SG-OPD: Sign-Gated On-Policy Distillation via Sign-Consistency Gating and Phased Teacher Sampling

Anonymous ACL submission

Abstract

On-policy distillation (OPD) trains a student on its own trajectories with dense per-token supervision from a stronger teacher, and often outperforms off-policy distillation and standard reinforcement learning. However, we find that its effectiveness implicitly relies on two assumptions that frequently break in practice: trajectory-level alignment between the student and the teacher, and uniform token-level reliability of the teacher's preferences. We therefore propose **Sign-Gated On-Policy Distillation** (**SG-OPD**), which uses a binary verifier as a trust signal for the teacher at two complementary granularities: *phased teacher sampling* mixes in verifier-endorsed teacher rollouts at cold-start, and a *sign-consistency gate* extrapolates the distillation update on tokens where the teacher agrees with the verifier-correct direction and interpolates it where it disagrees. Experiments on competition-level mathematical reasoning benchmarks show that SG-OPD consistently outperforms standard OPD, with average gains of 1.98 and 7.50 at the per-sample and per-question levels, respectively.

1 Introduction

The strong reasoning capabilities of large language models (Guo et al., 2025; Yang et al., 2025) come at steep computational cost, motivating distillation (Hinton et al., 2015) to compress them into smaller students. Off-policy distillation (Taori et al., 2023) trains on teacher-generated trajectories but suffers from exposure bias (Bengio et al., 2015) at inference time. On-policy distillation (OPD) (Agarwal et al., 2024; Lu and Lab, 2025) resolves this mismatch by sampling from the student and minimising a reverse KL to the teacher, yielding dense per-token supervision on the student's own distribution.

However, we observe that the effectiveness of OPD implicitly relies on two assumptions that frequently break in practice, which we attribute to the following two structural limitations:

- *Trajectory-level alignment is fragile.* In strong-to-weak settings, a weak student's cold-start rollouts lie far from the teacher's trajectory support, so the reverse-KL signal there is dominated by noise rather than informative supervision.
- *Token-level teacher reliability is not uniform.* Even for rollouts verified as correct, the teacher's per-token preferences may contradict the verifier-correct direction, causing the reverse-KL gradient to penalize tokens that support a valid trajectory.

We propose **Sign-Gated On-Policy Distillation (SG-OPD)**, which treats a binary verifier as a trust signal for the teacher at two complementary granularities, as illustrated in Figure 1. Specifically,

Sample level: Phased Teacher Sampling. To bridge the trajectory-level mismatch at cold-start, we adopt the optimization strategy inspired by mixed policy optimization(Zhang et al., 2026) instead of the conventional SFT-then-by-RL paradigm(Li et al., 2026c). We use the annealed schedule mixes in verifier-endorsed teacher rollouts early in training and decays to fully on-policy student rollouts later, so that the student's distillation targets are drawn from trajectories the verifier deems correct precisely when its own rollouts are not yet aligned with the teacher.

Token level: Sign-Consistency Gate. To enforce per-token reliability, we combine the verifier outcome with the reverse-KL advantage to label each token as either *consensus*, when the teacher agrees with a verifier-correct direction, or *conflict*, when the teacher would have moved the student away from a verifier-correct trajectory. Consensus tokens are extrapolated to amplify the trustworthy distillation signal, while conflict tokens are interpo-

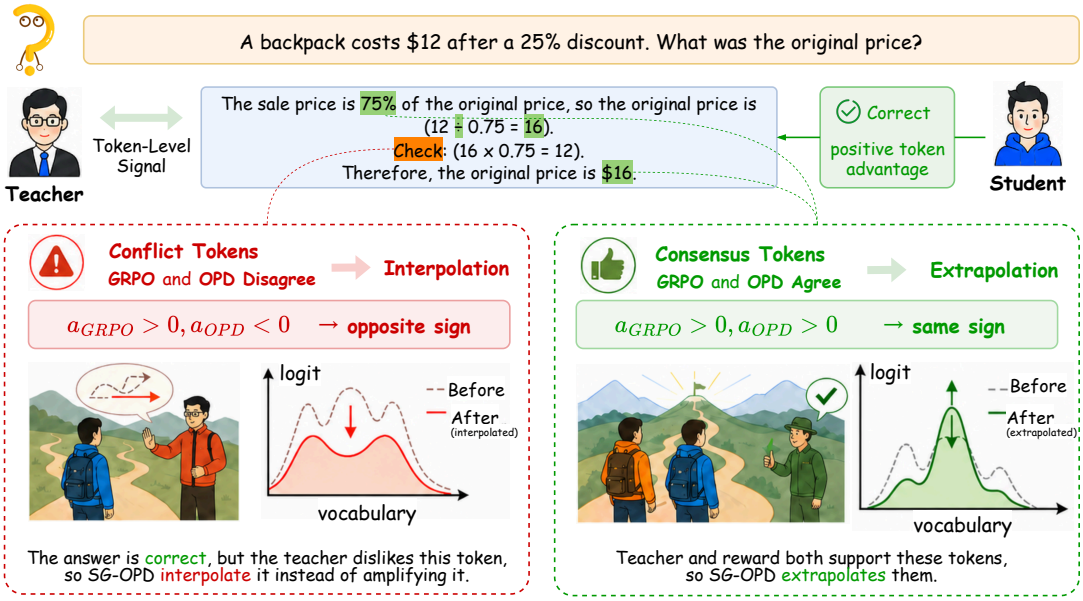

Figure 1: Illustration of token-level sign-consistency gating on a correct math rollout. Since the final answer is verified as correct, the outcome-level GRPO advantage a_1 is positive for the trajectory. However, the OPD reverse-KL advantage a_2 can still vary by token: consensus tokens such as "75%", "÷", and "16" have $a_1 a_2 > 0$ and are extrapolated, while a redundant verification token such as "Check:" can receive $a_2 < 0$ because the teacher assigns it lower probability, so SG-OPD routes it through interpolation.

lated to mute them.

Together, the two mechanisms let the student inherit teacher supervision where the teacher is reliable, and back off where it is not.

Under a strong-to-weak setup on competition-level math reasoning benchmarks, SG-OPD delivers stronger and more robust performance than existing baselines, remaining stable where uniform extrapolation collapses. Our main contributions are:

- We identify two implicit assumptions of on-policy distillation that frequently break in practice: a trajectory-level alignment assumption, where student rollouts and teacher trajectories are insufficiently aligned at cold-start, and a token-level reliability assumption, where the teacher's per-token preferences contradict verifier-correct directions even on rollouts judged correct by the verifier.
- We propose **SG-OPD**, which uses a binary verifier purely as a trust signal for the teacher, combining phased teacher sampling at the trajectory level with sign-consistency-gated extrapolation/interpolation at the token level.
- Experiments on competition-level mathematical reasoning benchmarks show that SG-OPD consistently improves over existing on-policy distillation baselines.

