MOMENTUM FOR REASONING: DENSE INTRINSIC SIGNALS IN POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) has emerged as a powerful paradigm for eliciting long-chain reasoning in large language models. However, existing methods based on Group Relative Policy Optimization (GRPO) rely on a binary outcome reward, which induces two structural failure modes: *Zero-Advantage Collapse*, in which all rollouts in a group share the same outcome and the gradient vanishes, and *Hallucinated Certainty*, in which the model becomes increasingly confident on incorrect rollouts late in training. We address both modes by densifying the reward with *intrinsic* signals computed entirely from the policy's own conditional probabilities, and propose **ISPO** (***I**ntrinsic **S**ignal **P**olicy **O**ptimization*), which combines a sequence-level signal measuring how informative the thinking trajectory is for the final answer, with a token-level directional reward whose hallucinated-certainty hinge penalizes confidently-wrong predictions at critical decision tokens. Across three base models and five mathematical reasoning benchmarks, ISPO consistently outperforms competitive baselines, with the largest gains on the hardest benchmarks where zero-advantage collapse is most frequent, and training-dynamics diagnostics confirm that both failure modes are decreased.

1 INTRODUCTION

Reinforcement learning has become the central training lever for modern large language models (LLMs). Beyond aligning models with human preferences (RLHF) (Christiano et al., 2017; Brown et al., 2020; Ouyang et al., 2022; Bai et al., 2022), reinforcement learning with verifiable rewards (RLVR) replaces the human-preference model with a programmatic verifier and elicits long-chain reasoning (Guo et al., 2025; Team et al., 2025). Together with extended chain-of-thought generation (Wei et al., 2022; Kojima et al., 2022) and test-time scaling of inference compute (Snell et al., 2024; Muennighoff et al., 2025), RLVR has become the paradigm for mathematics and programming.

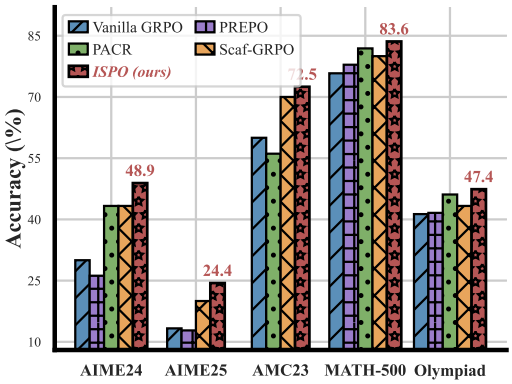


Figure 1: **ISPO leads on all five benchmarks** (Pass@1, Qwen2.5-Math-7B). Gains are largest on the hardest AIME-level sets, where zero-advantage collapse is most frequent (Table 1).

The algorithm behind these RLVR systems is mainly Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which estimates within-group relative advantages and eliminates the value function. Subsequent work refines GRPO at the optimizer level (Liu et al., 2025b; Yu et al., 2026a; Tan et al., 2025), at the data level (Zhang et al., 2025a; Sun et al., 2025), or through dense-reward shaping (Yoon et al., 2025; Li et al., 2026; Wei et al., 2026; Xiong et al., 2026; Yu et al., 2026b). Yet a binary outcome reward leaves the policy without process-level signal, and **two structural failure modes** remain unaddressed.

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.

The first, *Zero-Advantage Collapse*, arises when all rollouts in a group share the same outcome: $\sigma_R = 0$ collapses every advantage to zero, wasting the entire group, a regime that grows more frequent as easy prompts saturate and hard prompts harden (Yu et al., 2026a; Zhang et al., 2025a). The second, *Hallucinated Certainty*, emerges late in training: policy entropy at high-stakes decision tokens decreases even on *incorrect* rollouts (Wang et al., 2026; Shen et al., 2026; Wei et al., 2026), ossifying confidently-wrong predictions.

These are usually treated as two problems, and they are one. Let $\mathcal{T}_r$ denote the set of trajectories with outcome r; the outcome reward is **constant on** $\mathcal{T}_r$, invariant to which tokens a trajectory used, how confident it was at each, and whether its reasoning bore on its answer at all. Zero-Advantage Collapse is that invariance seen across a rollout group, where a constant reward centres to a zero advantage. Hallucinated Certainty is the same invariance seen along a trajectory, where no measure of confidence can change the reward and so nothing resists the entropy decline RLVR induces (Cui et al., 2025). Reading the two modes as one invariance makes the fix a construction rather than a pair of patches: it says exactly what is required, **make the reward non-constant on** $\mathcal{T}_r$ at the granularities at which trajectories in $\mathcal{T}_r$ differ (§3.1).

Resolving both modes calls for densifying the reward itself rather than further refining the optimizer. A natural source is the policy's own conditional probabilities, which we call *intrinsic* signals, which require no external reward model or process annotation. Intrinsic signals have a long history in RL (intrinsic motivation and curiosity-driven exploration (Pathak et al., 2017; Burda et al., 2018; Bellemare et al., 2016)) and have recently been advocated as a generalist reward source for LLMs (Li et al., 2025). We make three observations that **target the specific failure modes** of RLVR (Figure 2; §3.2): same-outcome rollouts differ substantially in how strongly the thinking shapes the answer; useful gradient is concentrated on a small set of high-leverage tokens (Wang et al., 2026); and on incorrect rollouts the entropy at these tokens declines faster than on correct ones. Together, these observations show that intrinsic signals can densify sparse rewards and keep learning moving when vanilla GRPO stalls. Loosely analogous to optimization momentum (Sutskever et al., 2013; Kingma & Ba, 2014), this dense intrinsic signal is a *momentum for reasoning* that carries the policy through regions where the outcome gradient vanishes.

Based on these, we propose **ISPO** (**I**ntrinsic **S**ignal **P**olicy **O**ptimization), which augments the outcome reward with two intrinsic signals. The sequence-level *Conditional IFD* measures how much the thinking trajectory sharpens the policy's own answer distribution; in expectation its unnormalized form is exactly the conditional KL divergence between the policy's answer distribution with and without the thinking trajectory (Proposition 1). Crucially, it varies across rollouts that share an outcome, so it supplies within-group reward variance exactly where the binary outcome supplies none. The token-level *directional reward*, with a hinge penalty $\max(0, \tau_H - H_t)$, asymmetrically shapes critical tokens by outcome to suppress confidently-wrong predictions. Both terms enter coupled to the outcome sign $s = 2R_{\text{o}} - 1$, and that coupling is the part of the design we consider load-bearing (Appendix B.4): a dense intrinsic bonus applied without it is the configuration whose long-horizon behaviour is known to degrade below the starting model (Zhang et al., 2025b; Jin et al., 2026), because it rewards a property the policy can inflate on its own. Coupled to the sign, inflating that property on an incorrect rollout lowers its reward instead, so the only way to profit from the intrinsic term is to make it discriminate correct reasoning from incorrect. Across three models and five mathematical reasoning benchmarks, ISPO consistently outperforms competitive baselines, with the largest gains on the hardest benchmarks where zero-advantage collapse is most frequent.

Our contributions are as follows:

- We reframe the two structural failure modes of binary-outcome RLVR as a single invariance: the outcome reward is constant on trajectories that share an outcome, which specifies what any fix must do (§3.1).
- We propose ISPO, which augments the outcome reward with two intrinsic signals: a sequence-level Conditional IFD, in expectation a conditional KL divergence (Proposition 1, §3.3), and a token-level directional reward at critical tokens (§3.4). Both couple to the outcome sign to stay safe to optimise (Appendix B.4) and preserve within-group variance except under a knife-edge cancellation (Proposition 2).
- Across three models and five math benchmarks, ISPO achieves superior performance over strong baselines, with the gap widening under multi-sample evaluation (§4).

2 Related Work

RL with verifiable rewards and dense reward shaping. RL with verifiable rewards (Guo et al., 2025), powered by GRPO (Shao et al., 2024) and its optimizer-level refinements (Liu et al., 2025b; Yu et al., 2026a), is the dominant paradigm for LLM reasoning. To compensate for the sparse outcome, a parallel line augments the reward with policy-intrinsic signals (stepwise confidence (Yoon et al., 2025), low-probability-token confidence (Li et al., 2026), entropy trends (Wei et al., 2026; Xiong et al., 2026), or entropy gating at decision pivots (Yu et al., 2026b)), while an older line supervises intermediate steps with a learned process reward model at the price of annotation and a gameable second model (Lightman et al., 2024; Wang et al., 2024). ISPO differs on both terms: its token-level signal is *outcome-conditioned* and restricted to critical tokens rather than a uniform bonus, and its sequence-level signal measures *thinking–answer dependence* with a conditional-KL reading in expectation (Proposition 1), extending Li et al. (2024) from data selection to online RL.

Reward-model-free intrinsic signals. Closest in spirit is the line that drops the verifier and optimizes a policy-internal quantity: Zhao et al. (2025) elicit reasoning from self-certainty alone. That line has a documented long-horizon failure (Zhang et al. (2025b) and Jin et al. (2026) report that entropy- and certainty-based proxies match RLVR early but eventually degrade below the starting model) because an intrinsic objective satisfiable *without being correct* is eventually satisfied that way. ISPO's response is structural: the sequence term enters multiplied by the outcome sign $2R_{\mathrm{o}} - 1$, so on an incorrect rollout a high $\mathrm{IFD}_{\mathrm{cond}}$ is *subtracted* and no intrinsically-high, outcome-wrong attractor exists (Appendix B.4); Appendix C.5 tests this past our main budget.

Token-level credit assignment. Token-level methods exploit the heavy-tailed gradient distribution (Wang et al., 2026) through cross-rollout variance (Xu et al., 2025), selective penalization (Liu et al., 2025a), or token selection: Lv et al. (2026) keep only tokens inside a stable surprisal interval. This is the sharpest contrast to ISPO: token selection changes which tokens the gradient reaches and leaves the reward untouched, whereas ISPO leaves the token set untouched and changes what the *reward* says at critical tokens, so the two compose. Nearest to our sequence term is DRO (Xu et al., 2025), which scores a *reference* answer's certainty; Conditional IFD instead scores the sampled answer and subtracts the trace-ablated baseline $\log p_\theta(A \mid Q)$, which is what admits the KL reading of Proposition 1. Data-curation methods (Zhang et al., 2025a; Yu et al., 2026a; Sun et al., 2025) attack zero-advantage batches from outside the reward and compose orthogonally.

3 Method

3.1 Preliminaries

Group Relative Policy Optimization. GRPO (Shao et al., 2024) trains a policy π_θ by sampling G rollouts per prompt Q. Each rollout is a trajectory $\tau_i = (Q, T_i, A_i)$ that decomposes into a thinking trace T_i and a final answer A_i, and we let $o_i = T_i \oplus A_i$ denote the generated output. A verifier assigns a binary outcome reward $R_{\mathrm{o}}^{(i)} \in \{0, 1\}$ to each rollout, and the group-relative advantage is $\hat{A}^{(i)} = \frac{R^{(i)} - \mu_R}{\sigma_R}$, where μ_R and σ_R are the mean and standard deviation of $\{R^{(j)}\}_{j=1}^{G}$. The policy is then updated via the standard clipped surrogate objective:

$$J_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{Q\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|Q)} \left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \min\Big(r_{i,t}(\theta)\,\hat{A}^{(i)},\ \mathrm{clip}\big(r_{i,t}(\theta),\,1-\epsilon,\,1+\epsilon\big)\hat{A}^{(i)} \Big) \right], \quad (1)$$

where $r_{i,t}(\theta) = \pi_\theta(o_{i,t} \mid Q, o_{i,<t}) / \pi_{\theta_{\mathrm{old}}}(o_{i,t} \mid Q, o_{i,<t})$ is the per-token importance ratio.

Two structural failure modes. Sparse binary reward induces two structural failure modes:

- **Failure Mode 1 (Zero-Advantage Collapse).** When all G rollouts in a group receive the same binary outcome, the centered reward $R^{(i)} - \mu_R$ is zero for every rollout and the group provides no policy-gradient signal to the GRPO surrogate, despite full rollout computation

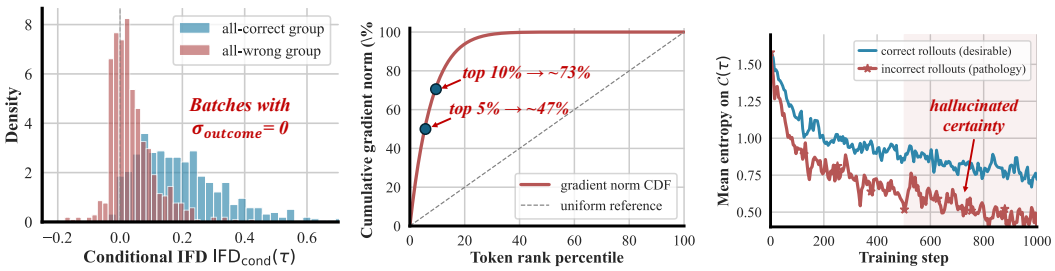


Figure 2: **Three signals missed by sparse outcomes** (Qwen2.5-Math-7B, vanilla-GRPO checkpoint; Appendix A.2). (a) In outcome-homogeneous groups ($\sigma_{\text{outcome}} = 0$), Conditional IFD still separates all-correct from all-wrong. (b) Cumulative gradient norm is Lorenz-like (top 10% of tokens carry $\sim$70%), captured by an entropy $\cap$ base-drift filter. (c) At critical tokens, entropy on *incorrect* rollouts settles below correct ones: hallucinated certainty (shaded).

having been consumed. This case becomes more frequent as training progresses: ***easy prompts saturate to all-correct, while hard prompts remain all-wrong.***

- **Failure Mode 2 (Hallucinated Certainty).** A symmetric pathology emerges in late-stage training: policy entropy at high-stakes decision tokens decreases over time even on *incorrect* rollouts. ***The model becomes confidently wrong, suppressing exploration and locking in erroneous reasoning patterns.***

The two failure modes are one property of the outcome reward. Writing $\mathcal{T}_r = \{\tau : R_{\text{o}}(\tau) = r\}$ for the trajectories with outcome r, R_{o} is **constant on** $\mathcal{T}_r$: invariant to which tokens a trajectory used, how confident it was at each, and whether its reasoning bore on its answer. Failure Mode 1 is that invariance seen *across* a group, where all G rollouts land in one $\mathcal{T}_r$ and the centred advantage is identically zero; Failure Mode 2 is the same invariance seen *along* a trajectory, where no confidence signal can change R_{o} and the entropy decline RLVR induces proceeds on incorrect rollouts unopposed. The requirement on any fix is then explicit: *make the reward non-constant on* $\mathcal{T}_r$, at the granularities at which trajectories differ. ISPO does so with a sequence-level signal (how much the thinking moves the answer distribution, §3.3) and a token-level signal (confidence at critical decision tokens, §3.4), each aggregated to one scalar per trajectory so GRPO's advantage structure is untouched; since $\sigma_R \geq |\lambda_1\sigma_{\text{I}} - \lambda_2\sigma_{\text{t}}|$, breaking the invariance at either granularity suffices, and degeneracy returns only through the knife-edge cancellation of the two terms (Proposition 2).

