IG-OPD: INFORMATION-GEOMETRIC CONTROL FOR ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) trains student language models using teacher feedback on self-generated trajectories, but such feedback can become unreliable when student-generated contexts deviate from teacher-preferred trajectories. In particular, the teacher distribution conditioned on a deviated student prefix may induce a correction direction that conflicts with the target associated with a selected task-valid reference prefix. We study this phenomenon from an information-geometric perspective by modeling next-token distributions on the probability simplex equipped with the Fisher metric. Using the square-root embedding of the categorical Fisher simplex, we derive an exact finite-step identity and a necessary-and-sufficient condition under which an idealized Fisher–Rao correction toward the teacher target conditioned on a deviated prefix moves away from a selected reference target. We further exploit the relationship between Fisher–Rao distance and the Bhattacharyya coefficient to derive a tractable teacher–student discrepancy signal. Guided by these geometry-motivated heuristics, we propose **IG-OPD**, which combines geometry-aware token weighting, discrepancy-guided prompt routing between off-policy supervised fine-tuning and OPD, and token-wise adaptive interpolation between forward and reverse KL objectives. Experiments across textual and multimodal tasks, teacher configurations, and student scales consistently improve over strong OPD baselines. Controlled-prefix diagnostics further demonstrate that larger teacher–student discrepancy is associated with more frequent reference-direction conflict.

1 INTRODUCTION

The rapid growth of large language models (LLMs) has increased the need to transfer their capabilities into smaller, deployable student models (Zhao et al., 2026b; Minaee et al., 2024; Brown et al., 2020). Knowledge distillation (KD) (Hinton et al., 2015; Xu et al., 2024), which trains a student to match the output behavior of a teacher, has become an important approach for LLM compression and post-training (Kim & Rush, 2016; Sanh et al., 2019). A key limitation of conventional off-policy distillation is *exposure bias* (Bengio et al., 2015; Ranzato et al., 2015; Zhang et al., 2019): training supervision is provided on fixed or teacher-generated contexts, whereas the student must condition on its own previous generations at inference time, causing errors to compound autoregressively. On-policy distillation (OPD) (Agarwal et al., 2024; Gu et al., 2024) mitigates this mismatch by allowing the student to generate its own rollouts and querying the teacher on student-visited contexts. At each context, the teacher's next-token distribution typically provides token-level supervision through a divergence-based objective, with existing methods differing in their divergence choices and treatment of individual tokens (Gu et al., 2024; Wen et al., 2023; Jin et al., 2026).

Despite this progress, OPD supervision can vary substantially in reliability across student-generated contexts. First, many existing methods apply uniform weights across token positions (Agarwal et al., 2024; Ko et al., 2024; Zheng et al., 2026). When a student trajectory deviates from a selected task-valid reference trajectory, subsequent teacher feedback is conditioned on the resulting prefix and may induce a correction direction that conflicts with the teacher target associated with the task-valid reference state. We refer to this phenomenon as *reference-direction conflict*. Entropy-based and token-selection methods (Jin et al., 2026; Huang et al., 2026; Xu et al., 2026c) partially address heterogeneous or uncertain supervision, but do not explicitly characterize such directional mismatch. Second, OPD can become less reliable under large teacher–student gaps, while off-policy warm-up

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.

may improve training stability (Li et al., 2026). Existing curriculum and difficulty-aware methods (Kumar et al., 2010; Zhang & Liu, 2025; Zhou et al., 2026; Xu et al., 2026b) motivate adaptive scheduling, but typically do not select the training strategy separately for each prompt using a distributional discrepancy signal. Third, a fixed forward KL (FKL) or reverse KL (RKL) objective imposes the same mass-covering or mode-seeking preference across positions (Gu et al., 2024; Wen et al., 2023; Luong et al., 2026). Although combining the two can be beneficial (Wu et al., 2025), how their relative contributions should vary across student-visited contexts remains underexplored.

To address these challenges, we view each next-token distribution as a point on the probability simplex endowed with the Fisher metric (Amari, 1985; 2016; Amari & Nagaoka, 2000; Nielsen, 2020). To isolate prefix-induced target conflict from parameterization and optimizer effects, we consider an idealized Fisher–Rao geodesic correction toward the teacher target conditioned on a deviated prefix. Using the spherical representation of the categorical Fisher simplex, we derive an exact finite-step displacement identity and a necessary-and-sufficient condition for this correction to increase the distance from a selected reference target. Because the directional conflict criterion requires an unavailable counterfactual reference target, we express the observable Fisher–Rao correction magnitude exactly through the Bhattacharyya coefficient, yielding a tractable teacher–student discrepancy signal. Finally, the simplex's dual affine structure distinguishes the conditional projection geometries induced by FKL and RKL, motivating adaptive rather than fixed KL orientation.

Motivated by these analyses, we propose **IG-OPD**, which adaptively controls three complementary aspects of distillation. First, *geometry-aware token weighting* attenuates geometrically large corrections to reduce the impact of potentially disruptive updates. Second, *discrepancy-guided prompt routing* assigns high-discrepancy prompts to off-policy supervised fine-tuning (SFT) and lower-discrepancy prompts to OPD, avoiding continued reliance on teacher feedback conditioned on substantially deviated student prefixes. Third, *adaptive KL orientation* interpolates between RKL and FKL according to the local teacher–student discrepancy. Experiments across textual and multimodal tasks, different student scales, teacher configurations, and strong-to-weak settings consistently improve over strong OPD baselines. Controlled-prefix diagnostics further show an empirical association between teacher–student discrepancy and reference-direction conflict.

In summary, our main contributions are as follows.

- An information-geometric analysis of OPD that provides an exact finite-step characterization of deviated-prefix reference conflict, derives a tractable Bhattacharyya representation of the observable Fisher–Rao correction magnitude, and distinguishes the conditional projection geometries induced by FKL and RKL.
- A unified framework, **IG-OPD**, that uses a tractable Bhattacharyya-based discrepancy signal to adapt token-level supervision strength, prompt-level SFT–OPD routing, and token-level KL orientation for more reliable OPD.
- Extensive evaluation across textual and multimodal benchmarks, different student scales, single- and multi-teacher settings, and strong-to-weak distillation, together with controlled-prefix diagnostics and ablation studies of the proposed geometric signals.

2 PRELIMINARY

On-Policy Distillation Given a prompt x, the teacher policy π_T and student policy π_θ define conditional next-token distributions at context $c_t = (x, y_{<t})$. A common RKL-based OPD formulation considers the following objective over student-generated rollouts:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\,\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[\frac{1}{|y|}\sum\nolimits_{t=1}^{|y|} D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid c_t)\|\pi_T(\cdot \mid c_t)\right)\right]. \tag{1}$$

The rollout distribution itself depends on the student, whereas the token-level KL terms are evaluated at the sampled contexts. In practice, many OPD implementations stop gradients through discrete rollout sampling and differentiate only the per-token divergences at the realized contexts (Yang et al., 2026; Ko et al., 2026).

Information Geometry At a fixed context, the teacher and student next-token distributions are points in the interior of the categorical probability simplex. Let $K = |\mathcal{V}|$; then

$$\mathcal{S}_K = \left\{p \in \mathbb{R}^K : p_i > 0,\ \sum\nolimits_{i=1}^{K} p_i = 1\right\}. \tag{2}$$

Equipped with the Fisher metric, $\mathcal{S}_K$ forms a $(K-1)$-dimensional Riemannian manifold $\mathcal{M}$ (Amari, 2016; Amari & Nagaoka, 2000). At $p \in \mathcal{M}$, its tangent space is

$$T_p\mathcal{M} = \left\{ u \in \mathbb{R}^K : \sum_{i=1}^K u_i = 0 \right\}, \tag{3}$$

whose elements represent infinitesimal probability perturbations that preserve normalization. For $u, v \in T_p\mathcal{M}$, the Fisher metric is $\langle u, v \rangle_p^F := \sum_{i=1}^K u_i v_i / p_i$. This metric induces the Fisher–Rao distance d_F, as well as the Riemannian exponential and logarithm maps, denoted by Exp_p and Log_p, which we use below to compare teacher–student correction directions. The categorical simplex is dually flat: probabilities provide mixture coordinates, while log-probability ratios $\xi_i(p) = \log(p_i/p_K)$, $i = 1, \ldots, K-1$, provide exponential coordinates. Accordingly, we use the m/e-affine structures for the conditional KL-projection identities in Section 3.3, and the Levi–Civita geometry of the Fisher metric for Fisher–Rao distances, logarithm maps, and directional comparisons.

3 Information-Geometric Analysis of OPD

We analyze OPD directly on the categorical probability simplex equipped with the Fisher metric, without assuming that the neural-model family is globally affine or dually flat. Under a fixed rollout distribution, the infinitesimal-step natural-gradient dynamics of token-level OPD form a Riemannian gradient flow. This perspective motivates intrinsic Fisher–Rao quantities for diagnosing conflicts between perturbed-prefix teacher targets and task-valid reference targets[1].

3.1 Reference-Direction Conflict

To isolate target conflict from parameterization and optimizer effects, we consider an idealized Fisher–Rao correction relative to a task-valid reference trajectory. At position t, let c_t^{ref} denote the reference context and c_t^{dev} a paired context whose prefix differs from the reference prefix because of an earlier deviation. Define $p_t = \pi_\theta(\cdot \mid c_t^{\mathrm{dev}})$, $p_{T,t}^{\mathrm{ref}} = \pi_T(\cdot \mid c_t^{\mathrm{ref}})$, and $p_{T,t}^{\mathrm{dev}} = \pi_T(\cdot \mid c_t^{\mathrm{dev}})$, where $p_{T,t}^{\mathrm{ref}}$ is the teacher target associated with the task-valid reference state. The corresponding correction directions are $u_t^{\mathrm{ref}} = \mathrm{Log}_{p_t}(p_{T,t}^{\mathrm{ref}})$ and $u_t^{\mathrm{dev}} = \mathrm{Log}_{p_t}(p_{T,t}^{\mathrm{dev}})$, both of which lie in $T_{p_t}\mathcal{M}$. For nonzero directions, define their normalized Fisher alignment as $\rho_t := \cos \angle_{p_t}^F(u_t^{\mathrm{ref}}, u_t^{\mathrm{dev}}) \in [-1, 1]$. The square-root embedding of the categorical Fisher simplex permits an exact finite-step characterization. Define $a_t = d_F(p_t, p_{T,t}^{\mathrm{ref}}) = \|u_t^{\mathrm{ref}}\|_{p_t,F}$ and $b_t = d_F(p_t, p_{T,t}^{\mathrm{dev}}) = \|u_t^{\mathrm{dev}}\|_{p_t,F}$.

Proposition 3.1 (Finite-step reference-displacement identity)**.** *Assume that p_t, $p_{T,t}^{\mathrm{ref}}$, and $p_{T,t}^{\mathrm{dev}}$ lie in the interior of the categorical probability simplex and that $a_t, b_t > 0$. For $0 \le \eta \le 1$, consider the Fisher–Rao correction toward the deviated-prefix teacher target, $p_t^+ = \mathrm{Exp}_{p_t}(\eta u_t^{\mathrm{dev}})$. Then*

$$\cos\left(\frac{d_F(p_t^+, p_{T,t}^{\mathrm{ref}})}{2}\right) = \cos\left(\frac{a_t}{2}\right)\cos\left(\frac{\eta b_t}{2}\right) + \sin\left(\frac{a_t}{2}\right)\sin\left(\frac{\eta b_t}{2}\right)\rho_t. \tag{4}$$

For any $0 < \eta \le 1$, the reference distance increases precisely when $\rho_t < \cot(a_t/2)\tan(\eta b_t/4)$. Since this threshold is positive, $\rho_t < 0$ guarantees an increase for every such η, and hence throughout the minimizing geodesic segment. The exact identity also yields the first-order expansion

$$\frac{1}{2}d_F^2(p_t^+, p_{T,t}^{\mathrm{ref}}) = \frac{1}{2}d_F^2(p_t, p_{T,t}^{\mathrm{ref}}) - \eta \left\langle u_t^{\mathrm{ref}}, u_t^{\mathrm{dev}} \right\rangle_{p_t}^F + \mathcal{O}(\eta^2). \tag{5}$$

Proposition 3.1 demonstrates that finite-step reference displacement is jointly governed by directional alignment and correction magnitude, with negative alignment guaranteeing an increase throughout the minimizing geodesic segment. This motivates geometry-aware control of OPD supervision, particularly more conservative treatment of geometrically large corrections rather than applying teacher feedback uniformly.