2 Related Work

On-policy distillation. Classical KD (Hinton et al., 2015) fits the student to a frozen teacher, typically via SFT on teacher responses (Taori et al., 2023) or sequence-level variants like SeqKD (Kim and Rush, 2016). OPD (Lu and Lab, 2025) instead samples from the student and minimizes the reverse KL, providing dense per-token feedback at the cost of student–teacher mismatch. Video-OPD (Li et al., 2026a) tackles the challenge of cross-modal misalignment. G-OPD (Yang et al., 2026) recasts OPD as a KL-regularized RL problem with a single global extrapolation factor λ, and AOPD (Jia et al., 2026) switches negative-advantage tokens from policy gradient to truncated forward-KL. These methods either share a global λ or condition only on a distillation-side signal. Several works (Li et al., 2026b) have sought to elucidate the factors driving the performance gains of the OPD approach.

Mixed Policy optimization. GRPO (Shao et al., 2024) and its successors (Guo et al., 2025; Yu et al., 2025) optimize a binary verifiable reward with group-normalized advantages, in the spirit of classical trust regions (Schulman et al., 2017). Recent work has been exploring mixed policy optimization by integrating SFT and GRPO. For ex-

ample, ExPO (Zheng et al., 2025) extrapolates model weights post-hoc and CHORD (Zhang et al., 2026) re-weights an SFT loss with a fixed prior, while teacher-trajectory mixing (Wulfmeier et al., 2024) bridges cold-start with off-policy expert data. DFT(Wu et al., 2026) is analogous.

3 Preliminaries and Failure Modes of OPD

Let π_θ be the student, π^* the frozen teacher, and π_{ref} the reference policy initialized from the student. A prompt $x \sim \mathcal{D}$ generates a student trajectory $y = (y_1, \ldots, y_T)$ with verifiable outcome reward $r(x, y) \in \{0, 1\}$.

Verifier signal (sample level). We define an outcome-level **verifier signal** $a_1(t) := (r(x, y) - \mu_x)/(\sigma_x + \epsilon)$ computed in the GRPO style, normalized across G rollouts of x with mean and std μ_x, σ_x. With binary r, a_1 is constant per trajectory; *we use it only as a trust signal for the teacher*, not as an optimization target.

OPD's per-token signal. The mechanism we want to stabilize is the OPD policy gradient. At each token y_t, OPD provides the **reverse-KL advantage** $a_2(t) := \log \pi_\theta(y_t) - \log \pi^*(y_t)$, derived from the on-policy reverse-KL objective under a per-token discount of 0 (Lu and Lab, 2025). The OPD policy gradient is then the dense per-token form

$$\nabla_\theta \mathcal{J}_{\text{OPD}} = \mathbb{E}\left[\sum_{t=1}^{T} a_2(t)\, \nabla_\theta \log \pi_\theta(y_t)\right]. \quad (1)$$

G-OPD extrapolation. Yang et al. (2026) reinterpret this objective as KL-regularized RL with extrapolation factor $\lambda \geq 1$ and obtain a **G-OPD** advantage $A_t^{\text{G-OPD}}(\lambda)$ generalising a_2; $\lambda = 1$ recovers OPD, $\lambda > 1$ extrapolates beyond the teacher (**ExOPD**), and larger λ degrades training in our setting. The full reverse-KL objective and the explicit G-OPD form are deferred to Appendix A.

Failure mode 1: trajectory-level alignment is fragile. OPD assumes that the student's rollouts and the teacher's trajectories are sufficiently aligned. In strong-to-weak settings, a weak student often produces trajectories that the teacher itself would find unlikely, so the reverse-KL signal at cold-start is unreliable and the early distillation update is dominated by noise rather than informative supervision. Pushing λ beyond a moderate value amplifies this noise: the "untrainable" regime of Yang et al. (2026) is observed precisely when the early student distribution sits far from the teacher's.

Failure mode 2: token-level teacher reliability is not uniform. OPD also assumes that the teacher is uniformly trustworthy along every student rollout. We find that this assumption breaks even on rollouts that the verifier judges correct: while $a_1(t) > 0$ for the entire trajectory, a_2 can still flip sign per token, indicating that the teacher would have suppressed a token that lies on a verifiably correct path. We refer to tokens with $a_1(t)\, a_2(t) \leq 0$ as *conflict* tokens; on these tokens, blindly amplifying the reverse-KL gradient drives the student away from a verified solution. Fig. 1 visualises this per-token pattern.

Empirical signature. Two observations make these concrete. **(i)** The reverse-KL signal at cold-start is dominated by trajectory-level mismatch, and uniformly increasing λ does not recover performance (Tab. 1). **(ii)** A non-trivial fraction of tokens remain in the conflict regime $a_1 a_2 \leq 0$ throughout training rather than only at cold-start (Fig. 7, Appendix E). Sec. 4 introduces a two-granularity framework that uses the verifier signal a_1 as a trust signal for the teacher to address both failure modes.

4 Method: SG-OPD

SG-OPD couples verifiable-reward RL with OPD, using a binary verifier as a trust signal at two levels: (i) a *sample-level* teacher anchor, **phased teacher sampling (PTS)**, adding an auxiliary loss on verified teacher rollouts; and (ii) a *token-level* sign-consistency mechanism, routing each token by whether the verifier-induced advantage and the OPD advantage agree in sign. Algorithm 1 summarizes the full training step.

Notation recap. We collect the symbols below (all introduced in §3).

Symbol	Meaning
$\pi_\theta, \pi^*, \pi_{\text{ref}}$	student / frozen teacher / reference policy
$x,\ y = (y_1, \ldots, y_T)$	prompt and student rollout
$r(x, y) \in \{0, 1\}$	verifiable outcome reward
$a_1(t)$	GRPO outcome advantage at token t
$a_2(t)$	OPD reverse-KL advantage at token t
g_t	token-level sign-consistency gate
ϕ_t	detached stability weight
$\alpha(t)$	phased teacher-anchor schedule
$\mathcal{L}_{\text{TLT}}$	token-level sign-consistency-gated loss
$\mathcal{L}_{\text{SLT}}$	sample-level teacher-anchor loss

Hyperparameters fall into five groups: *extrapolation strength* (λ_{high}, λ_{base}), *conflict fallback* (β and the fallback mode), *teacher sampling ratio* (ρ), *phased schedule* ($P_1, P_2, \alpha_0, \alpha_{\text{end}}$), and *stability clipping* (τ). Default values are listed in Appendix C.

4.1 Sample-Level Teacher Anchor: Phased Teacher Sampling (PTS)

Motivation. At cold-start the student attains low accuracy, so most on-policy rollouts are incorrect.then dominated by trajectories the teacher itself would not support, and the student can drift onto a low-reward manifold where the verifier supplies little corrective gradient. PTS addresses this sample-level failure mode by injecting a small number of verified teacher rollouts early in training and then annealing this anchor away.

Verified teacher rollouts. For each mini-batch $\mathcal{B}$, we reserve a fraction ρ of prompts for teacher sampling, yielding $\mathcal{B}_T$. The teacher generates $y^T \sim \pi^*(\cdot|x)$ on this subset, and only trajectories verified as correct are retained. The resulting sample-level teacher-anchor loss is

$$\mathcal{L}_{\text{SLT}} = \sum_{\substack{(x,y^T)\in\mathcal{B}_T \\ r(x,y^T)=1}} \sum_t \mathcal{L}_{\text{CE}}\big(\pi_\theta, y_t^T\big). \quad (2)$$

Incorrect teacher rollouts are discarded, so the anchor is defined by verifier agreement rather than teacher likelihood alone.