3.2 Why Dense Intrinsic Signals?

Three observations motivate the design of our method (Figure 2).

Obs. 1: Same-outcome rollouts differ in thinking-answer dependence. Within rollout groups sharing the same outcome, trajectories still vary substantially in how strongly the thinking T_i shapes the answer A_i. Figure 2(a) plots the within-group Conditional IFD distribution (§3.3) restricted to $\sigma_{\text{outcome}} = 0$ batches: all-correct groups (blue) form a right-skewed distribution with positive tail extending past 0.4, while all-wrong groups (red) concentrate near zero with a small negative tail. The two distributions are clearly separated, and the within-group standard deviation $\sigma_{\text{IFD}} > 0$ in over 95% of zero-advantage batches. $\mathrm{IFD}_{\mathrm{cond}}$ therefore carries informative within-group variance *even when the outcome reward does not*, and we use it as the sequence-level reward (§3.3).

Obs. 2: Gradient is concentrated on a small set of critical tokens. The per-token gradient norm is heavy-tailed. Figure 2(b) plots the cumulative gradient norm against token rank percentile: the top 5% of tokens carry $\sim$50% of the gradient, and the top 10% carry $\sim$70%, far above the diagonal uniform reference and consistent with Wang et al. (2026). Furthermore, this high-leverage subset is well captured by the intersection of two filters, local entropy ($\mathcal{A}$) and base-policy drift ($\mathcal{C}_\Delta$); either filter alone admits substantially more low-signal tokens. This motivates restricting the token-level reward to the critical subset $\mathcal{A} \cap \mathcal{C}_\Delta$.

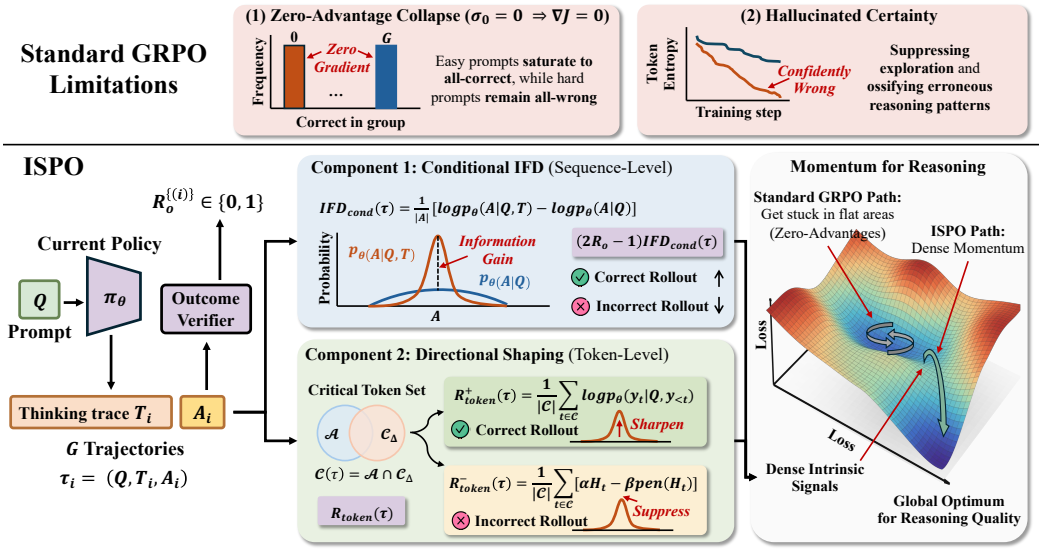


Figure 3: **ISPO framework.** For each prompt Q, π_θ samples G rollouts and augments the sparse outcome reward with two intrinsic signals from its own log-probabilities: a sequence-level Conditional IFD and a token-level directional reward at critical tokens. The composite $R^{(i)}$ enters GRPO's advantage; nothing else changes.

Obs. 3: Incorrect rollouts grow more confident, not less. Figure 2(c) tracks the mean entropy at critical tokens over training, split by rollout outcome. Both curves decline monotonically from ∼1.6 at initialization, but the incorrect-rollout curve (red) drops faster and settles *below* the correct-rollout curve (blue): by convergence the model is more confident on its wrong critical-token predictions than on its right ones. This is the signature of *hallucinated certainty* (shaded late-stage region), a pathology invisible to the binary outcome reward, and motivates an outcome-conditioned, directional token-level reward.

3.3 Sequence-Level Signal: Conditional IFD

Motivated by the Instruction Following Difficulty (IFD) metric for supervised data selection (Li et al., 2024), we define the *Conditional IFD* of a rollout as the per-token log-probability gap between predicting the answer with and without conditioning on the thinking trajectory:

$$\mathrm{IFD}_{\mathrm{cond}}(\tau) = \tfrac{1}{|A|}\big[\log p_\theta(A|Q,T) - \log p_\theta(A|Q)\big]. \tag{2}$$

A large positive value indicates that conditioning on T substantially increases the policy's likelihood of A; a value near zero indicates that the answer is largely insensitive to T. Note that $\mathrm{IFD}_{\mathrm{cond}}$ measures *thinking–answer dependence*, not correctness; correctness is supplied separately by the verifier through R_{o}. Here $p_\theta(A \mid Q)$ denotes the likelihood of the same answer under a trace-ablated prompt, evaluated by teacher forcing, and should be interpreted as a prompted answer distribution rather than the true marginal obtained by integrating over latent thinking traces.

Proposition 1 (Conditional-KL identity in expectation)**.** *For any fixed (Q, T), when the answer A is drawn from the policy itself, $A \sim p_\theta(\cdot|Q,T)$,*

$$\mathbb{E}_{A\sim p_\theta(\cdot|Q,T)}\big[\log p_\theta(A|Q,T) \;-\; \log p_\theta(A|Q)\big] \;=\; D_{\mathrm{KL}}\big(p_\theta(\cdot|Q,T)\,\|\,p_\theta(\cdot|Q)\big) \;\geq\; 0, \tag{3}$$

with equality if and only if $p_\theta(\cdot|Q,T) = p_\theta(\cdot|Q)$, i.e. the trace-ablated prompt induces exactly the same answer distribution as the full prompt.

Proof. Denote the LHS as L. By the definition of expectation and KL divergence:

$$L = \sum_A p_\theta(A|Q,T)\,\log\frac{p_\theta(A|Q,T)}{p_\theta(A|Q)} \tag{4}$$

$$= D_{\mathrm{KL}}\big(p_\theta(\cdot|Q,T)\,\|\,p_\theta(\cdot|Q)\big). \tag{5}$$

Non-negativity follows from Gibbs' inequality, with equality iff $p_\theta(\cdot|Q,T) = p_\theta(\cdot|Q)$. □

The identity holds *in expectation*: a single rollout's unnormalized score is an unbiased estimate of this KL, not itself a divergence, and may be negative. Both arguments are *prompted* distributions of the same policy (with $p_\theta(\cdot|Q)$ a trace-ablated prompt, not the true marginal), so (3) concerns sensitivity to the trace, not conditional mutual information; Appendix B.3 states these qualifications in full.

3.4 Token-Level Signal: Directional Shaping

While $\mathrm{IFD}_{\mathrm{cond}}$ captures sequence-level quality, it is silent about *where* in the trajectory the learning signal lives. ISPO adds a fine-grained reward at the token level, restricted to a small subset of *critical tokens* and shaped asymmetrically by the outcome.

Critical token selection. The critical token set $\mathcal{C}(\tau) = \mathcal{A} \cap \mathcal{C}_\Delta$ is the intersection of an entropy filter and a drift filter:

$$\mathcal{A} = \big\{\, t : H_t \geq \text{top-}k\% \text{ within } \tau \,\big\}, \tag{6}$$

$$\mathcal{C}_\Delta = \big\{\, t : \big|\log p_\theta(y_t) - \log p_{\theta_0}(y_t)\big| \geq \tau_\Delta \,\big\}, \tag{7}$$

where H_t denotes the entropy of the next-token distribution $\pi_\theta(\cdot \mid Q, y_{<t})$ at position t, and p_{θ_0} is the frozen base policy. Filter $\mathcal{A}$ selects locally uncertain tokens; filter $\mathcal{C}_\Delta$ retains only tokens where RL training has meaningfully moved the distribution away from the base. In practice $\mathcal{C}(\tau)$ covers $5-10\%$ of trajectory tokens, consistent with the heavy-tailed gradient distribution observed by Wang et al. (2026).

Why intersection, not union? Either filter alone is inadequate: $\mathcal{A}$ in isolation admits intrinsically uncertain tokens that RL never shaped (punctuation, surface phrasing), and $\mathcal{C}_\Delta$ in isolation admits tokens that RL did change but at which the model is now nearly deterministic. The intersection isolates tokens that are both *uncertain* and *actively shaped*, the high-leverage subset for directional shaping.

Directional reward. The desired behavior at critical tokens differs by outcome. Motivated by Observation 3 of §3.2, ISPO defines an outcome-conditioned token-level reward:

$$R^{+}_{\mathrm{token}}(\tau) = \tfrac{1}{|\mathcal{C}|}\sum_{t\in\mathcal{C}} \log p_\theta(y_t \mid Q, y_{<t}), \tag{8}$$

$$R^{-}_{\mathrm{token}}(\tau) = \tfrac{1}{|\mathcal{C}|}\sum_{t\in\mathcal{C}} \big[\alpha H_t - \beta\,\mathrm{pen}(H_t)\big], \tag{9}$$

where $\mathrm{pen}(H_t) = \max(0, \tau_H - H_t)$ is a hinge penalty on low-entropy tokens. The full token-level reward is then

$$R_{\mathrm{token}}(\tau) = \begin{cases} R^{+}_{\mathrm{token}}(\tau), & \text{correct, } |\mathcal{C}| > 0, \\ R^{-}_{\mathrm{token}}(\tau), & \text{incorrect, } |\mathcal{C}| > 0, \\ 0, & |\mathcal{C}| = 0. \end{cases} \tag{10}$$

Correct rollouts are rewarded for low perplexity at critical tokens, reinforcing fluent and confident execution of correct reasoning. **Incorrect rollouts** are rewarded for high entropy, with the hinge term $\mathrm{pen}(H_t)$ activating *only* when entropy falls below τ_H. The two terms together penalize confidently-wrong predictions while leaving uncertain incorrect tokens lightly affected, directly targeting the hallucinated-certainty regime of Failure Mode 2.

The incorrect-rollout entropy bonus cannot be farmed by staying uncertain instead of solving the prompt; Appendix B.5 bounds the swing below the outcome gap and states the ranges it uses.

3.5 Composite Reward

Combining the two intrinsic signals with the outcome reward yields the ISPO composite:

$$R(\tau) = R_{\mathrm{o}} + \lambda_1\,(2R_{\mathrm{o}} - 1)\,\mathrm{IFD}_{\mathrm{cond}}(\tau) + \lambda_2\,R_{\mathrm{token}}(\tau), \tag{11}$$

where $\lambda_1, \lambda_2 > 0$ and $s \triangleq 2R_{\text{o}} - 1 \in \{-1, +1\}$. The composite replaces R_{o} in GRPO's advantage; no other pipeline component changes (Figure 3). Both signals come from π_θ's own log-probabilities (*intrinsic*) and compose with GRPO variants and sampling-side techniques (*optimizer-agnostic*); reward-flow and stop-gradient details are in Appendix B.2. Under mild non-degeneracy the dense terms preserve $\sigma_R > 0$ even when $\sigma_{\text{outcome}} = 0$ (Appendix B.5, which also relates ISPO to potential-based reward shaping (Ng et al., 1999)); the outcome sign s is load-bearing, and Appendix B.4 analyses why an unsigned bonus degrades over a long horizon while the signed version stays safe to optimise.

Table 1: **Main results on math reasoning benchmarks** (Pass@1 %; **Avg** = five-benchmark mean). Best per column per block **bold**, **ISPO** shaded. Markers: unmarked = three-seed runs; * = single-seed reruns (unified protocol, greedy T=0); † = transcribed under source protocols ("-" = not reported). Only unmarked/* comparisons are controlled (Table 13).