3.2 Tractable Geometric Discrepancy

The alignment score ρ_t depends on the counterfactual reference target $p_{T,t}^{\mathrm{ref}}$ and is unavailable in standard OPD training. Using the definition of ρ_t and $\|\mathrm{Log}_p(q)\|_{p,F} = d_F(p, q)$, Eq. (5) gives

$$\frac{1}{2}d_F^2(p_t^+, p_{T,t}^{\mathrm{ref}}) - \frac{1}{2}d_F^2(p_t, p_{T,t}^{\mathrm{ref}}) = -\eta\rho_t\, d_F(p_t, p_{T,t}^{\mathrm{ref}}) d_F(p_t, p_{T,t}^{\mathrm{dev}}) + O(\eta^2). \tag{6}$$

[1] Formal proofs of all theoretical results in this section are provided in Appendix B.

Thus, the first-order effect factorizes into directional alignment, reference magnitude, and correction magnitude. Among these geometric factors, only $d_F(p_t, p_{T,t}^{\mathrm{dev}})$ is directly observable during distillation. It measures the Fisher–Rao magnitude of the deviated-prefix teacher correction and therefore controls the scale of its potential first-order effect even when its reference alignment is unknown. We accordingly utilize this observable correction magnitude as our geometric control variable. For categorical distributions, this Fisher–Rao correction magnitude can be computed exactly from the Bhattacharyya coefficient.

Proposition 3.2 (Bhattacharyya representation of Fisher–Rao distance)**.** *For p, q in the interior of the categorical simplex,*

$$d_F(p,q) = 2\arccos\left(\sum_{i=1}^{K}\sqrt{p_i q_i}\right) = 2\arccos\big(\mathrm{BC}(p,q)\big), \tag{7}$$

where $\mathrm{BC}(p,q)=\sum_i \sqrt{p_i q_i}$. For nearby distributions, $d_F^2(p,q)=8\big[-\log \mathrm{BC}(p,q)\big]+\mathcal{O}\big(d_F^4(p,q)\big)$.

We therefore define the nonnegative discrepancy score $\widehat{d}_t = -\log \mathrm{BC}(\pi_\theta(\cdot \mid c_t), \pi_T(\cdot \mid c_t))$, which is monotone in the observable correction magnitude $\|u_t^{\mathrm{dev}}\|_F = d_F(p_t, p_{T,t}^{\mathrm{dev}})$ and locally satisfies $\widehat{d}_t = d_F^2/8 + O(d_F^4)$. Thus, $\widehat{d}_t$ is an exact observable monotone measure of the correction-magnitude factor appearing in the first-order reference-displacement term, which can be utilized to control the scale of corrections whose reference alignment is unobservable[2].

3.3 Dual KL Projection Geometry

KL divergence is the Bregman divergence generated by negative entropy, $\Phi(p) = \sum_i p_i \log p_i$, with $D_\Phi(p,q) = D_{\mathrm{KL}}(p\|q)$. On the dually flat probability simplex, its two orientations admit distinct Pythagorean projection identities under the corresponding m- and e-affine assumptions (Amari, 2016). Because neural-model distribution families need not satisfy these assumptions, we state the identities conditionally and use them to distinguish the projection geometries of RKL and FKL in Proposition 3.3.

Proposition 3.3 (Dual KL Projections)**.** *Fix a context c, and let $p_T = \pi_T(\cdot \mid c)$ denote the teacher's next-token distribution in the interior of the categorical simplex.*

(i) Let $\mathcal{C}_m \subset \mathcal{M}$ be a nonempty m-affine family. Suppose $p_R = \arg\min_{p\in\mathcal{C}_m} D_{\mathrm{KL}}(p\|p_T)$ exists, is unique, and lies in the relative interior of $\mathcal{C}_m$. Then, for every $p \in \mathcal{C}_m$,

$$D_{\mathrm{KL}}(p\|p_T) = D_{\mathrm{KL}}(p\|p_R) + D_{\mathrm{KL}}(p_R\|p_T). \tag{8}$$

(ii) Let $\mathcal{C}_e \subset \mathcal{M}$ be a nonempty e-affine family. Suppose $p_F = \arg\min_{p\in\mathcal{C}_e} D_{\mathrm{KL}}(p_T\|p)$ exists, is unique, and lies in the relative interior of $\mathcal{C}_e$. Then, for every $p \in \mathcal{C}_e$,

$$D_{\mathrm{KL}}(p_T\|p) = D_{\mathrm{KL}}(p_T\|p_F) + D_{\mathrm{KL}}(p_F\|p). \tag{9}$$

Proposition 3.3 indicates that RKL and FKL define distinct constrained projection geometries under their respective affine assumptions, so neither orientation captures both projection preferences. This motivates adaptively combining their complementary distribution-matching behavior rather than using a fixed KL orientation throughout distillation.

4 IG-OPD: Information-Geometric Control

Analysis in Section 3 provides two design principles for reliable OPD. Proposition 3.1 motivates geometry-aware control of OPD, where we consider control at both the token and sample levels, while Proposition 3.3 motivates adapting the RKL–FKL combination due to their distinct projection geometries. IG-OPD instantiates these principles through geometry-aware token weighting, discrepancy-guided prompt routing, and adaptive KL orientation. All three controllers use the tractable Bhattacharyya-based discrepancy from Section 3.2, avoiding explicit Riemannian computations during training.

[2]We have also empirically evaluated the discrepancy signal's association with reference-direction conflict in Section 5.4.

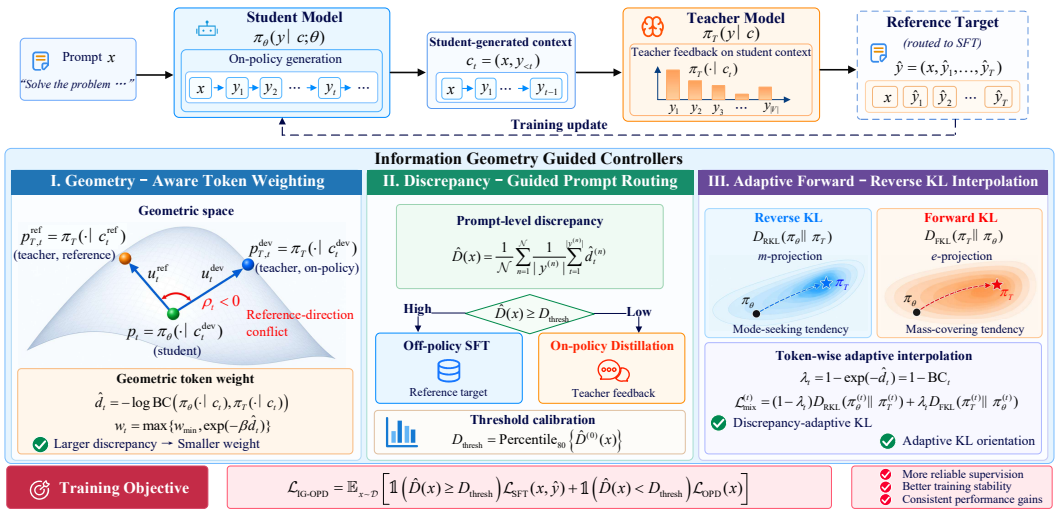


Figure 1: Overview of IG-OPD. IG-OPD routes each prompt either to off-policy SFT on teacher-generated trajectories or to OPD using teacher feedback at student-visited contexts. A unified Bhattacharyya/Fisher–Rao discrepancy signal drives geometry-aware token weighting, prompt-level SFT–OPD routing, and adaptive forward–reverse KL interpolation.

4.1 Geometry-Aware Token Weighting

Proposition 3.1 indicates that, under negative alignment, reference displacement increases with the Fisher–Rao correction length; thus, b_t controls the potential severity of a conflicting correction. Because $\widehat{d}_t$ is monotone in $b_t = \|u_t^{\text{dev}}\|_{p_t,F}$, we use it as a tractable magnitude signal and attenuate geometrically large token-level corrections via $w_t = \max\{w_{\min}, \exp(-\beta\widehat{d}_t)\}$. We set $\beta = 1$, for which the unclipped weight equals BC_t, and use $w_{\min} = 0.1$ to prevent potentially informative tokens from receiving vanishing weight.

We treat w_t as a stop-gradient scalar. In the idealized geodesic model, weighting reduces the traveled arc from ηb_t to $\eta w_t b_t$, thereby limiting finite reference displacement under negative alignment. In practical optimization, it reduces the influence of high-discrepancy token losses on the minibatch gradient, thereby controlling the magnitude of potentially disruptive corrections.

4.2 Discrepancy-Guided Prompt Routing

Empirical evidence suggests that successful OPD often relies on compatible teacher–student patterns[3] and that off-policy SFT training can rescue otherwise unsuccessful runs (Li et al., 2026; Liu et al., 2026). Our analysis (Proposition 3.1) demonstrates that potential reference displacement under directional conflict depends on the Fisher–Rao correction magnitude. Since the introduced discrepancy score $\widehat{d}_t$ is monotone in this magnitude, we route high-discrepancy prompts to SFT for better alignment and retain OPD for lower-discrepancy prompts to preserve its on-policy advantage.

With the token-level $\widehat{d}_t$, we operationalize this routing principle by aggregating it across token positions and student rollouts into the prompt-level discrepancy

$$\widehat{D}(x) = \frac{1}{\mathcal{N}} \sum_{n=1}^{\mathcal{N}} \frac{1}{|y^{(n)}|} \sum_{t=1}^{|y^{(n)}|} \widehat{d}_t^{(n)}. \tag{10}$$

Here $\widehat{D}(x)$ is the rollout-averaged, length-normalized prompt discrepancy. We route prompts according to

$$\text{Strategy}(x) = \begin{cases} \text{off-policy SFT}, & \widehat{D}(x) \geq D_{\text{thresh}}, \\ \text{on-policy OPD}, & \widehat{D}(x) < D_{\text{thresh}}. \end{cases} \tag{11}$$

[3]Experiments have been conducted to evaluate this phenomenon, which (Fig. 2(c)) shows that prompts with larger initial teacher–student discrepancy tend to achieve lower final accuracy under vanilla OPD.

Because the discrepancy scale varies across teacher–student configurations, we calibrate a threshold before training using a subset randomly sampled from the training data:

$$D_{\text{thresh}} = \text{Percentile}_{80}\left(\{\widehat{D}^{(0)}(x) : x \in \mathcal{D}_{\text{calib}}\}\right). \tag{12}$$

The threshold is calibrated with the initial student and then held fixed, while $\widehat{D}^{(k)}(x)$ is re-estimated from the current student, allowing prompts to switch between SFT and OPD during training.

4.3 Adaptive KL Orientation

Proposition 3.3 distinguishes the conditional projection geometries of RKL and FKL; hence, neither orientation alone captures both projection preferences. RKL is student-weighted and tends to concentrate on student-supported modes, whereas FKL is teacher-weighted and more strongly penalizes missing teacher-supported mass (Gu et al., 2024; Wen et al., 2023; Jung et al., 2025). Because the two orientations share the same local second-order Fisher geometry near agreement, we retain RKL in this regime and progressively increase the FKL contribution as the discrepancy grows. We control this interpolation using $\widehat{d}_t$, an observable monotone function of the Fisher–Rao distance. Since $\widehat{d}_t = -\log \text{BC}_t$, with $\text{BC}_t := \text{BC}(\pi_\theta^{(t)}, \pi_T^{(t)}) \in (0, 1]$, we define the parameter-free, bounded coefficient as follows:

$$\lambda_t = 1 - \exp(-\widehat{d}_t) = 1 - \text{BC}_t, \tag{13}$$

where $\lambda_t \in [0, 1)$. This controller is zero under exact teacher–student agreement, increases smoothly and monotonically with the Fisher–Rao discrepancy, and approaches one as the distributions separate. Consequently, the resulting token-level objective is

$$\mathcal{L}_{\text{mix}}^{(t)} = (1 - \lambda_t) D_{\text{RKL}}\left(\pi_\theta^{(t)} \,\|\, \pi_T^{(t)}\right) + \lambda_t D_{\text{FKL}}\left(\pi_T^{(t)} \,\|\, \pi_\theta^{(t)}\right). \tag{14}$$

In this manner, high-discrepancy tokens are optimized toward a more FKL-oriented objective, enabling the student to recover overlooked teacher-supported modes through more conservative updates (smaller w_t in Section 4.1).