Phased annealing. The teacher anchor is useful during cold-start but should not define the asymptotic training distribution. We therefore weight it by a three-phase cosine schedule,

$$\alpha(t) = \begin{cases} \alpha_0, & t \le P_1, \\ \alpha_{\text{end}} + \frac{\alpha_0-\alpha_{\text{end}}}{2}\left(1+\cos\left(\frac{\pi(t-P_1)}{P_2-P_1}\right)\right), & P_1 < t \le P_2, \\ 0, & t > P_2, \end{cases} \quad (3)$$

where α_0 is the cold-start weight, α_{end} is the value at the end of the transition window, and P_1, P_2 are the phase boundaries. Once $\alpha(t)=0$, the auxiliary anchor is removed and training becomes fully on-policy.

4.2 Token-Level Sign-Consistency and Stability Weighting

Motivation. PTS controls *which trajectories* the student visits. It does not resolve token-level sign conflict, which arises within a fixed student rollout: on some tokens, the verifier-induced GRPO advantage $a_1(t)$ and the teacher-induced OPD advantage $a_2(t)$ point in opposite directions. A uniform linear combination then amplifies an update direction opposed by one of the two signals. SG-OPD therefore routes tokens by sign agreement before forming the policy-gradient advantage.

Sign-consistency gate. We encode whether the two token-level signals agree with

$$g_t \;:=\; \mathbf{1}[\,a_1(t)\cdot a_2(t) > 0\,] \;\in\; \{0,1\}. \quad (4)$$

Here $g_t = 1$ marks *consensus* tokens, where the verifier and teacher push the sampled token in the same direction, while $g_t = 0$ marks *conflict* tokens.

Routed token advantage. Let

$$A_t^{\text{G-OPD}}(\lambda) = a_2(t) + (\lambda-1)\big(\log\pi_{\text{ref}}(y_t) - \log\pi^*(y_t)\big), \quad (5)$$

which recovers OPD at $\lambda=1$ and ExOPD at $\lambda>1$. On consensus tokens we apply stronger extrapolation,

$$A_t^{\text{cons}} = A_t^{\text{G-OPD}}(\lambda_{\text{high}}), \qquad \lambda_{\text{high}} > 1, \quad (6)$$

and on conflict tokens the default fallback is softened OPD,

$$A_t^{\text{conf}} = \beta\, a_2(t), \qquad \beta\in[0,1]. \quad (7)$$

Thus $\beta = 1$ recovers OPD on conflict tokens, whereas $\beta=0$ masks them. Alternative `preserve` and `grpo` fallbacks are evaluated in Tab. 3. The routed advantage is

$$A_t^{\text{SG}} = g_t \cdot A_t^{\text{cons}} + (1-g_t)\cdot A_t^{\text{conf}}. \quad (8)$$

When the gate is disabled, A_t^{SG} reduces to G-OPD, so this strictly generalizes uniform extrapolation.

Stability weighting. Large $|a_2(t)|$ values can cause a small number of OPD outliers to dominate the actor gradient. We therefore apply a detached clipping weight after the sign-consistency decision,

$$\phi_t \;=\; \min\big(1,\; \tau/|a_2(t)|\big), \quad (9)$$

where τ is a clipping hyperparameter. In implementation, ϕ_t is detached and batch-normalized to unit mean. It therefore changes the scale of the token update but not whether a token is classified as consensus or conflict. Ablations in Appendix F isolate this weighting from the sign-consistency gate.

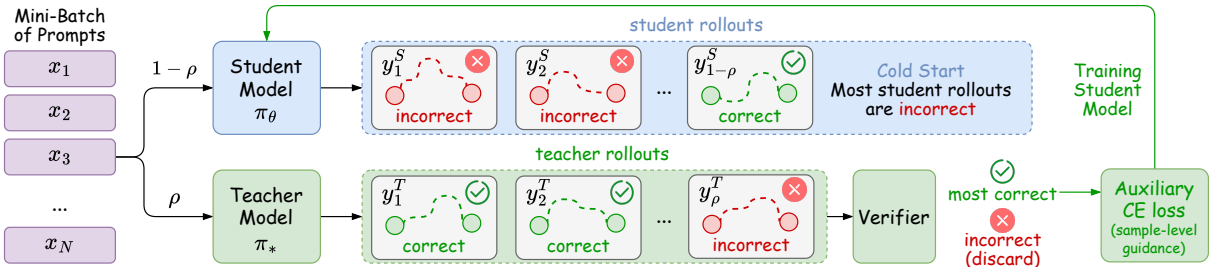

Figure 2: Sample-level Phased Teacher Sampling (PTS). A mini-batch is split into student on-policy rollouts and a small fraction of teacher rollouts. Verified teacher trajectories are retained and used as an auxiliary CE anchor, while incorrect teacher trajectories are discarded. The teacher-guidance weight is annealed from warm-up to zero, so the asymptotic training distribution remains on-policy.

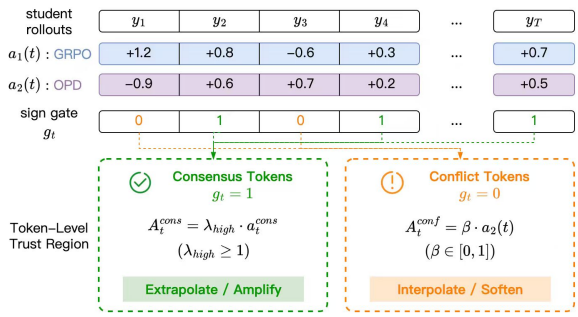

Figure 3: Overview of token-level sign-consistency gating. GRPO and OPD token advantages are routed by sign agreement: consensus tokens are extrapolated, while conflict tokens are softened by interpolation.

4.3 Combined Objective and Algorithm

The token-level policy-gradient loss is

$$\mathcal{L}_{\text{TLT}}(\theta) = -\,\mathbb{E}\left[\frac{1}{|y|}\sum_{t=1}^{|y|} \phi_t\, A_t^{\text{SG}}\, \log \pi_\theta(y_t \,|\, c_t)\right], \tag{10}$$

where the standard PPO importance-ratio clip (Schulman et al., 2017) replaces the $\log \pi_\theta$ factor in our implementation. The complete objective is

$$\mathcal{L}_{\text{SG-OPD}}(\theta) = \mathcal{L}_{\text{TLT}}(\theta) + \alpha(t)\, \mathcal{L}_{\text{SLT}}(\theta). \tag{11}$$

Equivalently, the token-level term can be written as the routed per-token expectation

$$\mathcal{L}_{\text{SG-OPD}}(\theta) = -\,\mathbb{E}\Big[\frac{1}{|y|}\sum_{t=1}^{|y|} \phi_t \big(g_t\, A_t^{\text{cons}} + (1-g_t)\, A_t^{\text{conf}}\big) \log \pi_\theta(y_t \,|\, c_t)\Big] + \alpha(t)\, \mathcal{L}_{\text{SLT}}(\theta). \tag{12}$$

The token-level gate is always active, while the sample-level teacher anchor is phased out by $\alpha(t)$.