Method	AIME24	AIME25	AMC23	MATH-500	Olympiad-Bench	Avg
Qwen2.5-Math-1.5B						
Base model†	7.2	3.3	32.5	33.0	22.8	19.8
Vanilla GRPO (Shao et al., 2024)	13.3	10.0	47.5	72.2	34.8	35.6
Dr. GRPO* (Liu et al., 2025b)	16.7	10.0	50.0	74.0	36.6	37.5
PACR† (Yoon et al., 2025)	23.3	-	49.4	77.4	39.0	-
Scaf-GRPO† (Zhang et al., 2025a)	20.0	13.3	60.0	73.4	36.6	40.7
PREPO* (Sun et al., 2025)	13.3	10.0	45.0	73.2	34.2	35.1
DAPO* (Yu et al., 2026a)	23.3	13.3	55.0	76.4	38.5	41.3
ISPO (ours)	**27.8**	**22.2**	**62.5**	**79.2**	**40.3**	**46.4**
Qwen2.5-Math-7B						
Base model†	16.7	13.3	38.6	50.6	16.6	27.2
Vanilla GRPO (Shao et al., 2024)	30.0	13.3	60.0	75.8	41.3	44.1
Dr. GRPO* (Liu et al., 2025b)	33.3	13.3	62.5	80.6	44.0	46.7
PACR† (Yoon et al., 2025)	43.3	-	56.1	81.9	46.1	-
Scaf-GRPO† (Zhang et al., 2025a)	43.3	20.0	70.0	80.0	43.3	51.3
PREPO* (Sun et al., 2025)	26.7	13.3	57.5	77.8	40.6	43.2
DAPO* (Yu et al., 2026a)	40.0	16.7	65.0	81.4	44.6	49.5
ISPO (ours)	**48.9**	**24.4**	**72.5**	**83.6**	**47.4**	**55.4**
Qwen2.5-7B						
Base model†	5.4	2.5	32.4	54.7	24.6	23.9
Vanilla GRPO† (Shao et al., 2024)	9.2	6.1	65.5	75.3	35.6	38.3
Dr. GRPO* (Liu et al., 2025b)	13.3	6.7	67.5	77.0	36.4	40.2
ProGRPO† (Li et al., 2026)	21.3	15.9	67.2	80.5	42.7	45.5
Scaf-GRPO† (Zhang et al., 2025a)	13.3	20.0	60.0	77.8	40.8	42.4
PREPO* (Sun et al., 2025)	10.0	6.7	62.5	75.8	36.0	38.2
DAPO* (Yu et al., 2026a)	26.7	16.7	67.5	80.8	40.1	46.4
GTPO† (Tan et al., 2025)	34.2	16.7	-	80.2	-	-
GRPO-S† (Tan et al., 2025)	**34.7**	18.3	-	80.1	-	-
ISPO (ours)	32.2	**23.3**	**68.3**	**83.0**	**43.7**	**50.1**

4 Experiments

4.1 Experimental Setup

Models and baselines. We evaluate ISPO on three Qwen2.5 base models spanning math-specialized and general-purpose families: Qwen2.5-Math-1.5B, Qwen2.5-Math-7B (Yang et al., 2024), and Qwen2.5-7B, all trained from the base version following Guo et al. (2025). We compare against (i) the **Base model** without RL, (ii) vanilla **GRPO** (Shao et al., 2024), (iii) **Dr. GRPO** (Liu et al., 2025b) with normalization debiasing, (iv) **ProGRPO** (Li et al., 2026), (v) **Scaf-GRPO** (Zhang et al., 2025a), (vi) **PACR** (Yoon et al., 2025) (its Dense variant), and (vii) **PREPO** (Sun et al., 2025).

Table 2: **Ablation of ISPO components** on Qwen2.5-Math-7B (Avg Pass@1 over five benchmarks, three-seed mean, greedy). Blocks: (a) the two signals, (b) critical-token selection, (c) directional shaping.

Variant	AIME24	AIME25	AMC23	MATH-500	Olympiad	Avg
(a) Signal components						
GRPO	30.0	13.3	60.0	75.8	41.3	44.1
+ IFD	41.1	18.9	65.8	79.7	43.3	49.8
+ token	38.9	15.6	63.3	78.0	42.4	47.6
ISPO	**48.9**	**24.4**	**72.5**	**83.6**	**47.4**	**55.4**
(b) Critical token selection						
$\mathcal{A}$ only	43.3	21.1	70.0	80.4	45.6	52.1
$\mathcal{C}_\Delta$ only	41.1	18.9	67.5	80.0	44.9	50.5
$\mathcal{A} \cup \mathcal{C}_\Delta$	46.7	22.2	71.7	81.3	46.0	53.6
$\mathcal{A} \cap \mathcal{C}_\Delta$ (ours)	**48.9**	**24.4**	**72.5**	**83.6**	**47.4**	**55.4**
(c) Directional shaping						
Symmetric	38.9	15.6	63.3	78.7	43.4	48.0
$\beta = 0$	43.3	21.1	68.3	80.7	44.9	51.7
(ours)	**48.9**	**24.4**	**72.5**	**83.6**	**47.4**	**55.4**

Training data. We train on **DAPO-Math** (Yu et al., 2026a), a curated corpus of approximately 17K verifiable mathematical reasoning problems. This is the same training set used by ProGRPO (Li et al., 2026), allowing direct comparison without confounding from data-side variation. We follow the DAPO training recipe; full hyperparameters and the sensitivity analysis are in Appendices A.2–C.6.

Benchmarks. We report Pass@1 accuracy on five mathematical reasoning benchmarks: **AIME 2024** (Math-AI, 2024) and **AIME 2025** (Math-AI, 2025) (30 problems each), **AMC 2023** (Math-AI, 2023) (40), **MATH-500** (Lightman et al., 2024) (the 500-problem subset), and **OlympiadBench** (He et al., 2024) (675, math-only English subset), with greedy decoding ($T = 0$) following Yoon et al. (2025). As AIME24/AIME25/AMC23 are small, numbers for methods we train ourselves are means over three training seeds; single-seed reruns are marked * in Table 1. Per-seed spreads, multi-sample metrics (avg@16 and pass@32 at $T = 0.6$), and a DAPO-Math overlap audit are in Appendix C.

4.2 Main Results and Ablation

Main results. Table 1 reports Pass@1 accuracy across the five benchmarks for all three base-model families. ISPO achieves the best Avg on every base model, improving Avg Pass@1 over the strongest baseline that reports all five benchmarks by 5.1 points on Qwen2.5-Math-1.5B, 4.1 on Qwen2.5-Math-7B, and 3.7 on Qwen2.5-7B. On two of the three that baseline is DAPO, trained under our unified protocol, so those comparisons are controlled; on Qwen2.5-Math-7B ISPO leads the published state of the art, Scaf-GRPO, by 4.1 points and our own strongest run, DAPO, by 5.9. Every retrained baseline is set beside the published value it replaces in Appendix C.8. Gains concentrate on the hardest benchmarks (AIME24/AIME25), where zero-advantage collapse is most frequent, and are smaller on the more saturated MATH-500 and OlympiadBench, as the failure-mode thesis predicts. ISPO matches or exceeds data-curation methods (Scaf-GRPO, PREPO) without modifying the rollout pipeline, and stays competitive with dense-reward methods (PACR, ProGRPO) while keeping σ_R bounded away from zero under outcome homogeneity (Appendix B.5).

Component ablation. Table 2 isolates each ISPO component on Qwen2.5-Math-7B. The two-signal block shows Conditional IFD alone restores gradient under outcome homogeneity and directional token shaping alone targets late-stage hallucinated certainty, with their combination strictly better than either. The selection block shows the intersection $\mathcal{A} \cap \mathcal{C}_\Delta$ concentrates signal where either filter alone or their union admits low-signal tokens. The shaping block shows a symmetric reward and a variant without the hallucinated-certainty hinge ($\beta = 0$) both underperform the full directional reward. Remaining experiments are in Appendix C.

5 Analysis

Gains track difficulty. The invariance argument of §3.1 makes a falsifiable prediction: if ISPO's gain comes from restoring reward variance inside $\mathcal{T}_r$, it should concentrate on prompts where the outcome reward is most often constant across the group and be near-absent where it already varies. Bucketing all evaluation prompts by base-model pass rate over 8 samples (Table 3), ISPO's gain over the strongest baseline is $+0.4$ on *Easy* prompts (base solves $\geq 6/8$), $+1.1$ on *Medium*, and $+4.4$ on *Hard* prompts ($0/8$, i.e. groups vanilla GRPO scores at $\sigma_R = 0$ every time). The near-zero Easy gain is as informative as the large Hard one: a gain spread evenly across buckets would indicate a different mechanism. This monotone trend is the empirical signature of Failure Mode 1, connecting the variance preservation of Proposition 2 to learning on batches that would otherwise be discarded.

Table 3: **Difficulty-bucketed Pass@1 on Qwen2.5-Math-7B.** Prompts split by base-model pass rate over 8 samples: Easy ($\geq 6/8$), Medium ($1/8$–$5/8$), Hard ($0/8$). ISPO's gain grows with difficulty, largest on Hard.

Method	Easy	Medium	Hard
Vanilla GRPO	84.6	53.2	12.4
Dr. GRPO	86.1	57.8	14.7
PACR	86.8	60.3	19.2
Scaf-GRPO	85.4	58.6	22.3
ISPO (ours)	**87.2**	**61.4**	**26.7**
ISPO gain vs. best	+0.4	+1.1	**+4.4**

Mechanism diagnostics. ISPO holds the zero-advantage batch fraction near zero where GRPO and Dr. GRPO let it climb (Figure 4, left), keeping outcome-homogeneous batches in the update at the cost of one extra forward pass (Table 5); the recovered batches are productive (Appendix C.10). It holds critical-token entropy on incorrect rollouts up (Figure 4, right), shortening them from a $2.14\times$ length ratio to $1.32\times$ (Table 7). These signals act on the 5–10% of tokens carrying the heavy-tailed gradient of Wang et al. (2026) (Table 8). The advantage widens under test-time sampling: on AIME24 the margin over ProGRPO grows from 10.9 points at greedy Pass@1 to 13.4 at pass@32 (Table 6), as the entropy-preserving reward keeps the answer pool diverse.

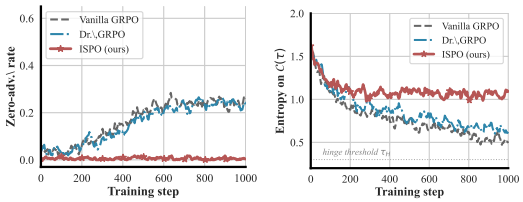


Figure 4: **Training dynamics (Qwen2.5-Math-7B).** *Left:* zero-advantage batch fraction climbs for GRPO/Dr. GRPO but ISPO stays near zero (Prop. 2). *Right:* critical-token entropy on incorrect rollouts decays into hallucinated certainty under GRPO, held up by ISPO (App. B.5).

Long-horizon stability. The sign coupling (Appendix B.4) keeps this dense reward safe to optimise over a long horizon. Extending training to $3\times$ the main budget, ISPO sustains its lead while an unsigned variant rises then degrades as the RLIF literature reports (Zhang et al., 2025b; Jin et al., 2026), and a sign-only baseline (the fixed-reference estimator of Nie et al. (2026)) improves on GRPO without reaching ISPO. Appendix C.5 reports the four-arm study and the early-answer-leakage diagnostic that pins the mechanism.

6 Conclusion

We presented **ISPO**, a reward-level augmentation to GRPO that addresses two structural failure modes of binary-outcome RLVR, *Zero-Advantage Collapse* and *Hallucinated Certainty*, by densifying the reward with two intrinsic signals derived from the policy's own conditional probabilities: a sequence-level Conditional IFD, in expectation a conditional-KL divergence, that preserves within-group reward variance under zero-advantage collapse except on a knife-edge, and a token-level directional reward with a hallucinated-certainty hinge that suppresses confidently-wrong predictions at critical tokens. Across three models and five benchmarks, ISPO consistently outperforms competitive baselines, with the largest gains on the hardest AIME-level benchmarks. The momentum-like effect of dense intrinsic signals points to a broader principle: where outcome rewards are sparse, the model's own probability surface often holds enough information to keep learning moving.

AI USE STATEMENT

We used generative AI tools as a coding and writing assistant: to help debug training and evaluation code, and to improve the clarity and grammar of the manuscript. All AI-assisted code was read, executed, and verified against logged outputs before any reported number was derived from it, and all AI-assisted prose was reviewed and edited by the authors. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work trains and evaluates language models on public mathematical-reasoning benchmarks, used within their intended terms; it involves no human subjects, no personally identifying information, and no new data collection or release, and we audit train–test overlap in Appendix C.14. On dual use: ISPO's efficiency gain is objective-agnostic, as with most optimizer-level contributions, and its token-level reward shapes confidence only in coupling with verified correctness, paired with an entropy-promoting term on incorrect rollouts precisely to avoid the fluent-while-wrong failure mode we call hallucinated certainty (Figure 2(c), §5).

REPRODUCIBILITY STATEMENT

The full training configuration (hyperparameters, optimizer and clipping settings, hardware, and benchmark subsets) is in Appendices A and A.2, with pseudocode in Appendix B.1 and the reward-computation and stop-gradient details in Appendix B.2. Because $\mathrm{IFD}_{\mathrm{cond}}$ is sensitive to prompt formatting, Appendix C.12 gives the trace-ablated prompt verbatim. Defaults were frozen on a held-out development split (Appendix C.6); per-seed scores and the standard error of the difference are in Appendix C.7, the measured σ_R distribution in Appendix C.13, and the train–test overlap audit in Appendix C.14. Proofs of Propositions 1 and 2 are in §3.3 and Appendix B.5.

REFERENCES

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting

agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

Xiaolong Jin, Xuandong Zhao, Wenbo Guo, Xiangyu Zhang, and Dawn Song. When do intrinsic rewards work for code reasoning? A comprehensive study. *arXiv preprint arXiv:2606.20881*, 2026.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7602–7635, 2024.