4.4 Overall Training Objective

IG-OPD integrates the three modules hierarchically: prompt routing selects between teacher-forced SFT and OPD, while within the OPD branch token weighting controls supervision strength and adaptive KL orientation balances RKL and FKL. For a prompt x, the resulting objective is

$$\mathcal{L}_{\text{IG-OPD}}(x) = \begin{cases} \mathcal{L}_{\text{SFT}}(x), & \widehat{D}(x) \geq D_{\text{thresh}}, \\ \mathcal{L}_{\text{OPD}}(x), & \widehat{D}(x) < D_{\text{thresh}}, \end{cases} \tag{15}$$

where $\mathcal{L}_{\text{SFT}}$ denotes standard teacher-forced cross-entropy on a teacher-generated trajectory from the original prompt, and the OPD-branch loss is

$$\mathcal{L}_{\text{OPD}}(x) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)}\left[\frac{1}{|y|}\sum\nolimits_{t=1}^{|y|} w_t \left((1 - \lambda_t) D_{\text{RKL}}\left(\pi_\theta^{(t)} \,\|\, \pi_T^{(t)}\right) + \lambda_t D_{\text{FKL}}\left(\pi_T^{(t)} \,\|\, \pi_\theta^{(t)}\right)\right)\right]. \tag{16}$$

Here, the rollout, w_t, and λ_t are evaluated using the current student but treated as fixed during each gradient update; gradients are propagated only through the student distributions appearing in the KL terms. The complete training procedure is summarized in Algorithm 1 in the Appendix.

Appendix C provides complementary local analyses of the OPD branch. Under standard local convexity and smoothness assumptions, positive token weights preserve local contraction, and the adaptive KL coefficient varies smoothly with the discrepancy.

5 Experiments

We evaluate IG-OPD on textual and multimodal tasks under both conventional and strong-to-weak distillation settings. Our evaluation covers comparisons with strong OPD baselines, empirical analyses of reference-direction conflict and its relation to geometric discrepancy, and component ablations. Unless otherwise specified, all experiments use three random seeds, and the main tables report mean performance across seeds. Implementation details are provided in Appendix E.

Method	Math Reasoning					Code Generation			
	AIME 2024	AIME 2025	HMMT Feb.	HMMT Nov.	Avg.	HumanEval$^+$	MBPP$^+$	LiveCodeBench	Avg.
Student†	23.0	19.3	12.3	9.2	16.0	77.4	65.3	17.7	53.5
Teacher†	60.1	55.1	32.5	38.5	46.6	85.2	69.8	26.6	60.5
Single-Teacher Distillation									
MiniLLM	53.8$_{-6.3}$	49.6$_{-5.5}$	29.1$_{-3.4}$	34.7$_{-3.8}$	41.8$_{-4.8}$	83.0$_{-2.2}$	68.5$_{-1.3}$	25.2$_{-1.4}$	58.9$_{-1.6}$
DistiLLM	55.3$_{-4.8}$	51.2$_{-3.9}$	30.5$_{-2.0}$	36.3$_{-2.2}$	43.3$_{-3.3}$	83.3$_{-1.9}$	68.8$_{-1.0}$	25.8$_{-0.8}$	59.3$_{-1.2}$
OPD†	57.9$_{-2.2}$	52.4$_{-2.7}$	30.2$_{-2.3}$	37.8$_{-0.7}$	44.6$_{-2.0}$	82.6$_{-2.6}$	68.8$_{-1.0}$	25.7$_{-0.9}$	59.0$_{-1.5}$
EOPD	56.0$_{-4.1}$	52.1$_{-3.0}$	31.2$_{-1.3}$	36.6$_{-1.9}$	44.0$_{-2.6}$	83.6$_{-1.6}$	69.4$_{-0.4}$	26.4$_{-0.2}$	59.8$_{-0.7}$
ToDi	57.3$_{-2.8}$	53.3$_{-1.8}$	32.1$_{-0.4}$	37.4$_{-1.1}$	45.0$_{-1.6}$	84.1$_{-1.1}$	69.9$_{+0.1}$	26.5$_{-0.1}$	60.2$_{-0.3}$
TIP	58.3$_{-1.8}$	53.7$_{-1.4}$	32.2$_{-0.3}$	38.0$_{-0.5}$	45.6$_{-1.0}$	84.5$_{-0.7}$	70.3$_{+0.5}$	27.2$_{+0.6}$	60.7$_{+0.2}$
SCOPE	58.8$_{-1.3}$	54.4$_{-0.7}$	32.9$_{+0.4}$	38.2$_{-0.3}$	46.1$_{-0.5}$	85.2$_{0.0}$	70.2$_{+0.4}$	27.6$_{+1.0}$	61.0$_{+0.5}$
ExOPD†	62.7$_{+2.6}$	56.1$_{+1.0}$	33.9$_{+1.4}$	39.3$_{+0.8}$	48.0$_{+1.4}$	86.9$_{+1.7}$	70.7$_{+0.9}$	28.6$_{+2.0}$	62.1$_{+1.6}$
Uni-OPD†	63.3$_{+3.2}$	57.0$_{+1.9}$	34.8$_{+2.3}$	39.8$_{+1.3}$	48.7$_{+2.1}$	88.3$_{+3.1}$	71.6$_{+1.8}$	29.7$_{+3.1}$	63.2$_{+2.7}$
IG-OPD (ours)	**64.6**$_{+4.5}$	**58.8**$_{+3.7}$	**36.2**$_{+3.7}$	**41.1**$_{+2.6}$	**50.2**$_{+3.6}$	**89.6**$_{+4.4}$	**73.0**$_{+3.2}$	**31.3**$_{+4.7}$	**64.6**$_{+4.1}$
Multi-Teacher Distillation									
Off-policy SFT	58.1$_{-2.0}$	53.4$_{-1.7}$	30.9$_{-1.6}$	34.7$_{-3.8}$	44.3$_{-2.3}$	86.6$_{+1.4}$	69.5$_{-0.3}$	26.3$_{-0.3}$	60.8$_{+0.3}$
OPD†	60.9$_{+0.8}$	55.2$_{+0.1}$	33.4$_{+0.9}$	38.3$_{-0.2}$	47.0$_{+0.4}$	86.3$_{+1.1}$	70.9$_{+1.1}$	23.4$_{-3.2}$	60.2$_{-0.3}$
ExOPD†	61.0$_{+0.9}$	56.0$_{+0.9}$	34.4$_{+1.9}$	39.2$_{+0.7}$	47.7$_{+1.1}$	86.3$_{+1.1}$	70.6$_{+0.8}$	29.0$_{+2.4}$	62.0$_{+1.5}$
Uni-OPD†	62.3$_{+2.2}$	57.2$_{+2.1}$	34.9$_{+2.4}$	39.6$_{+1.1}$	48.5$_{+1.9}$	88.0$_{+2.8}$	72.6$_{+2.8}$	30.1$_{+3.5}$	63.6$_{+3.1}$
IG-OPD (ours)	**63.7**$_{+3.6}$	**58.9**$_{+3.8}$	**36.3**$_{+3.8}$	**41.5**$_{+3.0}$	**50.1**$_{+3.5}$	**89.6**$_{+4.4}$	**73.7**$_{+3.9}$	**31.6**$_{+5.0}$	**65.0**$_{+4.5}$

Table 1: Performance of the Qwen3-4B student model on mathematical reasoning and code generation tasks. † denotes results from Hou et al. (2026). The teacher models utilized are Qwen3-4B-Math-RL and Qwen3-4B-Code-RL, respectively. Each colored subscript denotes the change relative to the corresponding domain teacher (improvement; degradation). The best results are indicated in **bold**, while the second-best results are underlined.

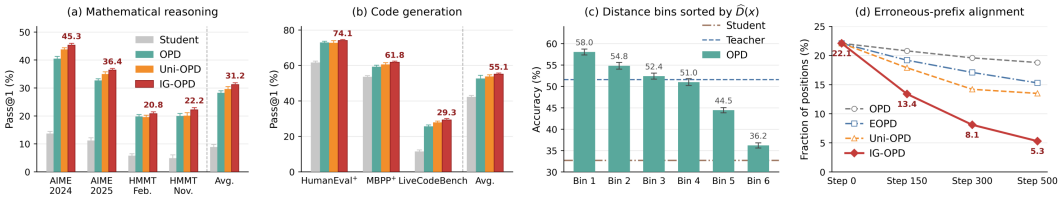


Figure 2: Multi-teacher distillation performance of the Qwen3-1.7B student on (a) mathematical reasoning and (b) code-generation benchmarks. The teachers are Qwen3-4B-Math-RL and Qwen3-4B-Code-RL. (c) Accuracy of vanilla OPD within equal-sized distance bins sorted by $\widehat{D}(x)$. (d) Fraction of paired positions with negative Fisher alignment ($\rho_t < 0$) between the idealized directions induced by controlled perturbed prefixes and their selected verified references during training.

5.1 Results for Textual Distillation

Table 1 summarizes the performance of the Qwen3-4B student across mathematical reasoning and code generation benchmarks. IG-OPD achieves the best overall performance among the evaluated OPD methods in both single-teacher and multi-teacher settings. In the single-teacher setting, IG-OPD outperforms Uni-OPD, the strongest competing baseline, by an average of 1.5 points on mathematical reasoning (50.2 vs. 48.7) and 1.4 points on code generation (64.6 vs. 63.2), with gains on every individual benchmark. In the multi-teacher setting, the corresponding average improvements over Uni-OPD are 1.6 and 1.4 points, respectively. Figs. 2(a) and (b) report the mean performance and standard deviation for the Qwen3-1.7B student. IG-OPD improves over Uni-OPD by 1.6 points on mathematics and 1.3 points on code generation. We have verified that the 95% paired hierarchical bootstrap confidence intervals for IG-OPD's gains over the strongest baseline remain entirely above zero, indicating that the improvements are robust. Additional evaluations in Appendix F.3 further indicate that IG-OPD preserves the student's general capabilities, including language understanding, commonsense reasoning, and instruction following, across various downstream benchmarks.

5.2 Results for Multimodal Distillation

IG-OPD requires no architectural changes to the multimodal model, as its controllers are computed from teacher and student next-token distributions. Token-level control and prompt-level discrepancy aggregation are restricted to text-generation positions; image-input and other masked positions contribute to neither the distillation loss nor the routing score. Table 2 reports results for a Qwen3-VL-4B-Instruct student on multimodal mathematical reasoning, logical reasoning, and document

Method	Math Reasoning				Logic Reasoning				Document Understanding				
	Math Vision	Dyna Math	We Math	Avg.	Vista (Acc)	Vista (Format)	Visu Logic	Avg.	AI2D	Chart QA	Doc VQA	Info VQA	Avg.
Student†	33.8	62.2	67.5	54.5	49.9	66.4	25.1	47.1	81.7	73.5	94.9	79.8	82.5
Teacher†	47.2	65.3	79.5	64.0	52.5	73.8	27.4	51.2	82.5	76.4	95.1	81.6	83.9
Single-Teacher Distillation													
OPD†	47.5$_{+0.3}$	64.8$_{-0.5}$	77.5$_{-2.0}$	63.3$_{-0.7}$	49.8$_{-2.7}$	73.0$_{-0.8}$	26.1$_{-1.3}$	49.6$_{-1.6}$	82.4$_{-0.1}$	75.4$_{-1.0}$	95.2$_{+0.1}$	81.4$_{-0.2}$	83.6$_{-0.3}$
Uni-OPD†	47.8$_{+0.6}$	65.4$_{+0.1}$	78.3$_{-1.2}$	63.8$_{-0.2}$	53.1$_{+0.6}$	73.8$_{0.0}$	28.2$_{+0.8}$	51.7$_{+0.5}$	82.6$_{+0.1}$	75.8$_{-0.6}$	95.2$_{+0.1}$	81.2$_{-0.4}$	83.7$_{-0.2}$
IG-OPD (ours)	**48.5**$_{+1.3}$	**66.2**$_{+0.9}$	**79.2**$_{-0.3}$	**64.6**$_{+0.6}$	**54.2**$_{+1.7}$	**75.1**$_{+1.3}$	**29.4**$_{+2.0}$	**52.9**$_{+1.7}$	**83.1**$_{+0.6}$	**76.6**$_{+0.2}$	**96.1**$_{+1.0}$	**81.7**$_{+0.1}$	**84.4**$_{+0.5}$
Multi-Teacher Distillation													
OPD†	41.0$_{-6.2}$	60.9$_{-4.4}$	71.7$_{-7.8}$	57.9$_{-6.1}$	51.3$_{-1.2}$	72.3$_{-1.5}$	26.3$_{-1.1}$	50.0$_{-1.2}$	82.6$_{+0.1}$	75.0$_{-1.4}$	95.1$_{0.0}$	81.3$_{-0.3}$	83.5$_{-0.4}$
Uni-OPD†	45.5$_{-1.7}$	62.3$_{-3.0}$	76.1$_{-3.4}$	61.3$_{-2.7}$	54.0$_{+1.5}$	75.2$_{+1.4}$	27.5$_{+0.1}$	52.2$_{+1.0}$	83.0$_{+0.5}$	75.7$_{-0.7}$	95.3$_{+0.2}$	81.6$_{0.0}$	83.9$_{0.0}$
IG-OPD (ours)	**47.0**$_{-0.2}$	**63.4**$_{-1.9}$	**77.5**$_{-2.0}$	**62.6**$_{-1.4}$	**55.8**$_{+3.3}$	**76.5**$_{+2.7}$	**29.1**$_{+1.7}$	**53.8**$_{+2.6}$	**83.7**$_{+1.2}$	**76.6**$_{+0.2}$	**95.9**$_{+0.8}$	**82.3**$_{+0.7}$	**84.6**$_{+0.7}$

Table 2: Performance of Qwen3-VL-4B-Instruct student under math reasoning, logic reasoning, and document understanding benchmarks. Teacher models (i.e., Qwen3-VL-4B-Instruct-Math-RL, Qwen3-VL-4B-Instruct-Logic-RL, and Qwen3-VL-4B-Instruct-Document-RL) are developed through domain-specific RL.