Algorithm 1 One step of SG-OPD.

Require: Student π_θ, teacher π^*, reference π_{ref}, current step t, total steps T, ratio ρ, phase (P_1, P_2), $(\lambda_{\text{base}}, \lambda_{\text{high}}, \beta, \tau)$.

Sample mini-batch $\mathcal{B} = \{x_i\}$ from $\mathcal{D}$.
Split $\mathcal{B}$ into $\mathcal{B}_S$ and $\mathcal{B}_T$ with $|\mathcal{B}_T|/|\mathcal{B}| = \rho$.
$y_S \leftarrow \pi_\theta(\cdot|x_i)$ for $x_i \in \mathcal{B}_S$ ▷ on-policy rollouts
$y_T \leftarrow \pi^*(\cdot|x_j)$ for $x_j \in \mathcal{B}_T$ ▷ teacher rollouts
Verify $r(x, y) \in \{0, 1\}$ on all trajectories.
for token y_t in y_S **do**
  $a_1(t) \leftarrow A_t^{\text{GRPO}}$, $a_2(t) \leftarrow \log \pi_\theta(y_t) - \log \pi^*(y_t)$.
  $g_t \leftarrow \mathbf{1}[a_1 \cdot a_2 > 0]$ ▷ sign-consistency gate
  $A_t^{\text{SG}} \leftarrow$ Eq. (8); $\phi_t \leftarrow$ Eq. (9).
$\mathcal{L}_{\text{TLT}} \leftarrow$ PPO-clipped policy gradient with $A_t^{\text{SG}} \cdot \phi_t$.
Filter y_T by $r(x_j, y_T) = 1$.
$\mathcal{L}_{\text{SLT}} \leftarrow$ Eq. (2).
$\alpha(t) \leftarrow$ Eq. (3).
$\mathcal{L} \leftarrow \mathcal{L}_{\text{TLT}} + \alpha(t)\, \mathcal{L}_{\text{SLT}}$.
Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$.

Thus SG-OPD uses verified teacher rollouts to stabilize cold-start but restores a fully on-policy objective once the anchor is annealed away.

The sign-consistency gate modifies the actor-update path, whereas PTS adds the teacher-rollout and SFT-loss path. The full algorithm is implemented in the verl framework (Sheng et al., 2024); we will release the implementation upon acceptance.

Our experiments are organized around four questions:

- **(Q1)** Does SG-OPD improve over OPD and ExOPD on competition-level math reasoning under an identical recipe? (§5.2, Tab. 1)
- **(Q2)** Do the sample-level and token-level mechanisms each contribute on their own, and are they complementary? (§5.3)
- **(Q3)** Does the sign-consistency gate widen the safe range of consensus-token extrapolation strength, recovering performance at a λ_{high} where uniform extrapolation collapses? (§5.3,

Tab. 3)

- **(Q4)** Does SG-OPD raise training reward without collapsing policy entropy, as the token-level gate predicts? (§5.4, Fig. 5)

5 Experiments

5.1 Setup

Models. The student π_θ is Qwen3-1.7B-Non-Thinking, and the teacher π^* is the step-500 Qwen3-4B-Non-Thinking-RL-Math checkpoint. The reference π_{ref} is the student's initial state. All three share the same tokenizer, following the strong-to-weak setting in Table 3.

Training data. Training prompts come from DeepMath-103K (He et al., 2025) filtered to difficulty level ≥ 6, resulting in 57 K problems. For each prompt we sample $G = 8$ rollouts at temperature 1.0.

Evaluation. We evaluate on four competition-level math reasoning benchmarks: AIME24, AIME25, HMMT25-Feb, and HMMT25-Nov. We report avg@32 and pass@32 accuracy with sampling temperature $\mathcal{T} = 1.0$, top-$p = 1.0$, and a generation budget of 16,384 tokens; AVG always denotes the arithmetic mean over the benchmarks. This protocol matches Yang et al. (2026)'s Table 3.

Training and baselines. We train for 100 optimizer steps with the GRPO advantage estimator and no learned critic. SG-OPD default hyperparameters are selected via the ablations in §5.3; OPD and ExOPD baselines are trained under the same recipe. All schedule comparisons use the same optimizer and teacher-rollout budgets. Full optimization details, hyperparameters, and run-to-run variance are in Appendix C and G.

5.2 Main Results (Q1)

Key Finding 1

A binary verifier provides a reliable token-level trust signal for the teacher in on-policy distillation.

Table 1 summarizes the main results. **SG-OPD achieves the highest AVG under both metrics**: 29.53 avg@32 (+1.98 over OPD, +1.54 over ExOPD) and 59.17 pass@32 (+7.50 over OPD, +5.00 over ExOPD). The gains vary across benchmarks, and the per-benchmark breakdown reveals where SG-OPD's two mechanisms contribute most.

Per-benchmark trends (avg@32). On AIME-style benchmarks, the largest gain appears on AIME25 (+5.00 over OPD), where the sign-conflict fraction is also highest (§5.4); AIME24 shows a smaller but consistent +2.39 improvement. The HMMT subsets pose greater difficulty (avg@32 in the 18–20% range) and SG-OPD matches OPD exactly on HMMT25-Feb (18.02) and improves slightly on HMMT25-Nov (+0.52).

Pass@32 unlocks the largest improvements. As shown in the pass@32 block of Table 1, when counting trajectories that successfully solve the problem at least once, the exploration advantage of SG-OPD becomes more pronounced: SG-OPD reaches 76.67 on AIME24 (+6.67 over OPD) and 66.67 on AIME25 (+16.67). This is consistent with the design of SG-OPD (Sec. 4): the gate permits aggressive extrapolation on consensus tokens, expanding the set of trajectories the student can reach without sacrificing the verifier's outcome signal on conflict tokens.

5.3 Ablations

We conducted over 50 controlled training runs varying the token-level and sample-level hyperparameters. Four representative slices are reported here; the full table is in Appendix F.

Key Finding 2

The sign-consistency gate widens the safe range of consensus-token extrapolation strength.

Token-level gate (Q3). Tab. 3 sweeps the token-level knobs λ_{high}, fallback mode, and β. The configuration maximizing AVG employs $\lambda_{\text{high}} = 1.8$ with `interp` and $\beta = 1$; `preserve` is also effective but slightly more conservative. The mechanism is robust across configurations: *any* sign-consistency-gated configuration improves over uniform-λ ExOPD at the *same* λ. The contrast is sharpest at aggressive strengths: pushing uniform ExOPD to $\lambda = 1.8$ collapses AVG to 24.71 (-3.28 vs the best uniform setting $\lambda = 1.25$, and -2.84 below the OPD baseline 27.55), whereas the sign-consistency gate at the *same* $\lambda_{\text{high}} = 1.8$ reaches 28.78 (group (b) of Tab. 3), +0.79 above the best uniform ExOPD. Sign-gating thus turns a regime that uniform extrapolation renders untrainable into the best-performing setting.best overall run. The other results is reported in Appendix F.