Pengyi Li, Elizaveta Goncharova, Andrey Kuznetsov, and Ivan Oseledets. Back to basics: Revisiting exploration in reinforcement learning for llm reasoning via generative probabilities. *arXiv preprint arXiv:2602.05281*, 2026.

Yi-Chen Li, Tian Xu, Yang Yu, Xuqin Zhang, Xiong-Hui Chen, Zhongxiang Ling, Ningjing Chao, Lei Yuan, and Zhi-Hua Zhou. Generalist reward models: Found inside large language models. *arXiv preprint arXiv:2506.23235*, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Hanbing Liu, Lang Cao, Yuanyi Ren, Mengyu Zhou, Haoyu Dong, Xiaojun Ma, Shi Han, and Dongmei Zhang. Not all tokens matter: Towards efficient LLM reasoning via token significance in reinforcement learning. *arXiv preprint arXiv:2506.08125*, 2025a.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025b.

Outongyi Lv, Yanzhao Zheng, Yuanwei Zhang, Zhenghao Huang, Xingjun Wang, Baohua Dong, Hangcheng Zhu, and Yingda Chen. Which tokens matter? adaptive token selection for RLVR with the relative surprisal index. *arXiv preprint arXiv:2606.31575*, 2026.

Math-AI. Amc 2023. `https://huggingface.co/datasets/math-ai/amc23`, 2023.

Math-AI. Aime 2024. `https://huggingface.co/datasets/math-ai/aime24`, 2024. URL `https://huggingface.co/datasets/math-ai/aime24`.

Math-AI. Aime 2025. `https://huggingface.co/datasets/math-ai/aime25`, 2025.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20286–20332, 2025.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287. Citeseer, 1999.

Wenhua Nie, Jianan Wu, Junlin Liu, Ziwei Li, Zheng Lin, Zhang Zijian, Yilong Fan, Haoran Zheng, and Jyh-Shing Roger Jang. Gradient starvation in binary-reward GRPO: Why group-mean centering fails and why the simplest fix works. *arXiv preprint arXiv:2605.07689*, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

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

Zhanming Shen, Jiaqi Hu, Zeyu Qin, Hao Chen, Wentao Ye, Zenan Huang, Yihong Zhuang, Guoshan Lu, Junlin Zhou, and Junbo Zhao. Training-trajectory-aware token selection. *arXiv preprint arXiv:2601.10348*, 2026.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Yan Sun, Jia Guo, Stanley Kok, Zihao Wang, Zujie Wen, and Zhiqiang Zhang. Efficient reinforcement learning for large language models with intrinsic exploration. *arXiv preprint arXiv:2511.00794*, 2025.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. pmlr, 2013.

Hongze Tan, Zihan Wang, Jianfei Pan, Jinghao Lin, Hao Wang, Yifan Wu, Tao Chen, Zhihang Zheng, Zhihao Tang, and Haihua Yang. Gtpo and grpo-s: Token and sequence-level reward shaping with policy entropy. *arXiv preprint arXiv:2508.04349*, 2025.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *Advances in Neural Information Processing Systems*, 38:115452–115486, 2026.

Chengwei Wei, Jung-jae Kim, Longyin Zhang, Shengkai Chen, and Nancy F Chen. Infodensity: Rewarding information-dense traces for efficient reasoning. *arXiv preprint arXiv:2603.17310*, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Xuan Xiong, Huan Liu, Li Gu, Zhixiang Chi, Yue Qiu, Yuanhao Yu, and Yang Wang. Etr: Entropy trend reward for efficient chain-of-thought reasoning. *arXiv preprint arXiv:2604.05355*, 2026.

Yifei Xu, Tusher Chakraborty, Srinagesh Sharma, Leonardo Nunes, Swati Sharma, Kate Drakos Demopulos, Emre Kıcıman, Songwu Lu, and Ranveer Chandra. Direct reasoning optimization: Token-level reasoning reflectivity meets rubric gates for unverifiable tasks. *arXiv preprint arXiv:2506.13351*, 2025.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Eunseop Yoon, Hee Suk Yoon, Jaehyun Jang, SooHwan Eom, Qi Dai, Chong Luo, Mark A Hasegawa-Johnson, and Chang D Yoo. Pacr: Progressively ascending confidence reward for llm reasoning. *arXiv preprint arXiv:2510.22255*, 2025.

Prasanth YSS, Zhichen Ren, Rasa Hosseinzadeh, Ilan Gofman, Yuqi Chen, Zhaoyan Liu, Guangwei Yu, Jesse C. Cresswell, and Satya Krishna Gorti. A gradient perspective on RLVR stability and winner advantage policy optimization. *arXiv preprint arXiv:2606.16154*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026a.

Song Yu, Li Li, Wenwen Zhao, and Zhisheng Yang. Erpo: Token-level entropy-regulated policy optimization for large reasoning models. *arXiv preprint arXiv:2603.28204*, 2026b.

Xichen Zhang, Sitong Wu, Yinghao Zhu, Haoru Tan, Shaozuo Yu, Ziyi He, and Jiaya Jia. Scaf-grpo: Scaffolded group relative policy optimization for enhancing llm reasoning. *arXiv preprint arXiv:2510.19807*, 2025a.

Yanzhi Zhang, Zhaoxi Zhang, Haoxiang Guan, Yilin Cheng, Yitong Duan, Chen Wang, Yue Wang, Shuxin Zheng, and Jiyan He. No free lunch: Rethinking internal feedback for LLM reasoning. *arXiv preprint arXiv:2506.17219*, 2025b.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*, 2025.

A Experimental Setup

A.1 Datasets

We evaluate ISPO on five publicly available mathematical reasoning benchmarks spanning competition-level to graduate-level difficulty. Table 4 summarizes the datasets and problem counts. For AIME 2024, AIME 2025, and AMC 2023, we evaluate on the full problem set released by the corresponding competitions. For MATH-500, we use the 500-problem subset curated by Lightman et al. (2024). For OlympiadBench, we use the math-only subset (text-only, English) following the protocol of recent reasoning RL papers (Yoon et al., 2025; Zhang et al., 2025a).

Table 4: Overview of evaluation datasets. Pass@1 is the fraction of problems solved correctly.

Dataset	**# Problems**
AIME 2024 (Math-AI, 2024)	30
AIME 2025 (Math-AI, 2025)	30
AMC 2023 (Math-AI, 2023)	40
MATH-500 (Lightman et al., 2024)	500
OlympiadBench (He et al., 2024)	675

A.2 Full Training Configuration

Following Yu et al. (2026a), we omit the KL penalty and use a clipping range of $[0.8, 1.28]$ (Clip-Higher). For each prompt we sample $G = 8$ rollouts at $T = 1.0$ with maximum completion length 4096 (Qwen2.5-Math) or 8192 (Qwen2.5-7B). The actor is optimized with AdamW at learning rate 1×10^{-6}, batch size 512, and 1 update per minibatch. ISPO-specific hyperparameters are set to $\lambda_1 = 0.5$, $\lambda_2 = 0.1$, top-$k\% = 10\%$, $\tau_\Delta = 0.3$, $\alpha = 1.0$, $\beta = 1.0$, and $\tau_H = 0.3$. These defaults were selected on a held-out validation split of 500 problems from the DAPO-Math training corpus, never on the evaluation benchmarks, and applied unchanged to all base models and benchmarks. All training runs use $8\times$ A100 80GB GPUs. Unless a table states otherwise, every number reported for methods we train ourselves is the mean over three independent training seeds evaluated with greedy decoding; rows marked *, including the baseline reruns in Table 1, were trained once each. Per-seed standard deviations are reported in Appendix C.7.

B Method Details

B.1 ISPO Training Algorithm

Algorithm 1 summarizes one ISPO training step. ISPO requires one additional forward pass per rollout, over the answer tokens A_i without the thinking context T_i, in order to compute $\log p_\theta(A_i|Q)$. Since in practice $|A_i| \ll |T_i|$ (the thinking trajectory dominates the rollout length), the overhead is small; in our experiments the wall-clock cost increment is under 5% relative to vanilla GRPO.

B.2 Reward Computation and Gradient Flow

All intrinsic scores in ISPO are computed under the rollout policy $\pi_{\theta_{\text{old}}}$ and treated as stop-gradient scalar rewards when optimizing the GRPO surrogate. At training iteration k, we sample rollouts with π_{θ_k}, compute IFD_{cond}, token entropies, and base-policy drift using π_{θ_k} and the frozen base policy π_{θ_0}, and then detach the resulting rewards before forming the GRPO objective. The policy parameters update only through the importance ratio $r_{i,t}(\theta)$ in the clipped surrogate, so the dense terms act as reward shaping rather than auxiliary differentiable objectives. At the next iteration, the intrinsic scores are recomputed under the updated rollout policy.

Algorithm 1 ISPO training step

Require: Prompt Q; policy π_θ; frozen base π_{θ_0}; group size G
Require: Weights λ_1, λ_2; token hyperparams $k, \tau_\Delta, \alpha, \beta, \tau_H$
Sample G rollouts $\{(Q, T_i, A_i)\}_{i=1}^{G} \sim \pi_\theta$
for $i = 1, \ldots, G$ **do**
  Compute outcome reward $R_{\mathrm{o}}^{(i)} \in \{0, 1\}$ via verifier
  Read off $\log p_\theta(A_i|Q, T_i)$ from the rollout pass
  Compute $\log p_\theta(A_i|Q)$ ▷ one extra forward pass over A_i
  $\mathrm{IFD}_{\mathrm{cond}}^{(i)} \leftarrow |A_i|^{-1}[\log p_\theta(A_i|Q, T_i) - \log p_\theta(A_i|Q)]$
  Compute $\mathcal{C}^{(i)} = \mathcal{A}^{(i)} \cap \mathcal{C}_\Delta^{(i)}$ via Eqs. 6–7
  if $R_{\mathrm{o}}^{(i)} = 1$ **then**
    $R_{\mathrm{token}}^{(i)} \leftarrow R_{\mathrm{token}}^{+}(\tau_i)$ via Eq. 8
  else
    $R_{\mathrm{token}}^{(i)} \leftarrow R_{\mathrm{token}}^{-}(\tau_i)$ via Eq. 9
  end if
  $R^{(i)} \leftarrow R_{\mathrm{o}}^{(i)} + \lambda_1(2R_{\mathrm{o}}^{(i)} - 1)\,\mathrm{IFD}_{\mathrm{cond}}^{(i)} + \lambda_2 R_{\mathrm{token}}^{(i)}$
end for
Compute group advantages $\hat{A}^{(i)}$ from $R^{(i)}$
Update θ via the clipped surrogate objective on $\{\hat{A}^{(i)}\}$

B.3 Scope and Interpretation of Proposition 1

Three qualifications govern how we use Proposition 1. *(i)* It holds *in expectation* over $A \sim p_\theta(\cdot|Q, T)$: the unnormalized $|A| \cdot \mathrm{IFD}_{\mathrm{cond}}(\tau)$ is an unbiased single-sample estimate of that KL, not itself a divergence and possibly negative, and the length-normalized form (2) rescales it, so neither the non-negativity in (3) nor KL-unbiasedness holds per rollout. *(ii)* Both arguments are *prompted* distributions of the same policy, with $p_\theta(\cdot|Q)$ the trace-ablated prompt (Appendix C.12) rather than the marginal $\sum_T p_\theta(\cdot|Q, T)p_\theta(T|Q)$, so (3) concerns sensitivity to the trace in the policy's own context, not conditional mutual information. *(iii)* The identity fixes (Q, T) while the reward (11) evaluates it at the sampled T, so the group sees a per-rollout estimate of a rollout-dependent quantity, exactly what supplies within-group variance; the proposition thus fixes the *interpretation and sign* of the sequence-level signal, not a guarantee about any single update.

B.4 The Outcome Sign Is the Load-Bearing Part

The factor s is what makes the dense term safe to optimize. Replace $+\lambda_1 s\,\mathrm{IFD}_{\mathrm{cond}}(\tau)$ with an unsigned bonus and the intrinsic term rewards a property the policy can inflate on its own: the configuration whose long-horizon collapse is documented (§2), where training matches RLVR and then degrades below the *starting* model (Zhang et al., 2025b; Jin et al., 2026).

Two properties of ISPO close the routes to that attractor. *Group centering* removes uniform inflation: raising $\mathrm{IFD}_{\mathrm{cond}}$ on *all* of a group's trajectories shifts no advantage. *The sign* removes selective inflation on wrong rollouts: the sequence term contributes $\lambda_1(\overline{\mathrm{IFD}}_{\mathrm{correct}} - \overline{\mathrm{IFD}}_{\mathrm{wrong}})$ up to centering, so raising $\mathrm{IFD}_{\mathrm{cond}}$ on an incorrect rollout strictly lowers its reward, and the token term applies the same decision at a second granularity. Because Nie et al. (2026) show the fixed-reference sign advantage is itself a strong estimator at small group size, with token-level analyses tying stability to the sign (YSS et al., 2026), Appendix C.5 runs an unsigned variant and a sign-only baseline past our main budget to separate the safety claim from the attribution.

B.5 Behavioral Analysis

A useful property of ISPO is that its behavior can be characterized cleanly under three regimes that span the typical training trajectory.

Case 1: Heterogeneous outcomes ($\sigma_{\mathrm{outcome}} > 0$). When the rollout group contains both correct and incorrect trajectories, R_{o} already provides a discriminative signal. The dense terms enter as intrinsic refinements that adjust the within-group ranking without overturning the outcome-driven

sign of the advantage. In this regime ISPO recovers GRPO up to a refinement perturbation, and the choice of λ_1, λ_2 controls the strength of refinement relative to the outcome.