Method	Math Reasoning					Code Generation			
	AIME 2024	AIME 2025	HMMT Feb.	HMMT Nov.	Avg.	HumanEval$^+$	MBPP$^+$	LiveCodeBench	Avg.
Teacher†	72.1	61.4	42.5	57.1	58.3	81.9	77.2	23.4	60.8
Qwen3-4B									
Student†	23.0	19.3	12.3	9.2	16.0	77.4	65.3	17.7	53.5
OPD†	56.5	46.4	28.5	33.4	41.2	82.9	72.4	21.6	59.0
Uni-OPD†	55.9	50.2	29.8	35.6	42.9	83.1	71.3	28.0	60.8
IG-OPD (ours)	**57.6**	**52.1**	**31.1**	**36.8**	**44.4**	**84.8**	**73.5**	**29.3**	**62.5**
Qwen3-1.7B									
Student†	13.9	11.1	5.6	4.9	8.9	61.9	53.4	11.9	42.4
OPD†	35.7	27.6	17.2	14.6	23.8	67.1	56.7	23.4	49.1
Uni-OPD†	35.2	30.7	17.7	16.4	25.0	71.5	58.6	28.0	52.7
IG-OPD (ours)	**37.1**	**32.2**	**19.6**	**18.1**	**26.8**	**73.2**	**59.8**	**29.3**	**54.1**

Table 3: Results for the strong-to-weak distillation setting under math reasoning and code generation benchmarks. The teacher model is *Qwen3-30B-A3B-Instruct*, and the student models are the smaller *Qwen3-4B* and *Qwen3-1.7B*.

understanding benchmarks. IG-OPD outperforms both vanilla OPD and Uni-OPD in single- and multi-teacher settings. Averaged across benchmarks, it improves over Uni-OPD and vanilla OPD by 0.9 and 1.8 percentage points, respectively, with a single teacher; the corresponding gains increase to 1.2 and 3.2 points with multiple teachers. These results show that the gains of IG-OPD extend to the evaluated multimodal tasks, with larger improvements in the multi-teacher setting.

5.3 Strong-to-Weak Distillation

To evaluate strong-to-weak distillation, we use Qwen3-30B-A3B-Instruct as the teacher and Qwen3-4B and Qwen3-1.7B as students, with joint training on mathematical reasoning and code generation data. As shown in Table 3, IG-OPD achieves the best performance across both student sizes and all evaluated benchmarks. Relative to Uni-OPD, it improves the domain-level averages by 1.4–1.8 points, while the gains over vanilla OPD range from 3.0 to 5.0 points. These results show that the gains of IG-OPD persist in the stronger-teacher setting and across different student scales. A controlled shared-threshold diagnostic in Appendix F.9 further demonstrates that the strong-to-weak setting exhibits a right-shifted prompt-discrepancy distribution: under the same absolute cutoff, a larger fraction of prompts exceeds the threshold than in the same-size setting.

5.4 Geometric Diagnostics

We examine whether the geometric signals underlying IG-OPD exhibit the empirical relationships motivating its controllers. Fig. 2(c) demonstrates that vanilla OPD attains lower final accuracy in bins with larger initial prompt-level discrepancy $\widehat{D}(x)$. This association motivates SFT alignment for high-discrepancy prompts. At the token level, controlled-prefix diagnostics reveal a positive association between the tractable discrepancy $\widehat{d}_t$ and increasingly negative reference alignment[4]. At the Step-500 vanilla-OPD checkpoint, $\mathrm{Spearman}(\widehat{d}_t, -\rho_t) = 0.38$ ($p < 0.001$), indicating a moderate positive association between teacher–student discrepancy and more negative reference

[4]The controlled-prefix construction and diagnostic protocol are detailed in Appendix F.1.

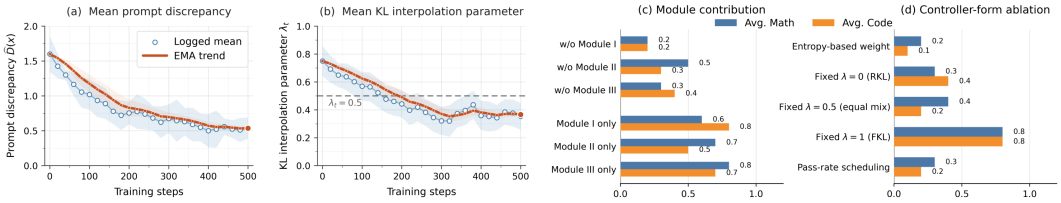


Figure 3: Training dynamics of IG-OPD. We track (a) the mean prompt-level discrepancy $\widehat{D}(x)$ under fixed-threshold routing and (b) the mean KL interpolation coefficient. (c) Performance degradation after removing individual modules or retaining only one module. (d) Performance degradation after replacing the proposed controllers with fixed or heuristic alternatives. Values denote the performance drop relative to full IG-OPD.

alignment. Beyond the controlled perturbation banks, Appendix F.12 presents naturally occurring mathematical and code-generation errors from actual trajectories, for which offline reference-pairing diagnostics also exhibit negative Fisher alignment following the erroneous transition. We further examine the magnitude interpretation derived in Section 3.2. For positions with $\rho_t < 0$, we measure the resulting signed change in the half-squared distance to $p^{\mathrm{ref}}_{T,t}$. Larger $\widehat{d}_t$ is associated with larger reference displacement ($\mathrm{Spearman} = 0.56$, $p = 0.0002$), consistent with its role as an observable correction-magnitude factor.

Fig. 2(d) further shows that IG-OPD exhibits a lower negative-alignment rate than the compared OPD methods. Finally, Fig. 3(a) indicates that the mean prompt-level discrepancy decreases during training, while Fig. 3(b) shows a decline in the parameter-free KL coefficient, indicating a gradual shift from greater FKL contribution toward RKL as teacher–student agreement improves. Together, these diagnostics support the use of $\widehat{d}_t$ as an observable magnitude-control signal while separately characterizing its empirical association with increasingly negative reference alignment.

5.5 Ablation Study

Fig. 3(c) evaluates the contributions of the three IG-OPD modules: geometry-aware token weighting (I), discrepancy-guided prompt routing (II), and adaptive KL orientation (III). Removing any component degrades performance, while variants retaining only a single component also remain below full IG-OPD, suggesting that the three modules provide complementary benefits. Fig. 3(d) further replaces the proposed controllers with entropy-based weighting, fixed KL mixtures, or pass-rate-based routing. Each alternative underperforms its corresponding IG-OPD controller, showing that the gains arise not only from combining the three mechanisms but also from their specific discrepancy-guided designs. IG-OPD also remains robust to moderate variations in its control hyperparameters, as detailed in Appendix F.7. Controlled substitutions in Sections F.5 and F.6 show that replacing the token-weighting or KL-orientation controller, or the shared Bhattacharyya/Fisher signal, consistently reduces performance, supporting their specific geometric formulations.

6 Conclusion & Discussion

This study introduces IG-OPD, an information-geometry-guided framework for improving vanilla OPD. Our analysis derives an exact finite-step characterization of when an idealized Fisher–Rao correction toward a deviated-prefix teacher target moves away from a selected reference target, expresses the observable correction magnitude through the Bhattacharyya coefficient, and distinguishes the conditional projection geometries induced by RKL and FKL. These results motivate three complementary modules: geometry-aware token weighting, discrepancy-guided prompt routing, and parameter-free adaptive KL orientation. Extensive experiments across textual and multimodal reasoning tasks, student scales, and teacher configurations demonstrate consistent improvements over strong distillation baselines, while controlled-prefix diagnostics support the empirical relationships underlying the proposed geometric signals.

Limitations and Future Work Our theory characterizes complementary geometric aspects of OPD under tractable conditional assumptions, rather than the full evolving training dynamics. Moreover, the counterfactual reference direction required for the alignment diagnostic is inaccessible during training, motivating future work on developing direct estimators of directional alignment.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5be69a584901a26c521c2b51e40a4c20-Paper-Conference.pdf.

Shun-ichi Amari. *Differential-geometrical methods in statistics*. Springer, New York, 1985.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Shun-ichi Amari. *Information geometry and its applications*. Springer, 2016.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions. *Advances in Neural Information Processing Systems*, 38, 2026.

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

Nicolas Boizard, Kevin El-Haddad, Celine Hudelot, and Pierre Colombo. Towards cross-tokenizer distillation: The universal logit distillation loss for llms. *Transactions on Machine Learning Research (TMLR)*, 2025.

Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

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. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025a.

Xiao Cui, Mo Zhu, Yulei Qin, Liang Xie, Wengang Zhou, and Houqiang Li. Multi-level optimal transport for universal cross-tokenizer knowledge distillation on language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23724–23732, 2025b.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8ac015d409635f196f9e3e9dcfb9a94e-Paper-Conference.pdf.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 138306–138322, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/df7b97ccf54c528516b62aa9a012d02a-Paper-Conference.pdf.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

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

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, Chengquan Zhang, Zhuotao Tian, Han Hu, Yi Yang, Fei Wu, and Hehe Fan. Uni-opd: Unifying on-policy distillation with a dual-perspective recipe. *arXiv preprint arXiv:2605.03677*, 2026.

Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. Selectkd: Selective token-weighted knowledge distillation for llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19287–19297, 2026.

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025.

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

Seongryong Jung, Suwan Yoon, DongGeon Kim, and Hwanhee Lee. Todi: Token-wise distillation via fine-grained divergence control. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pp. 235–251, 2016.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. In *International Conference on Machine Learning (ICML)*, 2024.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 31044–31062, 2025.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

M Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. *Advances in Neural Information Processing Systems*, 23, 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/e57c6b956a6521b28495f2886ca0977a-Paper.pdf.

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

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. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*, 36:21558–21572, 2023.

Tao Liu, Taiqiang Wu, Mao Zheng, Xuan Luo, Runming Yang, Xuewei Yang, Junjie Wang, and Yujiu Yang. Simple-opd: Demystifying warm-up for on-policy distillation. *arXiv preprint arXiv:2608.06802*, 2026.

Hoang-Chau Luong, Dat Ba Tran, and Lingwei Chen. Diversity-aware reverse kullback-leibler divergence for large language model distillation. *arXiv preprint arXiv:2604.00223*, 2026.

Jiaming Lv, Haoyuan Yang, and Peihua Li. Wasserstein distance rivals kullback-leibler divergence for knowledge distillation. *Advances in Neural Information Processing Systems*, 37:65445–65475, 2024. doi: 10.52202/079017-2090.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2582–2591, 2022.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Frank Nielsen. An elementary introduction to information geometry. *Entropy*, 22(10):1100, 2020.

Yifan Niu, Han Xiao, Dongyi Liu, Zelong Wang, Dihong Gong, Yasheng Wang, and Jia Li. Breaking the tokenizer barrier: On-policy distillation across model families. *arXiv preprint arXiv:2606.09456*, 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. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Peter Petersen. *Riemannian geometry*. Springer, 3rd edition, 2016.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20023–20070, 2025.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. In *NeurIPS Workshop on Energy Efficient Deep Learning*, 2019.

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.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024. doi: 10.52202/079017-3014.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. F-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10817–10834, 2023.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5737–5755, 2025.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 25966–26003, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/2bbc73b3d3c2de43743ce2d82c8f3d7d-Paper-Conference.pdf.

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

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, and Zhipeng Wang. Paced: Distillation and self-distillation at the frontier of student competence. *arXiv preprint arXiv:2603.11178*, 2026b.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026c.

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

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 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, 2026. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a4277440d50f1f15d2cb4c14f7e0c0d2-Paper-Conference.pdf.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D Lyng, Sanjit Singh Batra, and Robert E Tillman. Fast and effective on-policy distillation from reasoning prefixes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25553–25569, 2026a.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory (COLT)*, 2016.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025.

Kaichen Zhang, Keming Wu, Zuhao Yang, Bo Li, Kairui Hu, Bin Wang, Xingxuan Li, and Lidong Bing. Openmmreasoner: Pushing the frontiers in multimodal reasoning with an open and general recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19276–19286, 2026b.