Method	avg@32 (%)					pass@32 (%)				
	AIME24	AIME25	HMMT25-Feb	HMMT25-Nov	**AVG**	AIME24	AIME25	HMMT25-Feb	HMMT25-Nov	**AVG**
SFT (Taori et al., 2023)	30.93	25.62	16.04	18.33	22.73	48.33	33.67	24.67	31.41	34.52
SeqKD (Kim and Rush, 2016)	31.88	26.46	16.46	18.96	23.44	50.00	36.67	26.67	33.33	36.67
GKD (Agarwal et al., 2024)	37.71	32.19	17.40	18.96	26.57	66.67	50.00	36.67	46.67	50.00
OPD (Lu and Lab, 2025)	38.96	33.44	18.02	19.79	27.55	70.00	50.00	40.00	46.67	51.67
ExOPD (Yang et al., 2026)	40.83	33.12	**20.21**	17.81	27.99	70.00	53.33	**50.00**	43.33	54.17
SG-OPD (Ours)	**41.35**	**38.44**	18.02	**20.31**	**29.53**	**76.67**	**66.67**	43.33	**50.00**	**59.17**
Δ *vs OPD*	*+2.39*	*+5.00*	*±0.00*	*+0.52*	*+1.98*	*+6.67*	*+16.67*	*+3.33*	*+3.33*	*+7.50*

Table 1: Main results across four competition-level math reasoning benchmarks. We compare **SG-OPD** against SeqKD, GKD, OPD, and ExOPD under the strong-to-weak setting (Qwen3-1.7B distilled from Qwen3-4B-Non-Thinking-RL-Math, step 500). **Bold** marks the best result within each column, while underlined values denote the second best; Δ *vs OPD* reports the absolute improvement of SG-OPD over the OPD baseline.

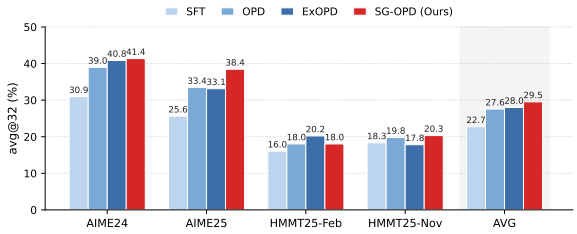

Figure 4: Per-benchmark avg@32 accuracy (%) under the strong-to-weak setting (Qwen3-1.7B distilled from Qwen3-4B-Non-Thinking-RL-Math). The light-to-dark blue gradient ranges over the off-/on-policy distillation baselines (SFT, OPD, ExOPD); **SG-OPD** (red) consistently leads on AIME and on average.

Sign-Gate	PTS	Per-benchmark avg@32				AVG
		A24	A25	H-F	H-N	
✗	✗	38.96	33.44	18.02	19.79	27.55
✗	✓	41.25	35.42	**18.23**	19.48	28.59
✓	✗	**41.88**	36.25	17.60	19.38	28.78
✓	✓	41.35	**38.44**	18.02	**20.31**	**29.53**

Table 2: Two-component ablation of SG-OPD on the four competition math benchmarks (avg@32, %). ✓ / ✗ indicate whether each component is enabled. **Bold** marks the best result within each column and underlined values denote the second best. The token-level **Sign-Gate** and the sample-level **PTS** target different failure modes, and the gains are roughly additive.

Sample-level anchor. Tab. 2 contrasts enabling the sample-level anchor (PTS) against the no-PTS baseline within the two-component grid; PTS alone raises AVG from 27.55 to 28.59. We further sweep the PTS internal knobs (ρ, P_1/P_2, correctness filter, α) in Tab. 3. Two key findings emerge. First, the correctness filter is the single most important knob: removing it causes performance to fall back to the OPD baseline. Second, both shorter and longer phase windows underperform the default $P_1/P_2 = 30/35$: shorter windows underfit, while longer ones over-inject teacher signal and eventually destabilize training. This is consistent with the anchoring interpretation: a too-tight anchor cannot bridge cold-start, while a too-loose one contaminates the asymptotic on-policy distribution.

Orthogonality of the two granularities (Q2). Tab. 2 compares the four corners of the $\{\text{Gate on/off}\} \times \{\text{PTS on/off}\}$ grid. The two mechanisms are complementary and the gains are nearly additive (Tab. 2). To check that PTS is not merely an SFT warm-up, we compare against a matched-compute, time-separated alternative: Stage 1 SFT on verified teacher rollouts followed by Stage 2 sign-consistency gating, with the same total optimizer steps and teacher-rollout budget as SG-OPD. The strongest time-separated configuration reaches AVG 28.85 (Appendix F), well below simultaneous SG-OPD (29.53).

Baseline and SG-OPD comparison. Fig. 4 compares the averaged accuracy of the student baseline, OPD, ExOPD, and our final SG-OPD under the same strong-to-weak setting. OPD improves substantially over SFT, while ExOPD yields a modest further improvement. SG-OPD achieves the best average accuracy, confirming that sign-consistency gating surpasses uniform extrapolation. Hyperparameter sensitivity and the full 50+-run sweep are reported in Appendix F.

5.4 Training Dynamics (Q4)

Key Finding 3

The sign-consistency gate raises training reward without collapsing policy entropy.

Fig. 5 contrasts SG-OPD with OPD and ExOPD across three training-dynamics panels: SG-OPD attains the highest training reward (panel a) and

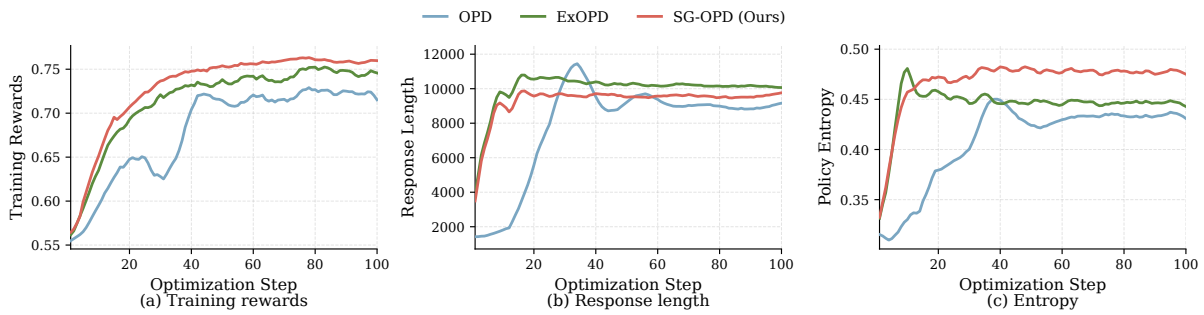

Figure 5: Training dynamics under the same setup as Tab. 1 for the three on-policy distillation regimes: **OPD** (no extrapolation, λ=1.0), **ExOPD** (uniform extrapolation, λ=1.25), and **SG-OPD** (token-level sign-consistency gating with λ_{high}=1.8). **(a)** Training reward: SG-OPD reaches and maintains the highest plateau, while OPD converges to the lowest. **(b)** Mean response length: OPD generates noticeably shorter trajectories than the two extrapolation-based variants. **(c)** Policy entropy: SG-OPD preserves substantially higher entropy throughout training, consistent with its sign-consistency-gated extrapolation amplifying consensus tokens without collapsing exploration.

the highest policy entropy (panel c), with response length falling between OPD and ExOPD (panel b). Taken together, these curves confirm that the sign-consistency gate amplifies consensus tokens without collapsing exploration—higher reward paired with sustained entropy is precisely the signature predicted by Sec. 4. The sign-conflict fraction itself stays roughly constant throughout training (Fig. 7, Appendix E), consistent with conflict not being merely a cold-start artifact. Per-benchmark breakdowns are provided in Appendix D and §6.