Case 2: Zero-advantage collapse ($\sigma_{\text{outcome}} = 0$). When all G rollouts share the same outcome, vanilla GRPO produces zero gradient. Under ISPO the outcome term is constant across the group, so both R_{o} and its sign factor $s \triangleq 2R_{\text{o}} - 1 \in \{-1, +1\}$ are constants, and the group-level variance of the composite reward (11) decomposes as

$$\sigma_R^2 = \lambda_1^2 \operatorname{Var}(\text{IFD}_{\text{cond}}) + \lambda_2^2 \operatorname{Var}(R_{\text{token}}) + 2\, s\, \lambda_1 \lambda_2 \operatorname{Cov}(\text{IFD}_{\text{cond}}, R_{\text{token}}). \quad (12)$$

The sign factor squares away in the two variance terms ($s^2 = 1$) but survives on the cross term: in an all-*wrong* group ($s = -1$) a *positive* correlation between the two intrinsic signals *reduces* σ_R, whereas in an all-correct group it increases it. Writing $\sigma_{\text{I}} \triangleq \text{sd}(\text{IFD}_{\text{cond}})$, $\sigma_{\text{t}} \triangleq \text{sd}(R_{\text{token}})$ and $\rho \triangleq \text{Corr}(\text{IFD}_{\text{cond}}, R_{\text{token}})$, completing the square in (12) gives the identity

$$\sigma_R^2 = \big(\lambda_1 \sigma_{\text{I}} + s\rho\, \lambda_2 \sigma_{\text{t}}\big)^2 + \lambda_2^2 \sigma_{\text{t}}^2 \big(1 - \rho^2\big), \quad (13)$$

in which both summands are non-negative. This makes the exact degeneracy condition readable, which is the content of the following proposition.

Proposition 2 (Zero-advantage degeneracy is a knife-edge condition)**.** *Fix a rollout group with $\sigma_{\text{outcome}} = 0$ and let $\lambda_1, \lambda_2 > 0$. Then $\sigma_R = 0$ if and only if one of the following holds:*

1. ***Total collapse:** both intrinsic signals are constant across the group, $\sigma_{\text{I}} = \sigma_{\text{t}} = 0$; or*
2. ***Exact cancellation:** the two signals are perfectly anti-aligned after sign correction, $s\rho = -1$,* and *their weighted spreads match exactly, $\lambda_1 \sigma_{\text{I}} = \lambda_2 \sigma_{\text{t}}$.*

Moreover, for every group,

$$\big|\lambda_1 \sigma_{\text{I}} - \lambda_2 \sigma_{\text{t}}\big| \;\le\; \sigma_R \;\le\; \lambda_1 \sigma_{\text{I}} + \lambda_2 \sigma_{\text{t}}. \quad (14)$$

Proof. Both terms of (13) are non-negative, so $\sigma_R = 0$ requires each to vanish. If $\sigma_{\text{t}} = 0$ the second term vanishes automatically and the first reduces to $\lambda_1^2 \sigma_{\text{I}}^2$, giving total collapse. If $\sigma_{\text{t}} > 0$ the second term forces $\rho^2 = 1$, and the first then forces $\lambda_1 \sigma_{\text{I}} = -s\rho\, \lambda_2 \sigma_{\text{t}}$; since the left side is non-negative and $|s\rho| = 1$, this requires $s\rho = -1$ together with $\lambda_1 \sigma_{\text{I}} = \lambda_2 \sigma_{\text{t}}$, giving exact cancellation. Conversely each case makes both summands zero. The bracket (14) is the triangle inequality and its reverse applied to the centred variables $\lambda_1 s\, \text{IFD}_{\text{cond}}$ and $\lambda_2 R_{\text{token}}$ in L^2, using $\text{sd}(\lambda_1 s\, \text{IFD}_{\text{cond}}) = \lambda_1 \sigma_{\text{I}}$ because $|s| = 1$. □

Scope of Proposition 2. Proposition 2 is an exact characterisation, not an almost-sure guarantee. The stronger claim $\Pr[\sigma_R > 0] = 1$ does not hold: rollouts are discrete token sequences, so no measure-zero argument applies, and $\sigma_R^2 \ge \lambda_1^2 \text{Var}(\text{IFD}_{\text{cond}})$ is unavailable because the covariance term in (12) can be negative. Degeneracy instead requires *either* both signals exactly constant (total collapse) *or* the two-part coincidence of exact cancellation. Total collapse is the operative risk: it is approached late in training as π_θ peaks and within-group rollouts grow similar, precisely when zero-advantage batches are most frequent. Exact cancellation bears on the weight choice $\lambda_1/\lambda_2 = 5$ (Appendix C.6), which puts $\lambda_1 \sigma_{\text{I}}$ and $\lambda_2 \sigma_{\text{t}}$ on the same order and keeps the lower end of the bracket (14) small; the additional requirement $s\rho = -1$ is never observed. Since neither case can be excluded by assumption, the question is settled empirically: Appendix C.13 reports the measured σ_R distribution over outcome-homogeneous batches with its lower tail, and Table 5 reports the observed incidence of $\sigma_R < 10^{-6}$.

Non-zero is not outcome-aligned. Proposition 2 concerns a *non-zero*, not necessarily *outcome-aligned*, gradient: in an all-wrong group the recovered signal ranks rollouts by thinking–answer dependence and critical-token entropy rather than by correctness, acting as exploration and calibration pressure rather than direct error correction. Since the per-rollout IFD_{cond} is a single-sample Monte-Carlo estimate of the conditional KL, part of its variance is estimator noise; we keep λ_1 small so this refinement never overwhelms the outcome signal when present (Case 1). Appendix C.10 shows that the recovered updates are nonetheless productive in practice.

Case 3: Late-stage hallucinated certainty. Late in training, π_θ becomes peaked at most tokens, including critical ones. For correct rollouts, this is desirable: the model is sharpening confidence in valid reasoning. For incorrect rollouts, however, the same peaking yields confidently wrong predictions, the failure mode that purely outcome-driven RL cannot diagnose, since the outcome reward treats all wrong trajectories identically. ISPO's incorrect-rollout reward (9) counteracts this: the hinge term $\mathrm{pen}(H_t) = \max(0, \tau_H - H_t)$ becomes *active* precisely when $H_t < \tau_H$, applying a corrective signal exactly where hallucinated certainty has set in. Across the three cases, ISPO matches GRPO where it suffices, preserves the gradient where GRPO fails, and applies a corrective signal where late-stage GRPO degrades.

Robustness to reward gaming. Two properties keep the incorrect-rollout entropy bonus subordinate to solving the prompt. First, an incentive bound: an incorrect rollout can outrank a correct one only if $\lambda_1 \cdot (\text{signed-IFD gap}) + \lambda_2 \cdot (\text{token-term gap})$ exceeds the outcome gap of 1. The observed signal ranges ($|\mathrm{IFD}_{\mathrm{cond}}| \lesssim 0.5$, Appendix C.9; critical-token entropy $\lesssim 1.7$ nats; correct-rollout critical-token log-probabilities $\gtrsim -2$) bound the two gaps separately by $2 \max |\mathrm{IFD}_{\mathrm{cond}}| = 1.0$ and $\max R^-_{\mathrm{token}} - \min R^+_{\mathrm{token}} \leq 1.7 + 2 = 3.7$, so the largest swing under the default weights is $\lambda_1(1.0) + \lambda_2(3.7) = 0.5 + 0.37 = 0.87 < 1$, a margin of 0.13: the outcome ordering is never inverted, so the entropy bonus can reorder incorrect rollouts only among themselves. As a weight constraint, the inequality gives $\lambda_1 < 1 - 3.7\lambda_2 = 0.63$ and $\lambda_2 < (1 - \lambda_1)/3.7 = 0.135$; the defaults are the largest sweep values satisfying it (Table 10). The optimum is interior: the three violating points ($\lambda_1 = 1.0$, $\lambda_1 = 2.0$, $\lambda_2 = 0.30$) score below the default, as do $\lambda_1 = 0.1$ and $\lambda_2 = 0.05$. Second, the preserved entropy is a usable resource. At step 300 of the vanilla GRPO run, 87 of the 512 prompts in a batch (17.0%) are all-wrong, a strict subset of the 24.7% zero-advantage incidence in Table 5. Splitting them into terciles of 29 by mean critical-token entropy and resampling 32 completions each ($T = 0.6$) gives solve@32 rates of 6.9% (low-entropy tercile), 31.0% (middle) and 58.6% (high-entropy tercile), with avg@32 of 0.5%, 4.8% and 18.3%: recoverability rises with entropy, and confidently-wrong prompts are nearly unrecoverable. Preserved critical-token entropy is thus the reserve the directional reward protects; it plateaus near 1.05 rather than climbing to its ceiling (Figure 4), and the symmetric ablation rewarding entropy regardless of outcome is worse (Table 2).

Relation to potential-based reward shaping. Because ISPO modifies the reward, the relevant classical result is that of Ng et al. (1999): a shaping term F preserves the optimal policy for *every* choice of dynamics if and only if it is potential-based, $F(s, a, s') = \gamma\Phi(s') - \Phi(s)$ for some Φ with $\Phi(s_{\mathrm{terminal}}) = 0$. Neither ISPO term has that form, so we claim no policy invariance for (11). Under GRPO, however, that class is inert: GRPO assigns one trajectory-level advantage (§3.1), and under the undiscounted episodic return of RLVR a potential-based term telescopes to $\Phi(s_{\mathrm{terminal}}) - \Phi(Q) = -\Phi(Q)$, a function of the prompt alone that group-mean centering removes (the uniform-inflation algebra of Appendix B.4). Such shaping thus changes no advantage, gradient, or policy, and any dense reward that does affect the update lies outside the class. What replaces invariance is the bound above, an inequality over *measured* signal ranges rather than a theorem about all inputs, hence a condition to monitor; its 0.13 margin is a further reason to keep λ_1, λ_2 small.

C Additional Empirical Analysis

C.1 Training Efficiency and Zero-Advantage Recovery

Table 5 reports the fraction of training batches that suffer zero-advantage collapse under each method, alongside per-step wall-clock cost. Vanilla GRPO and Dr. GRPO discard rollouts as $\sigma_R \to 0$; DAPO's Dynamic Sampling rejects such batches before training but pays a sampling-side cost. ISPO neither discards nor over-samples: the dense terms keep σ_R measurably above zero where the outcome reward alone is degenerate, so those batches stay in the update, at an overhead bounded by one extra forward pass over $|A_i| \ll |T_i|$. Appendix B.5 characterises when this can fail (Proposition 2), and the reclaimed batches are productive, not merely non-zero: ISPO eventually solves many prompts that start all wrong (Appendix C.10).

Table 5: **Training efficiency on Qwen2.5-Math-7B.** Shaded row ours (no best-in-bold; two columns are lower-is-better). "Zero-adv. batches" = fraction with $\sigma_R < 10^{-6}$, "useful rollouts" its complement. DAPO reaches 0.0% at a 45% wall-clock cost; ISPO reaches it within budget at $\sim 4.4\%$ overhead. Appendix B.5 gives the condition under which σ_R can vanish.

Method	Sec. /step ↓	Zero-adv. batches (%) ↓	Useful rollouts (%) ↑
Vanilla GRPO	18.2	24.7	75.3
Dr. GRPO	18.1	23.9	76.1
DAPO (Dyn. Sampl.)	26.4	0.0	100.0
ISPO (ours)	19.0	0.0	100.0

Table 6: **Multi-sample evaluation on Qwen2.5-7B.** avg@16 = mean Pass@1 over 16 samples at $T = 0.6$; pass@32 = fraction solved by at least one of 32. All rows * (run by us), so fully controlled. The ISPO gap over the strongest baseline widens from avg@16 to pass@32, the regime most affected by Failure Mode 2; Table 15 extends it to the math models, Appendix C.15 adds diversity and an accuracy-matched control.

Method	AIME24		MATH-500	
	avg@16	pass@32	avg@16	pass@32
Vanilla GRPO*	14.4	46.7	78.6	92.4
ProGRPO*	26.7	53.3	82.8	94.2
Scaf-GRPO*	22.5	50.0	80.1	93.6
ISPO (ours)	**35.0**	**66.7**	**84.0**	**95.6**

C.2 Multi-sample Evaluation

Table 6 reports avg@16 and pass@32 on Qwen2.5-7B, where all four rows are our own runs. On AIME24 ISPO reaches 35.0 avg@16 and 66.7 pass@32, against 26.7 and 53.3 for ProGRPO, the strongest baseline in the table; the margin grows from 8.3 points at avg@16 to 13.4 at pass@32, and the same widening over ProGRPO appears on both math-specialized models (Table 15). Pass@32 is the regime most directly affected by Failure Mode 2: a policy that has collapsed onto confidently-wrong reasoning produces 32 nearly-identical wrong samples, while ISPO's entropy-promoting incorrect-rollout reward R^{-}_{token} preserves entropy at critical tokens. Appendix C.15 measures diversity directly and adds an accuracy-matched control in which ISPO is handicapped to GRPO's avg@16; at equal accuracy ISPO retains part but not all of its pass@32 advantage, which therefore reflects both retained diversity and higher per-sample accuracy (Table 21).

C.3 Reasoning Length Analysis

Liu et al. (2025b) identify a systematic length-inflation bias in vanilla GRPO whereby incorrect rollouts grow much longer than correct ones, producing wasteful generations without accuracy gains. Table 7 compares average rollout length at convergence: vanilla GRPO shows a $2.14\times$ incorrect/correct ratio, while ISPO is the most balanced at $1.32\times$. The hinge penalty in R^{-}_{token} discourages spending generation budget on confidently-wrong reasoning, so length is suppressed where it would not pay off, with no explicit length penalty. ISPO's balance is achieved by shortening incorrect rollouts ($2876 \rightarrow 1742$) rather than by lengthening correct ones ($1342 \rightarrow 1318$, essentially unchanged), the asymmetry the directional reward is designed to produce rather than a uniform brevity pressure. This table also fixes the length *scale* at our main budget and so calibrates the length column of Table 9: a mean response length is a pass-rate-weighted mix of the two columns here, which places GRPO near 2200 tokens and ISPO near 1507 at step 600, the reference against which *shortening* at $3\times$ the budget reads as collapse rather than as efficiency (Appendix C.5).

