44	9.17	5.00	29.95
OPD + entropy bonus (Schulman et al., 2017)	66.87	39.69	9.58	4.58	30.18
OPD + advantage shaping (Cheng et al., 2025)	67.90	36.56	8.75	<u>5.83</u>	29.76
OPD + forward KL at all positions	67.58	39.39	8.75	5.00	30.18
OPD + random-position forward KL	67.33	36.88	7.92	5.42	29.39
BD-KD (Amara et al., 2024)	67.45	38.75	9.17	4.58	29.99
AKL (Wu et al., 2024)	66.43	37.81	7.50	4.58	29.08
ToDi (Jung et al., 2025)	67.43	38.75	8.33	<u>5.83</u>	30.09
ExOPD (Yang et al., 2026c)	66.45	39.38	<u>10.00</u>	5.00	30.21
EOPD	**68.50**	<u>40.00</u>	8.33	5.42	<u>30.56</u>
SP-OPD	<u>68.13</u>	**40.31**	**10.83**	5.42	**31.17**

Table 8: Benchmark-level Avg@8 (%) against additional baselines. Bold and underlined values mark the best and second-best results.

Method	MATH-500	AMC23	AIME24	AIME25	Macro Pass@8↑
OPD (Lu et al., 2025)	84.80	70.00	<u>20.00</u>	<u>16.67</u>	47.87
GKD ($\lambda = 0.5$) (Agarwal et al., 2024)	84.40	67.50	10.00	10.00	42.98
GKD ($\lambda = 0.8$) (Agarwal et al., 2024)	84.20	70.00	13.33	10.00	44.38
GKD ($\lambda = 1$; on-policy forward KL) (Agarwal et al., 2024)	85.00	67.50	**23.33**	13.33	47.29
OPD + entropy bonus (Schulman et al., 2017)	86.00	**75.00**	<u>20.00</u>	<u>16.67</u>	49.42
OPD + advantage shaping (Cheng et al., 2025)	85.20	**75.00**	**23.33**	**20.00**	<u>50.88</u>
OPD + forward KL at all positions	86.20	**75.00**	**23.33**	**20.00**	**51.13**
OPD + random-position forward KL	84.80	67.50	<u>20.00</u>	13.33	46.41
BD-KD (Amara et al., 2024)	<u>86.80</u>	<u>72.50</u>	<u>20.00</u>	<u>16.67</u>	48.99
AKL (Wu et al., 2024)	85.20	70.00	16.67	13.33	46.30
ToDi (Jung et al., 2025)	<u>86.80</u>	**75.00**	<u>20.00</u>	<u>16.67</u>	49.62
ExOPD (Yang et al., 2026c)	85.20	**75.00**	<u>20.00</u>	<u>16.67</u>	49.22
EOPD	**87.80**	65.00	16.67	13.33	45.70
SP-OPD	85.60	**75.00**	<u>20.00</u>	<u>16.67</u>	49.32

Table 9: Benchmark-level Pass@8 (%) against additional baselines. Bold and underlined values mark the best and second-best results.

G BENCHMARK-LEVEL RESULTS ACROSS STUDENT SCALES

The aggregate gains in Table 2 arise across multiple benchmarks. The two tables below separate per-sample reliability (Avg@8) from solution coverage (Pass@8). SP-OPD improves both macro metrics at every scale, with consistently larger gains in coverage.

Student	Method	MATH-500	AMC23	OlympiadBench	AIME24	AIME25	Macro
0.6B	KD	47.80	23.43	18.15	2.19	0.83	18.48
	GRPO	51.83	25.25	18.91	4.58	0.83	20.28
	OPD	50.09	24.69	20.15	2.50	1.25	19.74
	EOPD	50.63	29.38	19.78	2.92	2.50	21.04
	ExOPD	50.55	29.06	19.43	3.75	0.83	20.72
	SP-OPD	51.60	29.06	19.94	3.33	2.50	21.29
1.7B	KD	63.80	38.11	27.66	10.06	3.36	28.60
	GRPO	68.83	40.62	29.89	9.17	4.58	30.62
	OPD	67.76	39.06	30.09	8.33	6.25	30.30
	EOPD	68.50	40.00	30.89	8.33	5.42	30.63
	ExOPD	66.45	39.38	29.76	10.00	5.00	30.12
	SP-OPD	68.13	40.31	31.30	10.83	5.42	31.20
4B	KD	74.73	48.13	37.33	12.50	12.08	36.95
	GRPO	80.47	58.44	43.37	14.85	12.66	41.96
	OPD	78.81	57.33	42.10	18.33	12.08	41.73
	EOPD	79.45	56.25	43.44	18.33	15.83	42.66
	ExOPD	76.08	54.37	41.06	20.00	12.92	40.88
	SP-OPD	79.23	60.63	43.52	17.92	18.33	43.92

Table 10: Five-benchmark Avg@8 comparison across student scales (%).

Student	Method	MATH-500	AMC23	OlympiadBench	AIME24	AIME25	Macro
0.6B	KD	69.60	52.50	36.59	6.67	6.67	34.41
	GRPO	74.40	55.00	34.96	10.00	10.00	36.87
	OPD	73.20	57.50	37.19	10.00	6.67	36.91
	EOPD	74.60	62.50	40.30	10.00	10.00	39.48
	ExOPD	76.20	57.50	40.59	16.67	3.33	38.86
	SP-OPD	75.60	62.50	40.59	13.33	10.00	40.41
1.7B	KD	84.20	70.00	48.40	20.00	16.67	47.85
	GRPO	84.60	70.00	50.81	20.00	16.67	48.42
	OPD	84.80	70.00	51.56	20.00	16.67	48.61
	EOPD	87.80	65.00	52.00	16.67	13.33	46.96
	ExOPD	85.20	75.00	51.85	20.00	16.67	49.74
	SP-OPD	85.60	75.00	53.93	20.00	16.67	50.24
4B	KD	92.20	85.00	62.52	30.00	23.33	58.61
	GRPO	91.00	75.00	60.74	30.00	26.67	56.68
	OPD	90.80	80.00	58.80	26.67	30.00	57.25
	EOPD	91.60	77.50	64.44	36.67	33.33	60.71
	ExOPD	92.00	85.00	61.19	36.67	36.67	62.30
	SP-OPD	92.80	85.00	63.41	36.67	46.67	64.91

Table 11: Five-benchmark Pass@8 comparison across student scales (%).

H Training–Evaluation Overlap Analysis

The 4B gain persists after removing every exact training–evaluation prompt match. We normalize training and evaluation text before matching. This finds no overlap between the MATH-lighteval training split used at 0.6B and 1.7B and the five-benchmark suite. The 14,116-example DAPO-Math split used at 4B contains 10 matches: five from AMC23, four from MATH-500, and one from OlympiadBench. Removing these problems leaves gains of +1.44 Avg@8 and +3.04 Pass@8 for SP-OPD over EOPD. The resulting five-benchmark macro Avg@8 and Pass@8 values are shown below.

Method	Avg@8	Pass@8
EOPD	42.65	60.72
SP-OPD	**44.09**	**63.77**

Table 12: Five-benchmark macro Avg@8 and Pass@8 (%) after removing exact training–evaluation prompt matches.