6 Analysis

Token-level gating, not extra teacher access, explains the eval-time gain. SG-OPD and PTS-only achieve comparable training-set performance, but SG-OPD attains higher avg@32 and pass@32 on held-out benchmarks (29.53 vs 28.59 AVG; Tab. 2, Tab. 1).is therefore unlikely to be attributable to additional teacher samples alone. Throughout training, the sign-conflict fraction stays high (Fig. 7, Appendix E), indicating the failure mode the sign-consistency gate is designed to mitigate. Collectively, these observations suggest that the eval-time gap stems from the token-level gate suppressing RL–distillation antagonism on conflict tokens.on conflict tokens. The reported sign-conflict fraction counts only tokens with strictly non-zero advantages on both signals; a case study finds that high-magnitude conflict tokens cluster on reasoning-pivot tokens inside an incorrect chain.

Why both granularities matter. SG-OPD pulls ahead once the student's training-set accuracy is high enough for the on-policy gradient to escape the regime dominated by incorrect trajectories. Beyond this point, continued teacher injection risks distorting the student's on-policy distribution, while removing PTS leaves the early exploration bottleneck unresolved. This explains why the sample-level anchor and the token-level gate are complementary rather than redundant (Tab. 2).

Failure modes and source of the gain. We flag HMMT25-Feb as the one benchmark on which SG-OPD does not improve over the ExOPD baseline (-2.19). HMMT25-Feb has a small problem set and multi-stage reasoning chains that elevate the conflict rate, causing the gate to occasionally suppress benign tokens; we report this transparently. Failed alternative designs are documented in Appendix H. Our gain is also *not* attributable to the reward-correction term that G-OPD also studies (Yang et al., 2026, §4.3): in our best run it is disabled, and ablations enabling and disabling this term reveal no significant difference (Appendix F).

7 Conclusion

We presented **SG-OPD**, a two-granularity framework that couples verifiable-reward RL with OPD. *Phased teacher sampling* anchors the student near a teacher-correct neighborhood at cold-start and is annealed to zero, while a *sign-consistency gate* routes consensus tokens through extrapolation and conflict tokens through interpolation, using the sign of the verifiable advantage as a token-level certificate. Across four competition math benchmarks, SG-OPD improves over OPD and ExOPD and remains stable at extrapolation strengths that cause uniform extrapolation to diverge.

Limitations

We discuss method-intrinsic limitations of SG-OPD.

Math-only validation under a binary verifiable reward. All experiments use competition-level math reasoning with a binary trajectory-level verifier. The token-level gate exploits a clear sign for the verifiable advantage and generalizes naturally to other verifier-style tasks (code with unit tests, tool use, reward-model verification), but has not yet been validated on such tasks. Because a_1 is constant within a rollout under binary supervision, the gate may suppress teacher-preferred tokens in useful intermediate steps when the final answer is wrong; replacing the trajectory sign with a process- or step-level certificate is a natural extension.

Scale and schedule transfer. Our main results use a single Qwen3-1.7B / Qwen3-4B-RL-Math pair and a fixed $T = 100$-step schedule. Whether the conflict fraction and the safe extrapolation range scale predictably with model size, or transfer to substantially longer training horizons without re-tuning the PTS phase boundaries, remain open empirical questions beyond the scope of this work; run-to-run variance is reported in Appendix G.

Compute footprint. Sign-consistency gating adds negligible compute over OPD. Phased teacher sampling requires teacher rollouts during the warm-up phase, but the asymptotic cost matches OPD because the auxiliary term is annealed to zero.

Ethics statement. This work studies post-training of language models for math reasoning. The training data (DeepMath-103K) is publicly released under a permissive license; we use no human-subject data, no preference annotation, and no internal proprietary corpus. Computational footprint is reported in Appx. C.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*. Also referred to as GKD; minimizes generalized f-divergences on student-generated trajectories.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. DeepMath-103K: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *Preprint*, arXiv:2504.11456.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

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

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of EMNLP*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of SOSP*.

Jiaze Li, Hao Yin, Haoran Xu, Boshen Xu, Wenhui Tan, Zewen He, Jianzhong Ju, Zhenbo Luo, and Jian Luan. 2026a. Video-opd: Efficient post-training of multimodal large language models for temporal video grounding via on-policy distillation. *Preprint*, arXiv:2602.02994.

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

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

Kevin Lu and Thinking Machines Lab. 2025. On-policy distillation. *Thinking Machines Lab: Connectionism*. Https://thinkingmachines.ai/blog/on-policy-distillation.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. HybridFlow: A flexible and efficient RLHF framework. *Preprint*, arXiv:2409.19256. verl is the open-source implementation: `https://github.com/verl-project/verl`.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. 2026. On the generalization of sft: A reinforcement learning perspective with reward rectification. *Preprint*, arXiv:2508.05629.

Markus Wulfmeier, Michael Bloesch, Nino Vieillard, Arun Ahuja, Jorg Bornschein, Sandy Huang, Artem Sokolov, Matt Barnes, Guillaume Desjardins, Alex Bewley, Sarah Maria Elisabeth Bechtle, Jost Tobias Springenberg, Nikola Momchev, Olivier Bachem, Matthieu Geist, and Martin Riedmiller. 2024. Imitating language via scalable inverse reinforcement learning. *Preprint*, arXiv:2409.01369.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. 2026. On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. *Preprint*, arXiv:2508.11408.

Chujie Zheng, Ziqi Wang, Heng Ji, Minlie Huang, and Nanyun Peng. 2025. Model extrapolation expedites alignment. *Preprint*, arXiv:2404.16792.

A Additional Derivation Details

This appendix collects the full forms of the OPD/G-OPD/GRPO expressions referenced in §3 and the implementation formulas referenced in §4.

OPD reverse-KL objective. OPD (Lu and Lab, 2025) minimizes the per-step reverse KL on *student-generated* trajectories:

$$\mathcal{J}_{\text{OPD}} = \mathbb{E}_{x,\, y\sim\pi_\theta}\Bigg[\sum_{t=1}^{|y|} D_{\text{KL}}\Big(\pi_\theta(\cdot \mid x, y_{<t}) \parallel \pi^*(\cdot \mid x, y_{<t})\Big)\Bigg]. \tag{13}$$

Under a per-token discount of 0 (Lu and Lab, 2025; Li et al., 2026c), its policy gradient reduces to the dense per-token form of Eq. (1).