Mengxiang Zhang and Lingyuan Liu. Staged knowledge distillation through least-to-most prompting: Optimizing teacher guidance via difficulty-aware training. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 8489–8501, 2025.

Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, and Jinan Xu. Dual-space knowledge distillation for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18164–18181, 2024.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4334–4343, 2019.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*, 2026a. URL https://openreview.net/forum?id=31wkVZXEaJ.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, et al. A survey of large language models. *Frontiers of Computer Science*, 20(12):2012627, 2026b.

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

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. Sod: Step-wise on-policy distillation for small language model agents. *arXiv preprint arXiv:2605.07725*, 2026.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Xiaoling Zhou, Shuaiyu Zhou, Zhemg Lee, Tao Chen, Xirui Li, Peng Chen, Jie Jiang, Wei Ye, and Shikun Zhang. Tailoring the training: Difficulty-aware learning strategy allocation for large language models. In *Forty-Third International Conference on Machine Learning*, 2026.

Zhou Ziheng, Jiaqi Li, Huacong Tang, Ying Nian Wu, and Demetri Terzopoulos. Less is more: Early stopping rollout for on-policy distillation. *arXiv preprint arXiv:2605.27028*, 2026.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, volume 2025, pp. 48337–48383, 2025.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, refine hypotheses, and provide feedback on experiment design. We have not used generative AI tools to generate synthetic datasets, implement methods, help develop theoretical models, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, clean and reformat datasets, or interpret results. Support for qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for summarizing or analyzing existing literature, brainstorming, sourcing or searching for information, editing the research paper to improve readability, and identifying relevant literature. We have reviewed all AI-assisted work. In particular, AI-generated text and suggestions were manually reviewed, revised, and verified by the authors; and literature-related outputs were cross-checked against the original sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of AI.

A ADDITIONAL RELATED WORK

On-Policy Distillation Recent advances in language-model distillation have explored on-policy supervision, divergence design, and adaptive use of student-generated trajectories. GKD (Agarwal et al., 2024) introduces explicit OPD on student-generated sequences, while MiniLLM (Gu et al., 2024) optimizes sequence-level reverse KL using student rollouts. DistiLLM (Ko et al., 2024) combines a skewed-KL objective with adaptive off-policy training, while DistiLLM-2 (Ko et al., 2025) introduces a contrastive strategy that increases the likelihood of teacher-generated responses while decreasing that of student-generated responses. Later work adapted supervision according to token- or trajectory-level signals: EOPD (Jin et al., 2026) uses teacher entropy; ToDi (Jung et al., 2025) selects between FKL and RKL using teacher–student probability ratios; TIP (Xu et al., 2026c) combines student uncertainty with teacher–student disagreement; and SCOPE (Zheng et al., 2026) assigns verified correct and incorrect rollouts to different objectives. Uni-OPD (Hou et al., 2026) further unifies complementary distillation objectives within an adaptive on-policy framework. Complementary work broadens OPD along optimization, efficiency, and curriculum dimensions, including KL-constrained reinforcement learning (Yang et al., 2026), cold-start alignment (Li et al., 2026), step-level reweighting for agentic tasks (Zhong et al., 2026), efficient prefix supervision (Zhang et al., 2026a; Ziheng et al., 2026), self-distillation without external teachers (Zhao et al., 2026a), and competence-aware prompt routing (Xu et al., 2026b); a broader overview is provided by Song & Zheng (2026). Different from existing studies, IG-OPD formalizes reference-direction conflict by comparing reference- and deviated-prefix teacher targets in the same student-anchored Fisher tangent space, and derives a tractable Bhattacharyya/Fisher–Rao discrepancy that is used as a geometric signal across the three control levels.

Information Geometry Information geometry provides dual m- and e-affine structures and the corresponding KL projection and Pythagorean identities (Amari, 1985; Amari & Nagaoka, 2000; Amari, 2016). The Fisher metric further provides the geometric foundation for natural-gradient optimization and Fisher–Rao distance (Amari, 1998; Martens, 2020), while Riemannian optimization theory offers tools for analyzing local optimization behavior (Zhang & Sra, 2016; Bonnabel, 2013; Petersen, 2016). Related geometric perspectives on distillation include optimal-transport-based distribution matching (Lv et al., 2024; Cui et al., 2025b) and representational alignment across tokenizers (Boizard et al., 2025; Niu et al., 2026; Zhang et al., 2024). Building on these foundations, we

use dual affine geometry to distinguish the projection structures associated with RKL and FKL, and Fisher–Rao geometry to characterize reference-direction conflict under deviated prefixes and derive a tractable Bhattacharyya-based discrepancy signal for adaptive OPD.

Connections to Policy Optimization OPD is conceptually related to KL-regularized policy optimization, where model updates are constrained relative to a reference distribution (Ouyang et al., 2022; Shao et al., 2024; Yu et al., 2026). In particular, on-policy optimization methods train on trajectories sampled from the evolving policy, paralleling OPD's use of student-generated trajectories. Related work on process supervision (Lightman et al., 2024) further emphasizes the value of dense intermediate feedback for reasoning. Unlike reward-based policy optimization, OPD provides token-level supervision directly through the teacher's conditional next-token distributions, without requiring a learned reward or process-reward model.

B Proofs for the Information-Geometric Analysis

B.1 Proof of Proposition 3.1

Proof. Use the square-root embedding $\phi(p) = 2(\sqrt{p_1}, \dots, \sqrt{p_K})$, which is an isometry from the categorical Fisher simplex to the positive orthant of the radius-2 sphere. Write

$$x = \phi(p_t), \qquad r = \phi(p^{\mathrm{ref}}_{T,t}), \qquad z = \phi(p^{\mathrm{dev}}_{T,t}),$$

and let

$$v^{\mathrm{ref}} = d\phi_{p_t}(u^{\mathrm{ref}}_t), \qquad v^{\mathrm{dev}} = d\phi_{p_t}(u^{\mathrm{dev}}_t).$$

The isometry gives

$$\|v^{\mathrm{ref}}\|_2 = a_t, \qquad \|v^{\mathrm{dev}}\|_2 = b_t, \qquad \frac{(v^{\mathrm{ref}})^\top v^{\mathrm{dev}}}{a_t b_t} = \rho_t. \tag{17}$$

Because all three distributions are interior, their square-root images have strictly positive coordinates. Hence $a_t, b_t \in (0, \pi)$, and the relevant spherical logarithms are unique.

The exponential map on a sphere of radius 2 gives

$$x^+ := \phi(p^+_t) = \cos\left(\frac{\eta b_t}{2}\right) x + 2\sin\left(\frac{\eta b_t}{2}\right) \frac{v^{\mathrm{dev}}}{b_t}, \tag{18}$$

$$r = \cos\left(\frac{a_t}{2}\right) x + 2\sin\left(\frac{a_t}{2}\right) \frac{v^{\mathrm{ref}}}{a_t}. \tag{19}$$

For $0 \le \eta \le 1$, Eq. (18) traces the minimizing great-circle segment from x to z; equivalently, its spherical-interpolation coefficients are nonnegative. It therefore stays in the positive orthant and corresponds to a valid categorical distribution.

Since $x^\top v^{\mathrm{ref}} = x^\top v^{\mathrm{dev}} = 0$ and $x^\top x = 4$, taking the inner product of Eqs. (18) and 19, and using Eq. (17), yields

$$\frac{(x^+)^\top r}{4} = \cos\left(\frac{a_t}{2}\right)\cos\left(\frac{\eta b_t}{2}\right) + \sin\left(\frac{a_t}{2}\right)\sin\left(\frac{\eta b_t}{2}\right)\rho_t. \tag{20}$$

The left-hand side equals $\cos(d_F(p^+_t, p^{\mathrm{ref}}_{T,t})/2)$, proving Eq. (4).

For the finite-step condition, set $A = a_t/2$ and $C = \eta b_t/2$. When $0 < \eta \le 1$, both $A, C \in (0, \pi/2)$. Moreover, x^+ and r remain in the positive orthant, so $d_F(p^+_t, p^{\mathrm{ref}}_{T,t})/2 \in [0, \pi/2)$. Since cosine is strictly decreasing on this interval, Eq. (4) implies

$$d_F(p^+_t, p^{\mathrm{ref}}_{T,t}) > a_t \iff \cos A \cos C + \sin A \sin C\, \rho_t < \cos A \tag{21}$$

$$\iff \rho_t < \frac{\cos A(1 - \cos C)}{\sin A \sin C} = \cot A \tan\left(\frac{C}{2}\right). \tag{22}$$

Its right-hand side is strictly positive, so $\rho_t < 0$ is sufficient for every $0 < \eta \le 1$.

Finally, let $d(\eta) = d_F(p_t^+, p_{T,t}^{\mathrm{ref}})$. Differentiating Eq. (4) at $\eta = 0$ gives

$$-\frac{1}{2}\sin\Big(\frac{a_t}{2}\Big)\, d'(0) = \frac{b_t}{2}\sin\Big(\frac{a_t}{2}\Big)\,\rho_t,$$

and hence $d'(0) = -b_t\rho_t$. Therefore

$$\frac{d}{d\eta}\frac{1}{2}d^2(\eta)\Big|_{\eta=0} = -a_t b_t \rho_t = -\left\langle u_t^{\mathrm{ref}}, u_t^{\mathrm{dev}}\right\rangle_{p_t}^F.$$

Taylor expansion at zero now gives Eq. (5). □

B.2 Proof of Proposition 3.2

Proof. Consider the square-root map

$$\phi(p) = 2(\sqrt{p_1}, \ldots, \sqrt{p_K}).$$

Its image lies in the positive orthant of the sphere of radius 2. For tangent vectors $u, v \in T_p\mathcal{M}$,

$$d\phi_p(u) = \left(\frac{u_i}{\sqrt{p_i}}\right)_{i=1}^K.$$

Hence

$$\langle d\phi_p(u), d\phi_p(v)\rangle_{\mathbb{R}^K} = \sum_i \frac{u_i v_i}{p_i} = \langle u, v\rangle_p^F,$$

so ϕ is an isometry onto the positive spherical orthant. The great-circle angle between $\phi(p)$ and $\phi(q)$ is

$$\arccos\left(\frac{\phi(p)^\top \phi(q)}{4}\right) = \arccos\left(\sum_i \sqrt{p_i q_i}\right).$$

Because the minimizing great-circle arc between two strictly positive points remains in the positive orthant, its spherical length equals the intrinsic Fisher–Rao distance on the simplex. Multiplying this angle by the sphere radius gives Eq. (7).

For the local expansion, let $d = d_F(p, q)$. The exact identity implies $\mathrm{BC}(p, q) = \cos(d/2)$. Therefore

$$-\log \mathrm{BC}(p, q) = -\log\cos(d/2) = \frac{d^2}{8} + \mathcal{O}(d^4),$$

which is equivalent to the local expansion stated in Proposition 3.2. □

B.3 Proof of Proposition 3.3

Proof. We work in $K-1$ independent mixture coordinates on the simplex, with $p_K = 1 - \sum_{i=1}^{K-1} p_i$. For $\Phi(p) = \sum_{i=1}^K p_i \log p_i$, the induced Bregman divergence is

$$D_\Phi(a, b) = D_{\mathrm{KL}}(a\|b),$$

and its dual coordinates are $\xi_i(p) = \log(p_i/p_K)$.

For part (i), the Bregman three-point identity gives

$$D_\Phi(p, p_T) = D_\Phi(p, p_R) + D_\Phi(p_R, p_T) + \langle p - p_R, \nabla\Phi(p_R) - \nabla\Phi(p_T)\rangle.$$

Because $\mathcal{C}_m$ is affine and p_R is a relative-interior minimizer, the first-order optimality condition implies

$$\langle p - p_R, \nabla\Phi(p_R) - \nabla\Phi(p_T)\rangle = 0, \text{for every } p \in \mathcal{C}_m.$$

Substituting $D_\Phi(a, b) = D_{\mathrm{KL}}(a\|b)$ proves Eq. (8).