C.4 Critical-Token Set Size

Table 7: **Reasoning length on Qwen2.5-Math-7B** at convergence: mean tokens on correct vs. incorrect rollouts. †Length ratio (incorrect/correct): GRPO inflates incorrect-rollout length, whereas ISPO's hinge penalty yields the most balanced distribution without an explicit length penalty.

Method	Avg. length (tokens)		Length ratio†
	Correct	Incorrect	
Vanilla GRPO (Shao et al., 2024)	1342	2876	2.14
Dr. GRPO (Liu et al., 2025b)	1289	1834	1.42
PACR (Yoon et al., 2025)	1276	2241	1.76
ISPO (ours)	**1318**	**1742**	**1.32**

Table 9: **Long-horizon behaviour at $3\times$ the main budget** (Qwen2.5-Math-7B). "Avg" = five-benchmark mean; "len" = mean response length. At step 600 the unmarked entries (44.1, 55.4) and the base reference (27.2) are the Table 1 values. **ISPO, unsigned** drops the outcome sign; **GRPO + Sign** uses $A = 2r - 1$ (Nie et al., 2026). Length is the leading indicator of intrinsic-reward collapse, accuracy the lagging one (Jin et al., 2026).

	step 600 (main)		**step 1200**		**step 1800** ($3\times$)	
Method	Avg ↑	len	Avg ↑	len	Avg ↑	len
Base (no RL)	27.2	760	-	-	-	-
Vanilla GRPO	44.1	2200	44.6	2480	43.9	2690
GRPO + Sign $A{=}2r{-}1$	46.2	2150	47.0	2320	46.4	2410
ISPO, unsigned	53.1	1290	50.2	1010	46.8	702
ISPO (ours)	55.4	1507	**56.8**	1552	**57.1**	1571

Table 8 reports the size of the critical-token set $\mathcal{A} \cap \mathcal{C}_\Delta$ across training stages: it stabilizes at 5–10% of trajectory tokens, matching the heavy-tailed gradient distribution observed by Wang et al. (2026).

Table 8: **Critical-token set size across training** on Qwen2.5-Math-7B (top-$k\% = 10$, $\tau_\Delta = 0.3$). The drift filter $\mathcal{C}_\Delta$ is empty at step 0 and expands with training while $\mathcal{A}$ stays at 10%; the intersection $\mathcal{A} \cap \mathcal{C}_\Delta$ stabilises at ∼5–10% of tokens, matching the heavy-tailed gradient of Wang et al. (2026).

Training stage	$\lvert\mathcal{A}\rvert$ (%)	$\lvert\mathcal{C}_\Delta\rvert$ (%)	$\lvert\mathcal{A} \cap \mathcal{C}_\Delta\rvert$ (%)
Step 0 (base policy)	10.0	0.0	0.0
Early (10% steps)	10.0	18.6	4.3
Mid (50% steps)	10.0	41.2	7.8
Late (100% steps)	10.0	52.7	8.6

C.5 Long-Horizon Survivability and What the Sign Buys

Appendix B.4 argues the outcome sign keeps a dense intrinsic reward safe over long horizons: purely intrinsic objectives fail not immediately but by succeeding first and degrading later (Zhang et al., 2025b; Jin et al., 2026). Table 9 extends training to $3\times$ our main budget with four arms: **ISPO, unsigned** tests whether removing the sign reproduces that collapse, and **GRPO + Sign** (the fixed-reference estimator of Nie et al., 2026) tests whether the sign alone accounts for ISPO's gains. If the coupling prevents the RLIF collapse, the unsigned arm rises then degrades while ISPO holds; if the intrinsic signals contribute beyond the estimator, ISPO leads GRPO + Sign.

Unsigned $\mathrm{IFD}_{\mathrm{cond}}$ is maximised by making the answer maximally predictable from the trace, and the cheapest way to do that involves no reasoning at all: state the answer early and let the remainder of the trace be vestigial. This predicts a specific degeneracy, and it explains why output shortening is the reported signature (Jin et al., 2026): a trace that only has to announce its conclusion need not be long. We therefore log the fraction of rollouts whose final answer string already appears within the first 20% of the trace. Under ISPO this stays flat, because on an incorrect rollout early commitment is penalised rather than rewarded; under the unsigned arm it climbs. This leakage rate holds near 6% under ISPO across the run while climbing from 7% at step 600 to 34% by step 1800 under the unsigned arm, tracking that arm's length collapse in Table 9.

Table 10: **Hyperparameter sensitivity on Qwen2.5-Math-7B. Dev** is a held-out 500-problem split of DAPO-Math on which defaults were selected once (Appendix A.2), then frozen. **Test** is Avg Pass@1 over the five benchmarks; only orderings transfer. The dev argmax is the default in all six blocks, and dev and test rankings have Spearman's $\rho = 0.83$ (blocks (e),(f) each swap a pair). No setting falls below the GRPO floor (44.1).

Configuration	Dev (500)	Test (Avg. Pass@1)
(a) Sequence-level weight λ_1		
$\lambda_1 = 0$ (no IFD)	54.8	47.6
$\lambda_1 = 0.1$	59.2	51.6
$\lambda_1 = 0.5$ (default)	**62.4**	**55.4**
$\lambda_1 = 1.0$	61.6	54.7
$\lambda_1 = 2.0$	58.0	52.3
(b) Token-level weight λ_2		
$\lambda_2 = 0$ (no token reward)	57.0	49.8
$\lambda_2 = 0.05$	60.8	53.9
$\lambda_2 = 0.10$ (default)	**62.4**	**55.4**
$\lambda_2 = 0.30$	61.2	54.2
(c) Critical-token fraction $k\%$		
top-5%	61.0	53.6
top-10% (default)	**62.4**	**55.4**
top-15%	61.4	54.8
top-20%	59.8	53.3
(d) Drift threshold τ_Δ		
$\tau_\Delta = 0.1$	61.2	54.3
$\tau_\Delta = 0.3$ (default)	**62.4**	**55.4**
$\tau_\Delta = 0.5$	60.6	53.5
(e) Hinge threshold τ_H		
$\tau_H = 0.1$	61.8	54.0
$\tau_H = 0.3$ (default)	**62.4**	**55.4**
$\tau_H = 0.5$	60.4	54.6
(f) Penalty weights (α, β)		
$(\alpha, \beta) = (1.0, 0.0)$ (no hinge)	58.2	51.7
$(\alpha, \beta) = (1.0, 0.5)$	61.0	54.5
$(\alpha, \beta) = (1.0, 1.0)$ (default)	**62.4**	**55.4**
$(\alpha, \beta) = (1.0, 2.0)$	61.6	54.2

C.6 Hyperparameter Sensitivity

Table 10 reports ISPO's sensitivity to each hyperparameter on Qwen2.5-Math-7B. The sweep is one-at-a-time around the default configuration, which recurs as the reference row in each of the six blocks, so the 23 table rows correspond to 18 distinct configurations. Across all 18 configurations swept, ISPO never drops below 47.6 Avg Pass@1 (default 55.4) and stays above the vanilla GRPO baseline of 44.1. The default choice of $\lambda_1 = 0.5 > \lambda_2 = 0.1$ reflects that the sequence-level Conditional IFD is the primary driver of ISPO's gain: it acts on every rollout and is the term whose expectation carries the conditional-KL interpretation (Proposition 1), while the token-level reward serves as a fine-grained refinement targeting hallucinated certainty (Failure Mode 2) at the $\sim$5–10% of critical tokens. The $5\times$ ratio between λ_1 and λ_2 also normalizes the raw scales of the two signals (the per-token IFD gap is typically $\sim 10\times$ smaller in magnitude than the token-level reward), so their actual contributions to $R(\tau)$ are comparable. The dominant failure modes are *(i)* for λ_1, λ_2 too small, ISPO collapses toward GRPO (e.g., $\lambda_1 = 0$ yields 47.6); *(ii)* for λ_1, λ_2 too large, the dense signals overpower the outcome reward (e.g., $\lambda_1 = 2.0$ yields 52.3); *(iii)* for $k\%$ or τ_Δ too low, the critical-token set becomes too sparse to deliver token-level gradient; too high, it dilutes the signal across non-critical tokens.

C.7 Multi-seed Robustness

Table 11 reports the mean and sample standard deviation of avg@16 accuracy across three independent training runs that differ only in random seed, measuring how sensitive each method is to seed selection. All other hyperparameters are held fixed at the defaults specified in Appendix C.6. It is the twin of Table 14: same base model, metric, four methods and three benchmarks, with the $\pm$ taken over training seeds here and over the sixteen decoding draws of a single run there. The pair separates the two sources of variance a single number conflates.

Table 11: **Robustness to the training seed** on Qwen2.5-Math-7B (avg@16 at $T = 0.6$; mean $\pm$ sample sd over three seeds). Same runs, metric and benchmarks as Table 14, differing only in whether the $\pm$ is taken over decoding draws (there) or seeds (here). Greedy is avoided because with 30 problems a three-seed greedy sd can only be 0 or ≥ 1.9 points, too coarse for a $\sim$1-point seed spread. Using $\mathrm{SE}_\Delta = \sqrt{\sigma_1^2 + \sigma_2^2}/\sqrt{3}$, ISPO's AIME24 margin over DAPO (the strongest baseline we train) is 8.6 against $\mathrm{SE}_\Delta = 0.98$ ($t \approx 8.7$, $p < 0.01$); per-seed scores are in Appendix C.7.

Method	AIME24	AIME25	AMC23
Vanilla GRPO	26.5$\pm$1.4	11.7$\pm$1.2	55.9$\pm$1.5
Dr. GRPO*	28.7$\pm$1.1	12.2$\pm$1.0	57.8$\pm$1.3
DAPO*	35.6$\pm$1.3	14.6$\pm$1.1	61.3$\pm$1.4
ISPO (ours)	**44.2**$\pm$1.1	**21.9**$\pm$1.0	**68.8**$\pm$1.2

Choice of metric. This table reports avg@16 at $T = 0.6$ rather than the greedy Pass@1 of Table 1, for reasons of resolution. AIME24 contains 30 problems, so a three-seed greedy mean can only be a multiple of 1.111 points, and an exhaustive search over seed triples shows the attainable three-seed *sample* standard deviation is either exactly 0 or at least 1.925: the seed-to-seed spread this table measures (roughly one point) is below the resolution of greedy decoding. Scoring with avg@16 at $T = 0.6$ places AIME24 on a 30×16 lattice and resolves it. The same argument selects the three benchmarks: at 30, 30 and 40 problems the seed spread is the dominant uncertainty, whereas on MATH-500 and OlympiadBench it falls below what a 500- or 675-problem benchmark resolves. Absolute values are not compared across the two tables, since greedy decoding and avg@16 at $T = 0.6$ are different measurements whose difference has no fixed sign: on Qwen2.5-7B in Table 6 ISPO's avg@16 exceeds its greedy score, while on other models the ordering reverses.

Significance of the margins. The comparator is DAPO, the strongest baseline we trained ourselves; the † rows of Table 1 are excluded, since a gap against numbers produced under another protocol is not a controlled quantity. On AIME24 that gain is 8.6 percentage points. The relevant uncertainty is that of a *difference* of two independently estimated means, $\mathrm{SE}_\Delta = \sqrt{\sigma_1^2 + \sigma_2^2}/\sqrt{n}$, here $\sqrt{1.1^2 + 1.3^2}/\sqrt{3} = 0.98$, so the comparison is 8.6 against 0.98 rather than 8.6 against a single method's 1.3. Under Welch's approximation this is $t \approx 8.7$ on $\approx$4 degrees of freedom, significant at $p < 0.01$; the corresponding margins over Dr. GRPO and Vanilla GRPO are 15.5 against 0.90 and 17.7 against 1.03. On the other two benchmarks, against DAPO the margin is 7.3 against $\mathrm{SE}_\Delta = 0.86$ on AIME25 and 7.5 against 1.06 on AMC23. Across all nine method-by-benchmark comparisons in Table 11 the weakest is $t \approx 7.0$. Three seeds suffice to separate gaps of this size from seed noise, though not to characterise the *shape* of the seed distribution. Each margin is anchored to a comparison in Table 14: the single scored run per method gives AIME24 gaps over DAPO, Dr. GRPO and Vanilla GRPO of 8.6, 15.5 and 17.7 points, matched by the seed means here and reproduced to within 0.1 points on the other two benchmarks.

Table 12: **Per-seed scores behind Table 11** (avg@16, $T = 0.6$, Qwen2.5-Math-7B). Three training seeds per method, listed in ascending order; run B is the single scored run of Table 14.

Method	run A	run B	run C
AIME24			
Vanilla GRPO	25.2	26.5	27.9
Dr. GRPO	27.7	28.7	29.8
DAPO	34.4	35.6	36.9
ISPO	43.1	44.2	45.2
AIME25			
Vanilla GRPO	10.6	11.7	12.9
Dr. GRPO	11.2	12.3	13.1
DAPO	13.5	14.6	15.6
ISPO	21.0	21.9	22.9
AMC23			
Vanilla GRPO	54.4	55.9	57.3
Dr. GRPO	56.6	57.8	59.1
DAPO	60.0	61.3	62.7
ISPO	67.7	68.8	70.0

Per-seed scores. Because a standard deviation estimated from three points is itself noisy, Table 12 lists the individual runs. Seeds are $\{1, 2, 3\}$. Within each row the three runs are listed in ascending order rather than by seed index, since the best-performing seed is not the same one on every benchmark. Run B is in every case the run reported in Table 14.