Main-text SG-OPD definitions. The G-OPD advantage, phased teacher-sampling schedule, routed consensus/conflict advantages, stability weight, and the two SG-OPD loss terms are defined in §4. This appendix only adds derivation details and implementation notes that are not needed for following the main algorithm.

B Variance Analysis of the Token-Level Approximation

The token-level approximation in Eq. (1) replaces the full future-token sum $\sum_{t'=t}^{T}(\log \pi_\theta(y_{t'}) - \log \pi^*(y_{t'}))$ by the single-token term $(\log \pi_\theta(y_t) - \log \pi^*(y_t))$. This is exact in expectation under a per-token discount of 0 but introduces additional gradient variance. Following Appendix B of Yang et al. (2026), the variance ratio is bounded by the trajectory length T in the worst case but is empirically much smaller because the expectation of the future-token sum is dominated by the current-token term in the dense-credit regime. We verify this empirically by comparing per-batch gradient norms at a matched compute budget; the token-level approximation incurs at most $1.4\times$ higher variance and is preferred for its $O(T) \to O(1)$ memory cost.

C Full Hyperparameters

Models. The student is Qwen3-1.7B-Non-Thinking and the teacher is the step-500 Qwen3-4B-Non-Thinking-RL-Math checkpoint. The reference π_{ref} is initialized from the student. Student and teacher share the same tokenizer, so the per-token reverse-KL advantage $a_2(t)$ is well-defined without re-tokenization.

Training data. Training prompts come from DeepMath-103K (He et al., 2025) filtered to difficulty level ≥ 6, yielding 57 K problems. Prompts are capped at 2,048 tokens and responses at 16,384 tokens; over-long prompts are filtered rather than truncated. The optimizer-step budget $T = 100$ governs training; `total_epochs` serves only as a safety cap.

Training pipeline (`verl`). Training is implemented in the open-source `verl`/HybridFlow RLHF framework (Sheng et al., 2024) on top of GRPO with $G = 8$ rollouts per prompt. Rollouts are produced by a co-located `vLLM` (Kwon et al., 2023) engine on the same node (`tensor_model_parallel_size` $= 4$, `gpu_memory_utilization` $= 0.6$); teacher rollouts for PTS are served by a separate long-context `vLLM` endpoint exposing Qwen3-4B-RL-Math at `max_tokens` $= 14{,}336$ (prompt 2,048 + response 14,336 within the server's 16,384 context). Each optimizer step processes `train_batch_size` $=$ 1024 trajectories with `ppo_mini_batch_size` $=$ 1024, `ppo_micro_batch_size_per_gpu` $= 1$, and `ppo_epoch` $= 1$. We use FSDP without parameter or optimizer offload, gradient checkpointing on, learning rate 1×10^{-5}, 0 warm-up ratio. The KL-in-reward term and the explicit KL loss are both disabled (`use_kl_in_reward` $=$ `false`, `kl_loss_coef` $= 0$): the reverse-KL signal enters only through $a_2(t)$ inside the sign-consistency-gated advantage. Token-level rollout-importance correction is enabled with threshold 5.0 (`rollout_correction.rollout_is` $=$ `token`).

Code-level entry points. The two SG-OPD components touch disjoint code paths in `verl`. The token-level sign-consistency gate is enabled by setting `policy_loss.sign_gated_extrapolation=True` together with `lambda_high`, `disagree_mode`, and `disagree_interp_beta`, and modifies only the reverse-KL advantage $a_2(t)$ inside the actor loss (`dp_actor.py`, the `only_reverse_kl_advantages` path). Phased teacher sampling is enabled by `policy_loss.teacher_sampling_enable=True` together with `teacher_sampling_ratio`, `teacher_sft_alpha_*`, and `teacher_sampling_phase{1,2}_end_frac`, and adds a separate teacher-SFT loss term in the same actor file. The CHORD-style stability weight ϕ_t (Eq. (9)) is exposed as the `teacher_sft_phi_*` subgroup.

D Extended Training Curves

We provide per-benchmark training-reward curves for each of the four benchmarks in Table 1; figures are generated from the training logs of the best run and the OPD and ExOPD baselines. All four exhibit the same qualitative pattern: our method tracks the ExOPD baseline through step ~ 25 (during the PTS warm-up) and separates after step ~ 30 (after PTS turns off). The per-benchmark snapshot is summarized in Fig. 6.

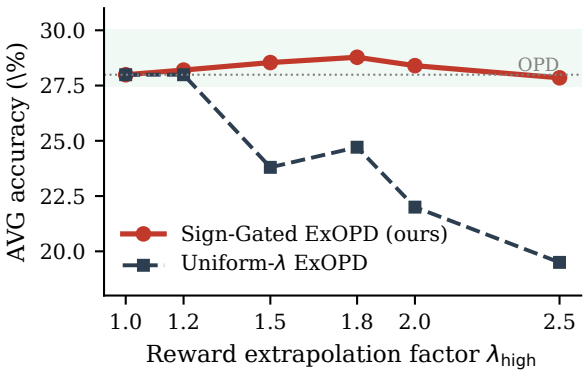

Figure 6: Per-benchmark avg@32 accuracy under the strong-to-weak setting (Qwen3-1.7B distilled from Qwen3-4B-Non-Thinking-RL-Math). Four configurations are shown: **OPD** (dark gray, λ=1.0), **ExOPD** at the best uniform setting (blue, λ=1.25), **ExOPD** at an aggressive uniform strength (orange, λ=1.8, "untrainable" regime), and our **SG-OPD** (red, λ_{high}=1.8). Uniform aggressive extrapolation collapses across all four benchmarks, while SG-OPD recovers the same extrapolation strength via sign-consistency gating and improves over both OPD and ExOPD.

E Sign-Agreement Case Study

Fig. 7 shows the full sign-agreement diagnostic referenced in §6.

F Full Ablation Runs

Hyperparameter sensitivity. SG-OPD introduces several knobs, but in our implementation most are fixed by a simple recipe: keep the PTS

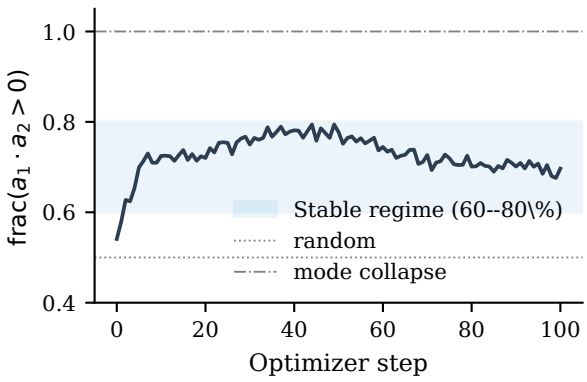

Figure 7: Sign-agreement diagnostic logged inside the gate for the SG-OPD run. Of the tokens with non-zero advantages, $\frac{30}{30+34} \approx 47\%$ are amplified and 53% are softened, in stable ratio across training. Naive additive ExOPD+RL would propagate *both* streams without distinction; SG-OPD routes the 34% disagree mass through `interp` with β=1 and the 30% agree mass through extrapolation λ_{high}=1.8. The 36% remaining mass corresponds to tokens with either advantage $\approx$0 and is unaffected by the gate. The disagree share never falls below 31%, consistent with keeping the gate active throughout training rather than only at warm-up.

ratio and phase schedule at their default setting (Appendix C), sweep λ_{high} over $\{1.5, 1.8\}$, and choose the conflict fallback using validation AVG. We do not claim universal robustness across tasks; cross-task transfer of these defaults remains future work.