For part (ii), let Φ^* denote the Legendre dual of Φ, and write $\xi_p = \nabla\Phi(p)$, $\xi_T = \nabla\Phi(p_T)$, and $\xi_F = \nabla\Phi(p_F)$. Define the feasible set in dual coordinates as

$$\Xi_e = \{\nabla\Phi(p) : p \in \mathcal{C}_e\}.$$

Because $\mathcal{C}_e$ is e-affine, Ξ_e is affine, and the relative-interior assumption on p_F implies that ξ_F lies in the relative interior of Ξ_e. Bregman duality gives

$$D_\Phi(a, b) = D_{\Phi^*}(\xi_b, \xi_a).$$

Consequently,

$$D_\Phi(p_T, p) = D_{\Phi^*}(\xi_p, \xi_T).$$

Thus, the definition of p_F is equivalent to

$$\xi_F = \arg\min_{\xi \in \Xi_e} D_{\Phi^*}(\xi, \xi_T).$$

Applying part (i) to Φ^* on the affine set Ξ_e therefore yields

$$D_{\Phi^*}(\xi_p, \xi_T) = D_{\Phi^*}(\xi_p, \xi_F) + D_{\Phi^*}(\xi_F, \xi_T) \tag{23}$$

$$= D_\Phi(p_T, p_F) + D_\Phi(p_F, p). \tag{24}$$

Using $D_\Phi(a, b) = D_{\mathrm{KL}}(a \| b)$ proves Eq. (9). □

C Proofs of the Conditional Properties of the OPD Branch

We analyze an idealized local surrogate of the OPD branch by conditioning on fixed rollout contexts and fixed controller values. Let $\{(c_j, p_T^{(j)}, w_j, \lambda_j)\}_{j=1}^M$ denote M fixed token positions, their teacher targets, token weights, and KL interpolation coefficients. Under our weighting rule, $w_j = \max\{w_{\min}, \exp(-\beta \widehat{d}_j)\}$, where $w_{\min} > 0$ is the fixed method-level weight floor (set to 0.1 in our experiments). We define

$$\overline{\mathcal{L}}(\theta) = \frac{1}{M} \sum_{j=1}^{M} w_j \left[(1 - \lambda_j) D_{\mathrm{RKL}}\left(p_\theta^{(j)} \,\|\, p_T^{(j)} \right) + \lambda_j D_{\mathrm{FKL}}\left(p_T^{(j)} \,\|\, p_\theta^{(j)} \right) \right], \tag{25}$$

where

$$p_\theta^{(j)} = \pi_\theta(\cdot \mid c_j), \qquad p_T^{(j)} = \pi_T(\cdot \mid c_j).$$

The contexts, weights, and interpolation coefficients are held fixed throughout the analyzed optimization sequence. This surrogate focuses on the local optimization behavior of the OPD branch, abstracting away the evolving rollout and routing dynamics of the full training procedure.

C.1 Conditional Local Convergence

For the fixed context collection $\mathcal{C} = \{c_j\}_{j=1}^M$ in Eq. (25), define the joint output map

$$\Phi(\theta) = \left(p_\theta^{(1)}, \ldots, p_\theta^{(M)} \right) \in \mathcal{P} := \prod_{j=1}^{M} \mathcal{S}_K. \tag{26}$$

For tangent vectors $u = (u^{(1)}, \ldots, u^{(M)})$ and $v = (v^{(1)}, \ldots, v^{(M)})$ at $z = (p^{(1)}, \ldots, p^{(M)}) \in \mathcal{P}$, equip the product manifold with the context-averaged Fisher metric

$$G_z(u, v) = \frac{1}{M} \sum_{j=1}^{M} \sum_{i=1}^{K} \frac{u_i^{(j)} v_i^{(j)}}{p_i^{(j)}}. \tag{27}$$

Thus, each factor of G is the fixed-context categorical Fisher metric defined in Section 2. Its pullback through Φ is

$$\begin{aligned} g_\theta(a, b) &= G_{\Phi(\theta)}\left(d\Phi_\theta[a], d\Phi_\theta[b] \right) = a^\top F_{\mathcal{C}}(\theta) b, \\ F_{\mathcal{C}}(\theta) &= \frac{1}{M} \sum_{j=1}^{M} \mathbb{E}_{Y \sim p_\theta^{(j)}} \left[\nabla_\theta \log p_\theta^{(j)}(Y) \nabla_\theta \log p_\theta^{(j)}(Y)^\top \right]. \end{aligned} \tag{28}$$

The pullback g is positive semidefinite and is singular whenever a nonzero parameter direction produces no first-order output change at all fixed contexts. We therefore do not assume that $F_{\mathcal{C}}(\theta)$ is nonsingular on the full neural-network parameter space. Instead, let $\mathcal{M}_{\mathrm{out}}$ be a local regular

model-output submanifold of $\mathcal{P}$ contained in the image of Φ, endowed with the restriction of G. Equivalently, $\mathcal{M}_{\text{out}}$ describes the locally identifiable quotient obtained after removing $\ker d\Phi_\theta$. The restricted metric is nondegenerate because it is induced from the Riemannian product metric G. For $z = \Phi(\theta) \in \mathcal{M}_{\text{out}}$, let $\ell_j(z)$ denote the mixed token loss in brackets in Eq. (25), and write

$$\mathcal{L}_0(z) = \frac{1}{M}\sum_{j=1}^{M} \ell_j(z), \qquad \overline{\mathcal{L}}(z) = \frac{1}{M}\sum_{j=1}^{M} w_j \ell_j(z). \tag{29}$$

With this output-space representation, $\overline{\mathcal{L}}(\Phi(\theta)) = \overline{\mathcal{L}}(\theta)$ for the surrogate in Eq. (25).

Theorem C.1 (Conditional local contraction)**.** *Let $\mathcal{U} \subset \mathcal{M}_{\text{out}}$ be a geodesically convex neighborhood of a minimizer z^* of $\overline{\mathcal{L}}$. Assume that:*

(A1) every ℓ_j is geodesically convex on $\mathcal{U}$;
(A2) $\mathcal{L}_0$ is geodesically μ_0-strongly convex and L_0-smooth on $\mathcal{U}$;
(A3) $w_{\min} \le w_j \le 1$ for some $w_{\min} > 0$;
(A4) for every k, the update geodesic

$$\gamma_k(s) = \operatorname{Exp}_{z_k}\left(-s\eta \operatorname{grad}^G \overline{\mathcal{L}}(z_k)\right), \qquad s \in [0,1],$$

remains in $\mathcal{U}$, and set $z_{k+1} = \gamma_k(1)$.

Then $\overline{\mathcal{L}}$ is geodesically μ-strongly convex with $\mu \ge w_{\min}\mu_0$ and is L_0-smooth. For $0 < \eta \le 1/L_0$,

$$\overline{\mathcal{L}}(z_k) - \overline{\mathcal{L}}(z^*) \le (1 - \eta w_{\min}\mu_0)^k \left[\overline{\mathcal{L}}(z_0) - \overline{\mathcal{L}}(z^*)\right]. \tag{30}$$

Theorem C.1 is deliberately conditional, which isolates a sufficient local regime in which positive token weights preserve contraction of the idealized output-model surrogate.

Proof. For any $z \in \mathcal{U}$ and tangent vector $v \in T_z\mathcal{M}_{\text{out}}$, linearity of the Riemannian Hessian, (A1), and $w_j \ge w_{\min}$ give

$$\begin{aligned}
\operatorname{Hess}^G \overline{\mathcal{L}}(z)[v,v] &= \frac{1}{M}\sum_{j=1}^{M} w_j \operatorname{Hess}^G \ell_j(z)[v,v] \\
&\ge \frac{w_{\min}}{M}\sum_{j=1}^{M} \operatorname{Hess}^G \ell_j(z)[v,v] \\
&= w_{\min} \operatorname{Hess}^G \mathcal{L}_0(z)[v,v] \\
&\ge w_{\min}\mu_0 \|v\|_{z,G}^2,
\end{aligned} \tag{31}$$

where $\|v\|_{z,G}^2 = G_z(v,v)$. Similarly, (A1), $w_j \le 1$, and the L_0-smoothness of $\mathcal{L}_0$ imply

$$\operatorname{Hess}^G \overline{\mathcal{L}}(z)[v,v] \le \operatorname{Hess}^G \mathcal{L}_0(z)[v,v] \le L_0 \|v\|_{z,G}^2. \tag{32}$$

Thus $\overline{\mathcal{L}}$ is geodesically μ-strongly convex on $\mathcal{U}$, with $\mu \ge w_{\min}\mu_0$.

The Riemannian descent lemma, the exponential-map update, and $\eta \le 1/L_0$ yield

$$\overline{\mathcal{L}}(z_{k+1}) \le \overline{\mathcal{L}}(z_k) - \frac{\eta}{2}\left\|\operatorname{grad}^G \overline{\mathcal{L}}(z_k)\right\|_{z_k,G}^2. \tag{33}$$

Geodesic μ-strong convexity also implies the gradient-dominance inequality

$$\left\|\operatorname{grad}^G \overline{\mathcal{L}}(z)\right\|_{z,G}^2 \ge 2\mu\left[\overline{\mathcal{L}}(z) - \overline{\mathcal{L}}(z^*)\right]. \tag{34}$$

Combining Eqs. (33) and (34) gives a one-step contraction by $1 - \eta\mu$. Since $\mu \ge w_{\min}\mu_0$, iterating this inequality proves Eq. (30). □

C.2 Deterministic Gradient Attenuation

Proposition C.2 (Deterministic gradient attenuation). *Let*

$$g_t = \nabla_\theta \mathcal{L}_{\text{mix}}^{(t)}$$

denote the unweighted token gradient in a fixed parameter-space norm, and assume $\|g_t\| \le \mathcal{G}$. For $\beta > 0$, $0 < w_{\min} \le 1$, and $\widehat{d}_t \ge 0$, let $w_t = \max\{w_{\min}, \exp(-\beta \widehat{d}_t)\}$ be treated as a stop-gradient controller. Then the weighted token contribution satisfies

$$\|w_t g_t\| \le \mathcal{G} \max\{w_{\min}, \exp(-\beta \widehat{d}_t)\}. \tag{35}$$

Thus, the upper bound decays exponentially with the discrepancy while $\widehat{d}_t \le \beta^{-1} \log(1/w_{\min})$, and saturates at $w_{\min}\mathcal{G}$ once the minimum-weight floor becomes active. The floor prevents token-level supervision from vanishing at large discrepancy.

Proof. Because w_t is detached, weighting the token loss scales its gradient by the fixed scalar w_t. Hence

$$\|w_t g_t\| = \max\{w_{\min}, \exp(-\beta \widehat{d}_t)\} \|g_t\| \le \mathcal{G} \max\{w_{\min}, \exp(-\beta \widehat{d}_t)\},$$

which proves Eq. (35). □

C.3 Stability of the KL Interpolation

Proposition C.3 (Interpolation stability). *Let $d \ge 0$ denote a generic discrepancy value, with $d = \widehat{d}_t$ in IG-OPD. The controller*

$$\lambda(d) = 1 - \exp(-d)$$

is strictly increasing, maps $[0, \infty)$ into $[0, 1)$, satisfies $\lambda(0) = 0$ and $\lim_{d\to\infty} \lambda(d) = 1$, and is 1-Lipschitz:

$$|\lambda(d_1) - \lambda(d_2)| \le |d_1 - d_2| \qquad \text{for all } d_1, d_2 \ge 0. \tag{36}$$

Proof. Direct differentiation gives

$$\lambda'(d) = \exp(-d).$$

Since

$$0 < \lambda'(d) \le 1,$$

the controller is strictly increasing, and the mean-value theorem proves Eq. (36). The endpoint properties follow directly from the exponential function. □

D Algorithm

The complete IG-OPD training procedure is summarized in Algorithm 1.

E Experimental Setup

E.1 Training Settings

Models For textual experiments, we use Qwen3-4B (Yang et al., 2025) as the primary student and Qwen3-1.7B for scale validation. Domain-specific teacher models are obtained by applying GRPO (Shao et al., 2024) to the same Qwen3-4B backbone on task-specific data, yielding Qwen3-4B-Math-RL and Qwen3-4B-Code-RL. For the strong-to-weak setting, we use Qwen3-30B-A3B-Instruct as the teacher. For multimodal experiments, we use Qwen3-VL-4B-Instruct (Bai et al., 2025) as the student. Domain-specific multimodal teachers are obtained by applying RL to the same backbone, yielding Qwen3-VL-4B-Instruct-Math-RL, Qwen3-VL-4B-Instruct-Logic-RL, and Qwen3-VL-4B-Instruct-Document-RL. Meanwhile, to assess the generalizability of our method across model architectures, we further conduct experiments on the LLaMA family (Grattafiori et al., 2024), using LLaMA-3.2-3B-Instruct as the student model and LLaMA-3.1-8B-Instruct as the teacher model.

Algorithm 1 IG-OPD Training.

Require: Student π_θ, teacher π_T, training set $\mathcal{D}$, calibration set $\mathcal{D}_{\text{calib}}$, token-weighting coefficient β, minimum token-weight floor $w_{\min}$, pilot rollout count $\mathcal{N}$, total OPD rollout count $R \geq \mathcal{N}$, and total training steps $\mathcal{T}$.