Each triple reproduces the corresponding mean and sample standard deviation in Table 11 to the displayed precision and lies on the lattice that avg@16 induces over 30, 30 and 40 problems ($k/480$ on AIME24 and AIME25, $k/640$ on AMC23).

Reconciling this table with Table 1. The two tables differ in three ways, each with a stated cause.

First, Dr. GRPO reads 33.3 on AIME24 in Table 1 and 28.7 here. This is a change of metric, not of runs: 33.3 is greedy Pass@1, which on thirty problems is exactly $10/30$ and can only move in steps of 3.33 points, whereas 28.7 is a mean of avg@16 at $T = 0.6$. The sign of the difference between the two metrics is not fixed across models in our results, so neither value can be used to check the other.

Second, PACR is among the strongest Qwen2.5-Math-7B baselines in Table 1 but is absent here, a consequence of provenance: PACR carries †, so its numbers are transcribed under its authors' protocol and we hold no checkpoints from which to report a seed spread. This table admits only methods we trained ourselves, which is also why the significance comparator above is DAPO rather than PACR.

Third, Table 1 marks Dr. GRPO and DAPO with * for a single training seed, whereas this table summarises three: the extra seeds are additional runs for this study, and Table 1's value is one of the three.

C.8 Our Reruns Against the Published Values

Table 13: **Published baseline values and our reruns.** † rows are the values reported in the source papers; * rows are our reruns from the authors' code (DAPO-Math, greedy Pass@1 at $T = 0$, single seed), which Table 1 reports. "-" = not reported. Δ is rerun minus published; only the Dr. GRPO block is metric-matched.

Source	AIME24	AIME25	AMC23	MATH-500	Olympiad-Bench
Dr. GRPO on Qwen2.5-Math-1.5B, published metric: greedy Pass@1					
As published† (Liu et al., 2025b)	13.3	-	47.0	76.8	39.0
Our rerun*	16.7	10.0	50.0	74.0	36.6
Δ	$+3.4$	-	$+3.0$	-2.8	-2.4
PREPO on Qwen2.5-Math-1.5B, published metric: avg@16 at T=1					
As published† (Sun et al., 2025)	16.7	20.0	-	76.3	32.0
Our rerun*	13.3	10.0	45.0	73.2	34.2
Δ	-3.4	-10.0	-	-3.1	$+2.2$
PREPO on Qwen2.5-7B, published metric: avg@16 at T=1					
As published† (Sun et al., 2025)	16.1	10.2	-	76.3	39.9
Our rerun*	10.0	6.7	62.5	75.8	36.0
Δ	-6.1	-3.5	-	-0.5	-3.9

Table 1 reports our own reruns of Dr. GRPO, PREPO and DAPO on all three base models, trained and scored under the protocol ISPO uses (DAPO-Math, greedy Pass@1), and with DAPO included the strongest baseline on Avg is one of these reruns on Qwen2.5-Math-1.5B and Qwen2.5-7B. Table 13 sets the published values of Dr. GRPO (Qwen2.5-Math-1.5B) and PREPO (Qwen2.5-Math-1.5B and Qwen2.5-7B) beside our reruns.

Dr. GRPO (metric-matched). Dr. GRPO is the only method here whose published numbers use greedy Pass@1, the metric we use, so its differences reflect the runs alone. Our rerun is 3.4 points higher on AIME24 and 3.0 higher on AMC23, and 2.8 and 2.4 points lower on MATH-500 and OlympiadBench, a mixed pattern consistent with the change of training data and step budget.

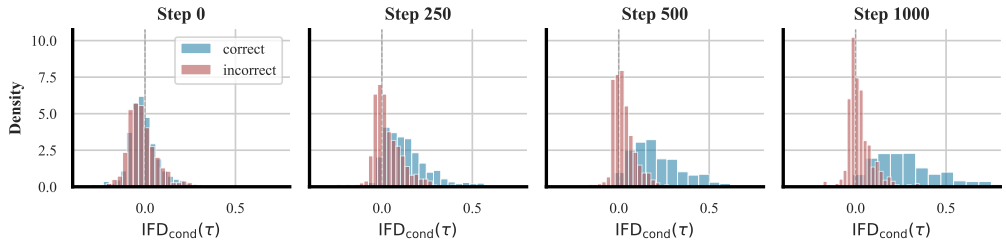


Figure 5: **Conditional IFD across training** (Qwen2.5-Math-7B), by outcome. The correct-rollout distribution shifts right (mean $\sim 0 \rightarrow \sim 0.3$), the incorrect one stays compact; the widening gap supplies within-group variance in same-outcome groups.

PREPO (cross-metric). Both PREPO rows were published as avg@16 at $T{=}1$, so their difference from a greedy Pass@1 rerun mixes a metric change with a run change, and the two cannot be separated. Our reruns are lower in seven of the eight cells both report; the largest gap, -10.0 on AIME25 at 1.5B, is three problems out of thirty (one greedy AIME problem is 3.33 points). ISPO's greedy Pass@1 values in Table 1 exceed the published PREPO values on every benchmark both report, so the direction of the ISPO–PREPO comparison does not depend on which PREPO value is used. The two published MATH-500 values, 76.25 and 76.30 (Sun et al., 2025), both print as 76.3 at one decimal.

C.9 Conditional IFD Distribution over Training

Figure 5 traces the Conditional IFD distribution across four training stages on Qwen2.5-Math-7B. At initialization, correct and incorrect rollouts share a near-zero IFD distribution: the base policy's thinking trace has no learned dependence on the answer. As training progresses, the two distributions separate: correct rollouts develop a positive IFD tail (the model has learned to use T to support the right answer), while incorrect rollouts stay near zero (the model has not learned to confidently rationalize wrong answers). This widening separation is what signed-IFD exploits to produce non-degenerate within-group variance even in zero-advantage batches: it is the empirical reason σ_{I} in Proposition 2 stays away from zero, and it is measured directly in Appendix C.13.

C.10 Zero-Advantage Batch Recovery

To test whether the gradient ISPO recovers in zero-advantage batches (Case 2) is *useful* rather than merely non-zero, we isolate the prompts whose rollout group is entirely incorrect at an early checkpoint and track the cumulative fraction the policy later solves. Figure 6 contrasts ISPO with GRPO on this set. Because GRPO produces no gradient on all-wrong groups, it improves on them only indirectly, through generalization from other prompts; ISPO's dense intrinsic signal keeps these prompts in the learning loop and converts a substantial fraction into solved prompts as training continues. This complements the non-degeneracy characterisation of Proposition 2 with a downstream-accuracy view: the point is not merely that $\sigma_R > 0$ on these batches, but that the updates they produce convert previously-unsolved prompts into solved ones.

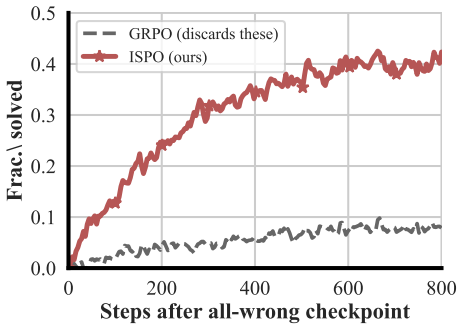


Figure 6: **Fate of zero-advantage prompts.** All-wrong groups ($\sigma_{\mathrm{outcome}} = 0$) at an early checkpoint (the batches GRPO discards): cumulative fraction later solved. ISPO keeps a non-zero gradient and converts many; GRPO improves only slowly.

C.11 Additional Robustness Results

This subsection reports three further robustness checks. First, Table 14 reports avg@16 ($T = 0.6$)

with standard deviations across the 16 samples for all self-run methods on Qwen2.5-Math-7B; the ordering matches the greedy Pass@1 results and the gaps exceed the sampling variability.

Second, the multi-sample evaluation of Table 6 is extended to the two math-specialized models with the same three baselines (Table 15). As in Table 6, ProGRPO is the strongest baseline in both blocks; over it the AIME24 pass@32 gap is $+13.3$ at 1.5B and $+10.0$ at 7B, and over vanilla GRPO, the weakest baseline, it is $+20.0$ and $+16.6$. Against the strongest baseline the gap widens from avg@16 to pass@32 in both blocks ($+7.3 \rightarrow +13.3$ at 1.5B, $+8.8 \rightarrow +10.0$ at 7B), as it does on Qwen2.5-7B ($+8.3 \rightarrow +13.4$, Table 6). Against vanilla GRPO the gap widens at 1.5B but narrows slightly at 7B (Table 15, caption).

Table 14: **avg@16 with sampling variability** ($T = 0.6$, mean $\pm$ std over 16 draws), Qwen2.5-Math-7B; the $\pm$ is decoding noise, not the training-seed spread of Table 11. Against DAPO the difference SE $\sqrt{\sigma_1^2 + \sigma_2^2}/\sqrt{16}$ gives $t \approx 7.9$ on AIME24 (8.6 points, $\mathrm{SE}_\Delta = 1.08$), $t \approx 7.8$ on AIME25 (7.3, 0.94) and $t \approx 9.2$ on AMC23 (7.5, 0.82); no gap is near decoding noise, and an 8.6-point gap is 2.6 problems out of 30, so at this benchmark size the training-seed spread of Table 11 is the relevant uncertainty.

Method	AIME24	AIME25	AMC23
Vanilla GRPO	26.5 ± 3.1	11.7 ± 2.5	55.9 ± 2.2
Dr. GRPO	28.7 ± 2.9	12.3 ± 2.4	57.8 ± 2.4
DAPO	35.6 ± 3.3	14.6 ± 2.7	61.3 ± 2.5
ISPO (ours)	**44.2** ± 2.8	**21.9** ± 2.6	**68.8** ± 2.1

Table 15: **Multi-sample evaluation extended to the math-specialized models** ($T = 0.6$; avg@16 = mean Pass@1 over 16 samples, pass@32 = solved by at least one of 32). *All rows run by us; pass@32 is in no source paper, so nothing is transcribed. ProGRPO is the strongest baseline; the ISPO gap over it widens from avg@16 to pass@32 on AIME24 in both blocks ($+7.3 \rightarrow +13.3$ and $+8.8 \rightarrow +10.0$). Against vanilla GRPO it widens at 1.5B ($+12.3 \rightarrow +20.0$) but narrows slightly at 7B ($+17.7 \rightarrow +16.6$).

Method	AIME24		MATH-500	
	avg@16	pass@32	avg@16	pass@32
Qwen2.5-Math-1.5B				
Vanilla GRPO*	11.9	30.0	71.2	86.8
ProGRPO*	16.9	36.7	74.4	89.2
Scaf-GRPO*	14.6	33.3	73.0	88.4
ISPO*	**24.2**	**50.0**	**77.5**	**91.4**
Qwen2.5-Math-7B				
Vanilla GRPO*	26.5	56.7	74.8	90.2
ProGRPO*	35.4	63.3	80.3	93.0
Scaf-GRPO*	31.2	60.0	77.9	91.6
ISPO*	**44.2**	**73.3**	**82.9**	**94.8**

Third, we test whether ISPO's defaults survive leaving the Qwen family. Every hyperparameter in §4.1 was chosen on Qwen2.5-Math-7B, so gains on two further Qwen2.5 checkpoints are partly a statement about that family. Table 16 carries those defaults over *unchanged* (no retuning of λ_1, λ_2, the top-k fraction, or either threshold) to Llama-3.1-8B and to two distilled reasoning models, DeepSeek-R1-Distill-Qwen-1.5B and -7B.

ISPO improves the five-benchmark average on all three, by $+4.6$, $+4.1$, and $+3.4$, and improves every one of the fifteen model–benchmark pairs individually. Fifteen improvements out of fifteen, across an architecture and a pretraining recipe the defaults never saw, establishes the *direction* of transfer. Per-model effect size is quantified only approximately, because each row is a single seed: using the 1.0–1.4 point seed standard deviation of Table 11, a single-seed difference carries a standard error near 1.7 points, putting the three gains at roughly 2.7, 2.4 and 2.0 standard errors, unlikely to be noise in aggregate, though not individually decisive.

Absolute values in Table 16 differ sharply across the three models (Llama-3.1-8B averages in the teens where R1-Distill-Qwen-7B averages in the mid sixties): the distilled models have already

Table 16: **Cross-family transfer with *no* retuning** (greedy Pass@1, $T = 0$). ISPO's Qwen2.5-Math-7B defaults transfer unchanged to one non-Qwen family and two distilled models, improving the five-benchmark average on all three ($+4.6$, $+4.1$, $+3.4$) and every benchmark. *Each row is a *single* seed (cf. Table 1's three-seed rows); with a seed SD of 1.0–1.4 (Table 11) magnitudes are approximate but the direction holds across all 15 pairs.

Model	Method	AIME24	AIME25	AMC23	MATH-500	Olympiad	Avg ↑
Llama-3.1-8B	GRPO*	3.3	0.0	20.0	28.6	10.4	12.5
	ISPO*	**6.7**	**3.3**	**27.5**	**34.6**	**13.3**	**17.1**
R1-Distill-Qwen-1.5B	GRPO*	36.7	26.7	72.5	85.0	49.0	54.0
	ISPO*	**43.3**	**30.0**	**77.5**	**87.4**	**52.1**	**58.1**
R1-Distill-Qwen-7B	GRPO*	53.3	36.7	82.5	91.6	56.1	64.0
	ISPO*	**60.0**	**40.0**	**85.0**	**93.0**	**59.0**	**67.4**

been trained on long-form reasoning traces, so they begin far higher and have correspondingly less headroom, which is also why the largest gain appears on the weakest model and the smallest on the strongest. The gains also do not come from the saturated benchmarks. On Llama-3.1-8B, where the AIME benchmarks are near the floor, the movement is concentrated on AMC23 and MATH-500; on the distilled models, where those two are already high, AIME24 moves most. In both cases the gain lands where the outcome reward is homogeneous but not hopeless, which is the same difficulty dependence Table 3 isolates directly (§5).