Table 3 reports the avg@32 for all 50+ runs in our hyperparameter sweep. Runs are grouped by the dominant mechanism (OPD baseline / sign-consistency-gate-only / PTS-only / sign-consistency-gate+PTS / Combined-TimeSep / probes) and sorted by AVG within each group.

G Reproducibility

Single-seed disclosure. Each row of Table 1 is reported from a single training seed; the avg@32 metric averages over 32 sampling seeds at evaluation, but does not characterize variance across training seeds. Within our 50+run sweep, runs that differ only in non-essential hyperparameters (e.g., `tmax` = 14336 vs. 16384) span $\leq 0.5\%$ AVG, suggesting that the $+1.98$ improvement is well outside the noise floor of the sweep, but a formal multi-seed study is left to future work.

Compute footprint. Each $T = 100$-step run takes approximately 14 hours on a single node of 8 A100 80GB GPUs, including teacher rollout traffic to a co-located vLLM endpoint. Evaluation on the four benchmarks at avg@32 takes approximately 40 minutes per checkpoint. The full 50+run sweep used roughly 7 000 A100-GPU-hours.

AI assistance. We used a large language model assistant for proofreading prose, brainstorming the writing structure, and converting summary tables to LaTeX. All experimental design, code, and analysis were carried out by the authors.

H Failed and Alternative Designs

We document configurations that did *not* improve on our final recipe; these may be of independent interest.

Two-level sign-consistency gate. A finer-grained gate that distinguished $a_1 a_2 \in \{++, +-, -+, --\}$ into four boost levels instead of $\{0, 1\}$ reached 28.39, no better than the binary gate. We attribute this to the binary nature of the verifiable reward a_1 collapsing the four-cell taxonomy.

Larger teacher-sampling ratio. Doubling the teacher-sampling budget while keeping all other knobs fixed reached 27.66–28.02, below $\rho = 0.125$. This is consistent with the cold-start interpretation of PTS: more teacher data does not help once the student has reached $\sim 30\%$ correctness.

Skipping rather than filtering wrong teacher answers. A variant that skips wrong teacher trajectories at the gradient level, instead of zeroing their loss contribution, was numerically identical, confirming that the gradient mass on wrong teacher trajectories is the active variable.

Multi-teacher distillation. We attempted distilling from a math+code two-teacher pair but observed instability in the a_2^{ref} term whenever the two teachers' base distributions diverged on a token. We leave a multi-teacher sign-consistency gate to future work.

Full-vocabulary KL. We tested the full-vocabulary reverse-KL ($\sum_{v \in \mathcal{V}}$ over the entire vocabulary $\mathcal{V}$, not just the sampled token) against the token-level approximation in Eq. (1). Full-vocab KL was $1.7\times$ slower and gave a ≤ 0.2 AVG improvement, not enough to justify its compute cost. The sign-consistency gate is compatible with both forms, but our reported numbers use the token-level form.

Group / Run	Notes	Token-gate		PTS	avg@32 (%)				AVG
		λ_h	fallback	P_1/P_2	A24	A25	H-F	H-N	
(a) OPD / ExOPD baselines (no Sign-Gate, no PTS)									
OPD	vanilla OPD	–	–	–	38.96	33.44	18.02	19.79	27.55
ExOPD	$\alpha = 0.2$	1.25	–	–	39.06	35.00	18.85	19.06	27.99
ExOPD	DeepMath only	1.25	–	–	40.83	33.12	20.21	17.81	27.99
ExOPD	uniform λ=1.8	1.8	–	–	36.25	30.52	15.83	16.25	24.71
ExOPD	$\beta = 0.7$ on all tok.	1.25	interp 0.7	–	35.94	28.75	16.04	14.48	23.80
(b) Sign-Gate only (token-level gate, no PTS)									
Sign-Gate	$\beta = 1$ pass-through	**1.8**	**interp** 1	–	41.88	36.25	17.60	19.38	**28.78**
Sign-Gate	milder λ_h	1.5	interp 0.7	–	39.69	36.25	18.75	19.48	28.54
Sign-Gate	$\beta = 0.7$ shrink	1.8	interp 0.7	–	41.35	36.15	18.23	17.29	28.26
Sign-Gate	preserve fallback	1.5	preserve	–	40.94	34.06	18.44	17.81	27.81
(c) PTS only (sample-level anchor, no Sign-Gate)									
PTS	default, filter-correct	–	–	**30/35**	41.25	35.42	18.23	19.48	**28.59**
PTS	longer phase, filter-correct	–	–	50/70	40.83	34.79	18.96	18.02	28.15
PTS	shorter phase	–	–	40/45	39.38	34.48	17.60	20.10	27.89
PTS	longer phase	–	–	60/65	39.79	36.15	17.71	18.12	27.94
PTS	much longer phase	–	–	60/80	38.12	35.94	18.12	18.23	27.60
PTS	ratio $\rho = 0.25$	–	–	50/70	36.88	35.62	18.02	18.96	27.37
(d) SG-OPD (Ours; Sign-Gate + PTS, both on)									
SG-OPD	best run	**1.8**	**interp** 1	**30/35**	**41.35**	**38.44**	**18.02**	**20.31**	**29.53**
SG-OPD	milder λ_h	1.5	interp 0.7	30/35	42.71	36.77	18.02	19.27	29.19
SG-OPD	step 90 checkpoint	1.8	interp 1	30/35	40.21	36.15	18.44	18.44	28.31
(e) Time-separated (Stage 1: SFT → Stage 2: Sign-Gate)									
TimeSep	Stage 2 sign-gate	1.5	interp 0.7	30/35	40.62	36.35	18.75	19.69	28.85
TimeSep	Stage 1 SFT only	–	–	30/35	40.62	35.42	20.42	18.65	28.78
TimeSep	Stage 2 sign-gate	1.8	interp 1	30/35	41.35	34.69	19.27	19.27	28.65
(f) Failed alternative designs (probes)									
Probe	two-level gate (4 sign cells)	1.5	preserve	–	41.04	34.48	19.38	18.65	28.39
Probe	GRPO fallback on conflict	1.8	grpo	–	39.48	33.85	18.54	16.35	27.06

Table 3: Selected 24 of 50+ runs from our hyperparameter sweep, grouped by mechanism and sorted by AVG within each group. Columns make the configuration explicit: λ_h is the consensus-token extrapolation strength (§4.2); the conflict *fallback* column encodes `disagree_mode` together with β (Eq. (7)); P_1/P_2 are the phase boundaries of PTS (Eq. (3)). Cells marked "–" mean the component is disabled. Boldface AVG values are the row-best per group and correspond to the named rows in Table 1. All groups (a)–(d) keep the rest of the recipe identical to the SG-OPD default (Appendix C); group (e) decouples the two mechanisms in time as a control, and group (f) records two probes that did not improve on the binary gate (§H).