Estimate $\{\widehat{D}^{(0)}(x) : x \in \mathcal{D}_{\text{calib}}\}$ using $\mathcal{N}$ initial student rollouts per prompt;
Set the fixed D_{thresh} using Eq. (12);
for $k = 1, \ldots, \mathcal{T}$ **do**
  Sample a prompt minibatch $\mathcal{B} \subset \mathcal{D}$;
  Initialize $\mathcal{L}_{\mathcal{B}} = 0$;
  for each $x \in \mathcal{B}$ **do**
    Sample a pilot set $\mathcal{Y}_x^{\text{pilot}} = \{y^{(n)}\}_{n=1}^{\mathcal{N}}$ from $\pi_\theta(\cdot \mid x)$ without differentiating through rollout sampling;
    for each $y^{(n)} \in \mathcal{Y}_x^{\text{pilot}}$ and position t **do**
      Set $c_t^{(n)} = (x, y_{<t}^{(n)})$;
      Evaluate $\pi_\theta(\cdot \mid c_t^{(n)})$ and $\pi_T(\cdot \mid c_t^{(n)})$ without gradient tracking;
      Compute and cache $\widehat{d}_t^{(n)}$;
    end for
    Compute $\widehat{D}(x)$ using Eq. (10);
    if $\widehat{D}(x) \geq D_{\text{thresh}}$ **then**
      Obtain the teacher-generated trajectory $y^T(x)$;
      Compute $\mathcal{L}_x^{\text{SFT}} = -\frac{1}{|y^T(x)|} \sum_{t=1}^{|y^T(x)|} \log \pi_\theta\big(y_t^T(x) \mid x, y_{<t}^T(x)\big)$;
      $\mathcal{L}_{\mathcal{B}} \leftarrow \mathcal{L}_{\mathcal{B}} + \mathcal{L}_x^{\text{SFT}}$;
    else
      Sample $R - \mathcal{N}$ additional student rollouts and combine them with $\mathcal{Y}_x^{\text{pilot}}$ to form $\mathcal{Y}_x^{\text{OPD}} = \{y^{(r)}\}_{r=1}^{R}$;
      for each $y^{(r)} \in \mathcal{Y}_x^{\text{OPD}}$ and position t **do**
        Set $c_t^{(r)} = (x, y_{<t}^{(r)})$;
        Evaluate $\pi_\theta(\cdot \mid c_t^{(r)})$ with gradient tracking and $\pi_T(\cdot \mid c_t^{(r)})$ without gradient tracking;
        Compute $\widehat{d}_t^{(r)}$ if not cached during routing;
        Set the detached controllers $w_t^{(r)} = \text{sg}\Big(\max\Big\{w_{\min}, \exp(-\beta \widehat{d}_t^{(r)})\Big\}\Big)$, and $\lambda_t^{(r)} = \text{sg}\Big(1 - \exp(-\widehat{d}_t^{(r)})\Big)$;
        Compute $\mathcal{L}_{\text{mix}}^{(r,t)}$ using Eq. (14);
      end for
      Compute $\mathcal{L}_x^{\text{OPD}} = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{|y^{(r)}|} \sum_{t=1}^{|y^{(r)}|} w_t^{(r)} \mathcal{L}_{\text{mix}}^{(r,t)}$;
      $\mathcal{L}_{\mathcal{B}} \leftarrow \mathcal{L}_{\mathcal{B}} + \mathcal{L}_x^{\text{OPD}}$;
    end if
  end for
  Update θ using the chosen optimizer on $\mathcal{L}_{\mathcal{B}}/|\mathcal{B}|$;
end for

Training Data For textual math reasoning, following Hou et al. (2026), we utilize 57K samples from DeepMath (He et al., 2026). For textual code generation, we use 25K samples from the code subset of Eurus-2-RL-Data (Cui et al., 2025a). For multimodal tasks, we employ math reasoning, logic reasoning, and document understanding data mainly from OpenMMReasoner-RL-74K (Zhang et al., 2026b). Specifically, for multimodal math reasoning tasks, 14.8K samples are drawn from the OpenMMReasoner-RL dataset[5], covering MMK12, We-Math-Standard, and We-Math-Pro subsets. For multimodal logic reasoning data, 14.8K samples are collected, spanning AlgoPuzzle, PuzzleVQA, and ThinkLiteVL-Hard subsets from the OpenMMReasoner-RL-74K dataset. For multimodal document understanding data, we include 14.6K document understanding samples, obtained

[5] https://huggingface.co/datasets/OpenMMReasoner/OpenMMReasoner-RL-74K

Prompt template

Math reasoning

```
<|im_start|>user
{question}
Please reason step by step, and put your final answer within \boxed{}.<|im_end|>
<|im_start|>assistant
```

Code reasoning

```
<|im_start|>user
{question}
Write Python code to solve the problem. Present the code in
```python
Your code
```
at the end.
You need to think first then write the Python code.<|im_end|>
<|im_start|>assistant
```

Multimodal reasoning

```
<|im_start|>user
<image>
{question}
Please solve the problem step by step and put your answer in one \boxed{}.<|im_end|>
<|im_start|>assistant
```

Figure 4: Prompt templates used for mathematical, code, and multimodal reasoning experiments.

by 15% sampling from the TQA subset of OpenMMReasoner with ChartQA (Masry et al., 2022) and InfographicsVQA (Mathew et al., 2022) training sets.

Setups We follow the general student-distillation settings of prior work (Hou et al., 2026; Yang et al., 2026), including the training data, optimizer, learning-rate schedule, sequence-length limits, and evaluation protocol. For all methods run in our pipeline, we use the same training and evaluation configuration unless a method requires its own method-specific hyperparameters. The training prompts are provided in Fig. 4. Each OPD update uses a batch size of 64 and 16 on-policy rollouts per prompt. To reduce routing overhead, $\widehat{D}(x)$ is first estimated from 4 rollouts. Prompts assigned to teacher-forced SFT require no further student generation, whereas those assigned to OPD receive 12 additional rollouts; the initial four rollouts are retained when computing the OPD loss. The domain-specific teachers are trained using GRPO with the configuration reported in Table 4.

For OPD, we inherit most hyperparameters, including the learning rate, optimizer, and sequence length, from the teacher RL configuration reported in Table 4, thereby ensuring that the student model is trained under an optimization regime consistent with that of the teacher models. As for the OPD-specific entries, the training batch size is set to 64, and the number of on-policy samples per prompt is set to 16. Moreover, we set the token-weighting coefficient to $\beta = 1.0$ in all experiments except the sensitivity analysis. For each training configuration, D_{thresh} is calibrated once before training as the 80th percentile of $\widehat{D}^{(0)}(x)$ over a calibration subset of 1K prompts sampled from the training data, following Eq. (12). We recompute $\widehat{D}^{(k)}(x)$ from current student rollouts when routing each prompt, so assignments remain adaptive even though the threshold is fixed. In the multi-teacher setting, we use deterministic domain-based teacher routing: each training prompt is assigned to the domain-specific teacher corresponding to its task. Prompts routed to the SFT branch are trained on

Setting	Value
Training Schedule	
Text teacher steps (Math / Code)	500 / 300
Multimodal teacher steps (Math / Logic / Document)	300 / 300 / 160
Optimization and System	
Learning rate	1×10^{-6}
LR schedule	Constant
Optimizer	Adam
Training batch size	128
Micro batch size / GPU	1
Tensor parallelism	2
ZeRO stage	3
Max prompt length	2,048
Max response length	16,384
GRPO Configuration	
Clip ratio	0.2
KL loss coefficient	0.0
Entropy coefficient	0.0
Rollout Sampling	
Samples per prompt (n)	8
Temperature	1.0
Top-p	0.95
Top-k	50

Table 4: Hyperparameter configuration used to train the domain-specific teacher models with GRPO.

the teacher-generated trajectories. The rule-based verifiers from DeepMath[6], PRIME[7], and OpenMMReasoner[8] are used for textual mathematics, code generation, and multimodal reasoning, respectively, when outcome labels are required for teacher training, outcome-dependent baselines, or correctness-based analyses.

E.2 Evaluation Settings

Evaluation Setup For textual math reasoning, we utilize AIME 2024 [9] and 2025[10], and HMMT 2025 February and November (Balunovic et al., 2026). For textual code generation, we utilize HumanEval$^+$ (Liu et al., 2023), MBPP$^+$ (Liu et al., 2023), and LiveCodeBench v6 (Feb.–May 2025) (Jain et al., 2025). For multimodal math reasoning, we utilize MathVision (Wang et al., 2024), DynaMath (Zou et al., 2025), and We-Math (Qiao et al., 2025). For multimodal logic reasoning, we utilize LogicVista (Xiao et al., 2024) and VisuLogic (Xu et al., 2026a). For document understanding, we consider AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), and InfoVQA (Mathew et al., 2022). For general capabilities, we utilize MMLU (Hendrycks et al., 2020), ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), TruthfulQA (Lin et al., 2022), Winogrande (Sakaguchi et al., 2021), GSM8K (Cobbe et al., 2021), CommonsenseQA (Talmor et al., 2019), and IFEval (Zhou et al., 2023), to assess the retention of general-purpose capabilities after distillation.

For all textual benchmarks, we utilize the `vLLM` inference engine for efficient sampling, with a temperature of 1.0, top-p of 1.0, a maximum generation length of 16,384 tokens, and a fixed random

[6] https://github.com/zwhe99/DeepMath
[7] https://github.com/PRIME-RL/PRIME
[8] https://github.com/EvolvingLMMs-Lab/OpenMMReasoner
[9] https://huggingface.co/datasets/AI-MO/aimo-validation-aime
[10] https://huggingface.co/datasets/opencompass/AIME2025

Category	Task	Reported Metric
Multimodal Math Reasoning	MathVision Test	`mathvision_standard_eval`
	DynaMath Reasoning	`dynamath_average`
	We-Math TestMini Reasoning	`acc_score`
Multimodal Logic Reasoning	LogicVista Reasoning	`acc_score`, `format_score`
	VisuLogic	`visulogic_acc`
Document Understanding	AI2D	`exact_match`
	ChartQA	`relaxed_human_split`
	DocVQA Val	`anls`
	InfoVQA Val	`anls`

Table 5: Multimodal evaluation benchmarks and the corresponding LMMs-Eval metrics used for the main reported results.

seed of 42. We sample 32 solutions per problem for mathematical reasoning and 4 solutions per problem for code generation. For all main results, we report the average accuracy across sampled solutions (i.e., pass@1). To assess functional correctness, we employ `Math-Verify` as a rule-based verifier for mathematical tasks, while for code generation, we leverage the `EvalPlus` and `LiveCodeBench` frameworks. For multimodal evaluations, we adopt the `LMMs-Eval` framework (Zhang et al., 2025) and strictly adhere to its official evaluation protocols and configurations. A comprehensive summary of the metrics reported for each benchmark is provided in Table 5.

Baselines We compare IG-OPD with the following distillation baselines:

- **Off-policy SFT** applies standard teacher-forced cross-entropy to teacher-generated trajectories.
- **OPD** applies uniformly weighted token-level KL on student-generated trajectories.
- **MiniLLM** (Gu et al., 2024) optimizes sequence-level reverse KL on student-generated sequences using a policy-gradient estimator.
- **DistiLLM** (Ko et al., 2024) combines a skew-KL objective with an adaptive off-policy strategy.
- **EOPD** (Jin et al., 2026) augments clipped reverse-KL OPD with forward-KL supervision at positions with high teacher entropy.
- **ToDi** (Jung et al., 2025) adaptively combines forward and reverse KL at each vocabulary element according to the teacher–student probability log-ratio.
- **SCOPE** (Zheng et al., 2026) routes rollouts by correctness, applying teacher-perplexity-weighted KL to incorrect trajectories and student-perplexity-weighted MLE to correct ones.
- **TIP** (Xu et al., 2026c) selects informative token positions using student entropy and teacher–student divergence.
- **Uni-OPD** (Hou et al., 2026) combines offline difficulty-aware and online correctness-aware data balancing with outcome-guided margin calibration.
- **ExOPD** (Yang et al., 2026) sets the reward-scaling factor above one in generalized OPD, extrapolating beyond direct teacher imitation.

F Additional Experiments

This section provides supplementary empirical analyses of IG-OPD. Section F.1 describes the controlled prefix-pairing protocol and Fisher-alignment computation. Sections F.2 and F.3 report additional textual-distillation comparisons and evaluate general-capability preservation. Section F.4 evaluates the cross-architecture generalizability of IG-OPD on the LLaMA model family. Sections F.5 and F.6 provide controller- and signal-substitution ablations. Section F.7 examines hyperparameter sensitivity, followed by the effect of the pilot rollout count in Section F.8. Sections F.9 and F.10 present routing and reference-alignment diagnostics. Finally, Sections F.11 and F.12 report computational overhead and case studies.