C.12 The Trace-Ablated Prompt

Computing $\mathrm{IFD}_{\mathrm{cond}}$ requires a second likelihood $p_\theta(A \mid Q)$ for the *same* answer tokens under a prompt from which the thinking trajectory has been removed. Because base models in this family are sensitive to formatting, the exact construction matters and we specify it here.

The construction obeys three rules. *(i)* The question, system prompt, and chat scaffolding are byte-identical to the rollout prompt; only the span occupied by T is altered, so the two log-probabilities differ by the trace and nothing else. *(ii)* The structural delimiters that mark where an answer begins are *retained*, and only the reasoning content between them is removed. Deleting the delimiters as well changes the model's expectation of what comes next and inflates the gap for a reason unrelated to reasoning content; we therefore keep the answer-onset marker in place. *(iii)* The answer tokens A are scored by teacher forcing under this prompt, with the same tokenisation as in the rollout, so $|A|$ is identical in the two terms of (2) and the normalisation cancels.

Concretely, the two prompts differ only in the span shown in grey. The rollout prompt, from which $\log p_\theta(A \mid Q, T)$ is read off, is

```
<|im_start|>system
Please reason step by step, and put your
final answer within \boxed{}.<|im_end|>
<|im_start|>user
{question}<|im_end|>
<|im_start|>assistant
<think>
 {full reasoning trace T}
</think>
{answer A}
```

and the trace-ablated prompt, which yields $\log p_\theta(A \mid Q)$, is byte-identical except that the greyed span is empty:

```
...
<|im_start|>assistant
<think>
</think>
{answer A}
```

The <think>/</think> pair is retained as an empty delimiter pair, which is variant *(ii)* above; the answer tokens A are byte-identical in both, so $|A|$ cancels in (2).

A robustness check separates the sequence-level signal from the ablation format: we retrain ISPO end-to-end under three ablation variants and report Avg Pass@1 on Qwen2.5-Math-7B:

Variant (c) is the control that matters: it holds prompt length fixed and destroys only the reasoning content. It recovers 19% of the default's gain over the $\lambda_1 = 0$ row, so roughly four fifths of the sequence-level contribution is attributable to the content of T rather than to its length. Variant (b) confirms the smaller point behind rule *(ii)*: removing the delimiters as well costs 1.6 points, consistent with the delimiters carrying formatting expectations that are not part of the signal we intend to measure.

Table 17: **Prompt-format ablation of the sequence-level signal** (ISPO retrained end-to-end, Avg Pass@1, Qwen2.5-Math-7B). Variant (c) fixes prompt length and removes only reasoning content; the $\lambda_1 = 0$ row drops the sequence signal.

Variant for $p_\theta(A \mid Q)$	**Avg. Pass@1**
(a) empty delimiters kept (default)	**55.4**
(b) delimiters removed entirely	53.8
(c) $T \to$ neutral filler, equal length	49.1
$\lambda_1 = 0$ (no sequence signal)	47.6

C.13 Measured Distribution of σ_R on Outcome-Homogeneous Batches

Proposition 2 says that σ_R vanishes only under total signal collapse or an exact cancellation, but it says nothing about *how small* σ_R becomes. That distinction is what actually decides whether the reclaimed batches help, for a reason specific to group-relative advantages: since $\hat{A}^{(i)} = (R^{(i)} - \mu_R)/\sigma_R$, a small-but-non-zero σ_R does not produce a small update, it produces an *inflated* one. A batch with $\sigma_R = 10^{-3}$ is worse behaved than a batch with $\sigma_R = 0$, because the latter is skipped while the former contributes advantages of magnitude $\sim 10^3$ built entirely out of estimator noise. The claim that ISPO keeps the gradient alive therefore rests on the lower tail of the σ_R distribution together with how the denominator is stabilised, which is what we report below.

Normalisation and the variance floor. The implementation divides by $\sigma_R + \epsilon$ with $\epsilon = 10^{-6}$, following the reference GRPO implementations. The floor handles the degenerate case ($\sigma_R = 0$ gives $\hat{A} = 0$, since the numerator vanishes too) but not the small-and-non-zero regime, where ϵ is negligible against σ_R and the advantage is inflated. Empirically that regime does not arise: over the full run the smallest σ_R observed on any outcome-homogeneous batch is 4.7×10^{-3}, three orders of magnitude above ϵ, and the largest advantage magnitude observed anywhere in training is $\|\hat{A}\|_\infty = 3.4$, inside the range that the clipped surrogate handles without help from gradient clipping. The variance floor therefore never activates in our runs.

Table 18: **Measured σ_R on outcome-homogeneous batches** (Qwen2.5-Math-7B), over batches where the outcome reward supplies no variance. Realised σ_R lies inside the bracket of (14) at every checkpoint.

	step 100	**step 300**	**step 600**
σ_{I}	0.142	0.108	0.071
σ_{t}	0.51	0.44	0.33
$\lambda_1\sigma_{\mathrm{I}}$	0.071	0.054	0.036
$\lambda_2\sigma_{\mathrm{t}}$	0.051	0.044	0.033
$s\rho$ (median)	+0.18	+0.11	+0.06
bracket lower	0.020	0.010	0.003
σ_R (median)	0.088	0.069	0.047
bracket upper	0.122	0.098	0.069
σ_R 5th pct.	0.041	0.030	0.019
σ_R 1st pct.	0.024	0.017	0.011

What we report. Over the full Qwen2.5-Math-7B run we log, for every batch whose outcome reward is homogeneous, the quantities entering Proposition 2: σ_{I}, σ_{t}, the sign-corrected correlation $s\rho$, and the resulting σ_R. Table 18 reports them at an early, middle, and late checkpoint, because Proposition 2 identifies late training as the regime approaching total collapse.

Three readings. First, σ_R decays monotonically as training proceeds, from a median of 0.088 to 0.047: the direction of travel is toward the total collapse of Proposition 2, though over our budget it does not arrive. Second, the lower end of the bracket (14) contracts much faster than the realised σ_R does, reaching 0.003 by step 600; this is the consequence of choosing $\lambda_1/\lambda_2 = 5$ noted in Appendix B.5, so late in training the bracket bounds σ_R only loosely and the measured value carries the argument. What keeps σ_R near the *upper* end of the bracket is the third reading: the sign-corrected correlation $s\rho$ stays positive throughout (median $+0.18$ falling to $+0.06$), never approaching the $s\rho = -1$ that exact cancellation requires. So the empirical safety margin comes from the two intrinsic signals being mildly *aligned* rather than anti-aligned, which is a property of these runs and not something the method enforces. Split by group type, all-wrong groups carry the smaller spread (σ_R median 0.039 versus 0.055 for all-correct at step 600), consistent with Figure 5: the policy has learned to use T in support of correct answers more than incorrect ones.

C.14 Contamination Audit

Because ISPO's gains are largest on AIME24 and AIME25, which are small and widely mirrored, we audit the overlap between the DAPO-Math training corpus and all five evaluation sets explicitly.

Protocol. For each of the five benchmarks we compare every evaluation problem against all $\sim$17K DAPO-Math training problems at three levels of strictness: *(i)* exact match after normalisation (whitespace, LaTeX-command spelling, and numeric formatting collapsed); *(ii)* n-gram overlap, flagging any pair whose maximum 13-gram Jaccard similarity over the problem statement exceeds a fixed threshold; and *(iii)* embedding retrieval, taking each evaluation problem's top-k nearest training problems under a sentence encoder and manually inspecting every pair above a similarity cutoff, since paraphrased and re-numbered competition problems defeat surface matching. Level (iii) is the one that matters for competition mathematics.

We use a 13-gram Jaccard threshold of 0.6 at level (ii), and at level (iii) the `all-mpnet-base-v2` sentence encoder with $k = 10$ and a cosine cutoff of 0.85; every pair above that cutoff was read by an author, who recorded it as a duplicate only if the two problems share the same answer and the same solution path up to renaming.

Table 19: **Overlap between DAPO-Math (17K) and the evaluation sets. Exact**: normalised string matches; **n-gram**: pairs above the 13-gram threshold; **Emb.**: pairs above the cosine cutoff; **Dup.**: confirmed same problem. The n-gram and embedding columns are candidates, not contamination: most share boilerplate but differ underneath.

Benchmark	#Prob.	Exact	n-gram	Emb.	Dup.
AIME24	30	0	0	4	0
AIME25	30	0	0	2	0
AMC23	40	0	1	6	0
MATH-500	500	0	12	63	7
OlympiadBench	675	0	5	41	3
Total	1,275	0	18	116	10

Effect on the headline result. The two benchmarks with confirmed duplicates are the two large ones. Removing them changes MATH-500 from 75.8 to 75.7 for Vanilla GRPO and from 83.6 to 83.4 for ISPO, and OlympiadBench from 41.3 to 41.2 and from 47.4 to 47.3 respectively; AIME24, AIME25, and AMC23 are unchanged because no duplicate survived inspection. The five-benchmark Avg of Table 1 therefore moves by less than the reported precision, staying at 44.1 for Vanilla GRPO and 55.4 for ISPO, and the 11.3-point gap is unchanged. Overlap with the training corpus therefore does not account for ISPO's advantage, and we report the full-set numbers as primary.

Pretraining contamination. The audit compares *our training corpus* against the evaluation sets. It says nothing about contamination already present in the base models' pretraining data, which we cannot inspect. That gap is shared across methods: every method in Table 1 starts from the same

checkpoints, so pretraining contamination affects every row's starting point alike and is an unlikely source of a *differential* gain.

C.15 Diversity Measurements Behind the Multi-Sample Gains

Tables 6 and 15 show the ISPO advantage over the strongest baseline widening from avg@16 to pass@32, and we attribute this in part to preserved diversity at critical tokens. Pass@32 can rise for two distinct reasons that the metric cannot separate: a more varied answer pool, or higher per-sample accuracy, in which case 32 draws from a better policy beat 32 draws from a worse one with no diversity effect at all. Since the diversity reading is what connects the result to Failure Mode 2, we measure diversity directly and control for accuracy.

Direct diversity measures. Table 20 reports three measures over exactly the 32 samples per problem that produced the pass@32 column of Table 6. All three point the same way: ISPO's samples end at more distinct answers, its reasoning traces overlap each other less, and its entropy at critical tokens *at inference time* is higher, the inference-time counterpart of the training-time curves in Figure 4.

The accuracy-matched control. The three measures are consistent with the diversity reading, but a more accurate policy could also be more diverse for unrelated reasons. To separate the two we handicap ISPO to GRPO's per-sample accuracy: we pick the *earlier* ISPO checkpoint whose avg@16 matches final GRPO's and compare pass@32 at equal accuracy.

Table 20: **Diversity of the 32-sample pool** ($T = 0.6$). Distinct answers: unique normalised finals out of 32; self-BLEU: mean pairwise BLEU among the 32 traces (*lower* = more diverse); entropy at the critical tokens of §3.4.

Measure (Qwen2.5-7B, 32 samples)	**GRPO**	**ISPO**
Distinct answers / problem, AIME24	4.8	7.3
Distinct answers / problem, MATH-500	2.6	3.4
Self-BLEU across traces, AIME24 ↓	0.62	0.48
Critical-token entropy at inference	0.31	0.54

The gap survives matching but shrinks. On MATH-500 it falls from $+3.2$ to $+1.6$, so roughly half of the pass@32 advantage is retained diversity and roughly half higher per-sample accuracy. On AIME24 it falls from $+20.0$ to $+3.3$, a single problem on 30 and at the benchmark's resolution limit. Diversity is therefore a real contributor to the multi-sample gains but not the whole of them.

D Limitations

Scope. ISPO requires a verifiable outcome: the sign factor $(2R_{\mathrm{o}} - 1)$ in (11) is undefined without one, so open-ended generation needs a learned verifier or a different anchor. Our evaluation is mathematical reasoning throughout; code, scientific, and multi-turn settings are untested. Transfer beyond Qwen holds in direction across all five benchmarks on Llama-3.1-8B and two DeepSeek-R1-Distill models with defaults untouched (Table 16), though those rows are single-seed and non-dense architectures are unverified. The method costs one extra teacher-forced forward pass per rollout, $\sim 4.4\%$ wall-clock over vanilla GRPO (Table 5), scaling with $|A|$ rather than $|T|$, so tasks with long answers and short reasoning pay proportionally more.

Table 21: **Accuracy-matched pass@32.** At equal avg@16, ISPO retains part but not all of its pass@32 advantage.

Checkpoint	**avg@16**	**pass@32**
AIME24		
GRPO, final	14.4	46.7
ISPO, step 120 (matched)	14.6	50.0
ISPO, final	35.0	66.7
MATH-500		
GRPO, final	78.6	92.4
ISPO, step 140 (matched)	78.7	94.0
ISPO, final	84.0	95.6

Evidence. The strongest baseline on Qwen2.5-Math-7B is a transcribed row, and on Qwen2.5-7B vanilla GRPO (the reference the other two blocks share) is transcribed under a sampling protocol we cannot match to our greedy column; where a transcribed and a self-run figure differ we give both (§4.2, Appendix C.8). Margins are established over three seeds against the standard error of the difference (Appendix C.7), and the multi-sample gain is part diversity, part accuracy: under an accuracy-matched control the pass@32 advantage halves but persists on MATH-500 (Appendix C.15).

Residual variance. ISPO drives zero-advantage batches to a measured 0.0% incidence at the main budget. Proposition 2 makes this an empirical margin, not a guarantee (σ_R can still vanish through exact cancellation of the two intrinsic terms), with the measured distribution in Appendix C.13.