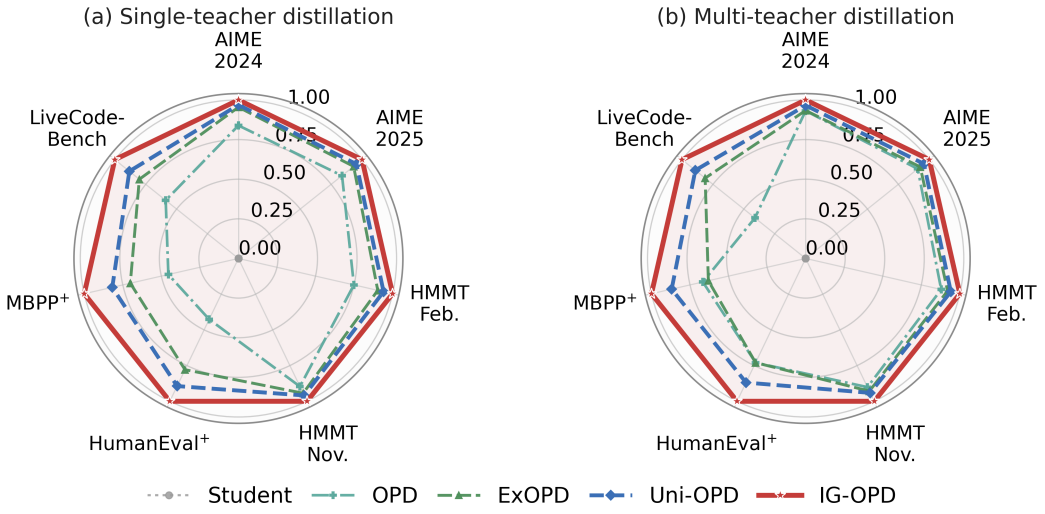


Figure 5: Performance comparison of different distillation methods under (a) single-teacher and (b) multi-teacher settings. Scores are normalized by the best performance achieved on each benchmark.

F.1 Controlled Prefix Pairing and Fisher-Alignment Diagnostic

This section describes the controlled offline diagnostics used for Fig. 2(d) and the token-level analyses in Section 5.4. We construct separate paired-prefix banks for mathematical reasoning and code generation. Both banks are used only for offline evaluation and are held fixed across methods, checkpoints, and random seeds. For the mathematical-reasoning bank, we start from a verified reference solution and modify a numerical value, arithmetic operator, or equation at a verifier-checkable reasoning step while leaving the preceding trace unchanged. A rule-based verifier followed by manual review certifies the reference step as valid and the deviated step as invalid; ambiguous cases and alternative valid solutions are excluded. We compare the contexts immediately following the corresponding complete reasoning steps rather than aligning trajectories by token index. Thus, both the location of the controlled deviation and the correspondence between the paired reasoning steps are known by construction. The bank contains 200 problems and 220 candidate perturbations. After excluding 14 ambiguous cases, 206 accepted paired contexts remain; removing six pairs with a zero-norm Fisher tangent vector yields 200 valid paired positions. As for the code-generation bank, we additionally construct a unit-test-verified controlled bank from the HumanEval$^+$ and MBPP$^+$ problems. Starting from a reference program that passes its complete test suite, we introduce one localized, syntax-preserving mutation to a boundary condition, comparison or arithmetic operator, loop range, index offset, or state-update expression. The unchanged reference program must pass all tests, whereas the mutated program must fail at least one test without a syntax error or timeout. Mutations that remain functionally equivalent under the available tests, admit an alternative valid interpretation, or alter multiple semantic units are excluded. As in the mathematical bank, the paired contexts are taken immediately after the complete mutated line or block rather than matched by token index. From 100 candidate mutations, unit-test verification and manual review retain 86 paired contexts; removing six zero-norm pairs yields 80 valid positions, split equally between HumanEval$^+$ and MBPP$^+$. All methods are evaluated at Steps 0, 150, 300, and 500 using the same tokenizer, decoding settings, domain-specific controlled banks, and random seeds. Fig. 2(d) reports the equal-weight average of the fractions of valid positions with $\rho_t < 0$ in the two banks.

To compute the alignment ρ_t defined in Section 3.1, we use the isometric square-root embedding of the categorical simplex into the radius-2 sphere,

$$\phi(p) = 2(\sqrt{p_1}, \ldots, \sqrt{p_K}). \tag{37}$$

Method	MMLU	ARC	HellaSwag	TruthfulQA	Winogrande	GSM8K	CommonsenseQA	IFEval	Avg.
Student†	68.3	80.7	68.4	54.8	66.6	84.2	**75.8**	88.9	73.5
Math Teacher†	**68.4**	**80.8**	**68.5**	54.3	66.0	86.7	75.4	89.2	73.7
Code Teacher†	68.3	80.2	68.3	54.8	65.7	85.8	75.7	89.7	73.6
OPD†	68.3	80.3	68.4	54.6	66.5	88.6	75.5	89.2	73.9
Uni-OPD†	68.3	80.3	68.3	54.6	66.0	88.6	75.7	89.2	73.9
IG-OPD	68.3	80.6	**68.5**	**55.0**	**66.8**	**89.0**	75.6	**89.9**	**74.2**

Table 6: General downstream performance of the Qwen3-4B student after multi-teacher distillation. † denotes results reported in Hou et al. (2026). Avg. is the unweighted average of the eight benchmark scores and is included as a descriptive summary.

Let $z_t = \phi(p_t)$, $z^s_{T,t} = \phi(p^s_{T,t})$, $s \in \{\mathrm{ref}, \mathrm{dev}\}$, where p_t is the student distribution evaluated at the deviated context and serves as the common base point. Define

$$\varphi^s_t = \arccos\left(\frac{z_t^\top z^s_{T,t}}{4}\right), \qquad v^s_t = \frac{\varphi^s_t}{\sin\varphi^s_t}\left(z^s_{T,t} - \cos\varphi^s_t z_t\right), \tag{38}$$

where $v^s_t = \mathrm{Log}_{z_t}(z^s_{T,t})$, with the continuous extension at $\varphi^s_t = 0$. Under the radius-2 embedding, $\|v^s_t\|_2 = 2\varphi^s_t = d_F(p_t, p^s_{T,t})$. Since the square-root embedding is an isometry between the Fisher–Rao simplex and the radius-2 sphere, $v^s_t = d\phi_{p_t}(u^s_t)$, where $u^s_t = \mathrm{Log}_{p_t}(p^s_{T,t})$. Hence, the alignment ρ_t can be computed as

$$\rho_t = \frac{(v^{\mathrm{ref}}_t)^\top v^{\mathrm{dev}}_t}{\|v^{\mathrm{ref}}_t\|_2 \, \|v^{\mathrm{dev}}_t\|_2}. \tag{39}$$

For numerical stability, we clip the argument of $\arccos$ to $[-1, 1]$ and exclude positions at which either tangent vector has zero norm. We report the fraction of valid positions satisfying $\rho_t < 0$, which operationalizes the idealized reference-direction conflict characterized in Proposition 3.1.

As for reference-displacement diagnostic, for every valid controlled position with $\rho_t < 0$, we restore the same Step-500 vanilla-OPD model and optimizer state and perform one actual optimizer update using the standard vanilla-OPD loss with $p^{\mathrm{dev}}_{T,t}$ as the teacher target. Let $p^{\mathrm{pre}}_t = \pi_\theta(\cdot \mid c^{\mathrm{dev}}_t)$ and $p^{\mathrm{post}}_t = \pi_{\theta^+}(\cdot \mid c^{\mathrm{dev}}_t)$ denote the full-vocabulary student distributions before and after this update, respectively. We measure the signed change in reference distance as

$$\Delta^{\mathrm{ref}}_t = \frac{1}{2} d^2_F\big(p^{\mathrm{post}}_t, p^{\mathrm{ref}}_{T,t}\big) - \frac{1}{2} d^2_F\big(p^{\mathrm{pre}}_t, p^{\mathrm{ref}}_{T,t}\big). \tag{40}$$

Restoring the checkpoint and optimizer state before each diagnostic update prevents order effects across controlled positions. We then relate the pre-update discrepancy $\widehat{d}_t$ to Δ^{ref}_t on the combined mathematical-reasoning and code-generation banks.

F.2 More Comparison Results for Textual Distillation

To facilitate a more intuitive comparison, Fig. 5 presents radar-chart visualizations of the results reported in Table 1. Across both single- and multi-teacher distillation settings, IG-OPD consistently achieves the strongest performance on all mathematical reasoning and code-generation benchmarks, demonstrating its effectiveness in both domain-specific knowledge transfer and knowledge integration from multiple expert teachers.

F.3 General Capability Preservation

Table 6 evaluates the Qwen3-4B student on eight general-purpose benchmarks. The resulting model does not exhibit an apparent degradation on any evaluated task. The largest observed gains over the original student occur on GSM8K and IFEval, whereas the changes on the remaining benchmarks are comparatively small. We interpret the results as evidence that IG-OPD preserves broad language understanding, commonsense reasoning, and instruction-following capabilities while improving the skills most closely related to its reasoning-oriented training data.

Method	AIME24		AIME25	
	Pass@8	Avg@8	Pass@8	Avg@8
OPD	10.2	2.0	6.9	1.5
EOPD	<u>13.3</u>	<u>3.1</u>	<u>10.4</u>	<u>2.2</u>
IG-OPD	**14.8**	**4.2**	**12.3**	**3.4**

Table 7: Comparison results on the LLaMA series, with LLaMA-3.2-3B-Instruct as the student model and LLaMA-3.1-8B-Instruct as the teacher model.

Variant	Math Reasoning					Code Generation			
	AIME 2024	AIME 2025	HMMT Feb.	HMMT Nov.	Avg.	HumanEval$^+$	MBPP$^+$	LiveCodeBench	Avg.
IG-OPD w/ TIP controller	64.3	57.9	35.6	40.6	49.6	89.0	72.7	30.4	64.0
IG-OPD w/ ToDi controller	64.2	58.1	35.8	40.8	49.7	89.3	72.6	30.7	64.2
IG-OPD	**64.6**	**58.8**	**36.2**	**41.1**	**50.2**	**89.6**	**73.0**	**31.3**	**64.6**

Table 8: Controller-substitution results under the Qwen3-4B single-teacher textual-distillation settings of Table 1. "w/ TIP" replaces only Module I with Soft-OR top-50% token selection; "w/ ToDi" replaces only Module III with the vocabulary-token-level ToDi controller.

F.4 Generalization Evaluation

To evaluate the generalizability of the proposed IG-OPD across different model families and architectures, we conduct additional experiments on the LLaMA series (Grattafiori et al., 2024). Specifically, we adopt LLaMA-3.2-3B-Instruct as the student model and LLaMA-3.1-8B-Instruct as the teacher model. Pass@8 and Avg@8 are employed as the evaluation metrics. As shown in Table 7, IG-OPD consistently outperforms vanilla OPD and EOPD, demonstrating that the effectiveness of IG-OPD is not restricted to a particular model family and that the proposed method can generalize well to different model architectures.

F.5 Controller-Substitution Ablation with TIP and ToDi

We conduct a component-substitution experiment to distinguish the benefit of the complete IG-OPD design from that of inserting an existing adaptive controller into the same architecture. Each variant replaces exactly one module while retaining the other two. IG-OPD w/ TIP replaces the continuous Fisher/BC token weighting in Module I with TIP's Soft-OR controller, which ranks positions using student entropy and teacher–student disagreement and retains the top 50% for distillation (Xu et al., 2026c). Prompt routing and adaptive KL orientation remain unchanged. IG-OPD w/ ToDi retains our token weighting and prompt routing but replaces the position-level KL interpolation in Module III with ToDi's vocabulary-element-level controller, which adjusts the forward- and reverse-KL contributions according to the teacher–student probability ratio (Jung et al., 2025). We evaluate the three variants using the Qwen3-4B single-teacher textual distillation settings of Table 1. Table 8 shows that both external controllers remain competitive within the IG-OPD architecture, while the complete method achieves the best result on every benchmark. These findings demonstrate the effectiveness and complementarity of the three information-geometry-inspired components.

F.6 Signal-Substitution Ablation

We isolate the contribution of the Bhattacharyya/Fisher signal from that of the three-level IG-OPD architecture. Specifically, we retain prompt routing, token weighting, and adaptive KL orientation exactly as in the full method and replace only the token-level discrepancy. In addition to $s_t^{\mathrm{BC}} = -\log \mathrm{BC}(p_t, p_{T,t})$, we also consider the following two manners:

$$s_t^{\mathrm{JS}} = \tfrac{1}{2} D_{\mathrm{KL}}(p_t \,\|\, m_t) + \tfrac{1}{2} D_{\mathrm{KL}}(p_{T,t} \,\|\, m_t)\,, \qquad m_t = \tfrac{1}{2}(p_t + p_{T,t}), \tag{41}$$

$$s_t^{\mathrm{symKL}} = \tfrac{1}{2}[D_{\mathrm{KL}}(p_t \,\|\, p_{T,t}) + D_{\mathrm{KL}}(p_{T,t} \,\|\, p_t)]\,. \tag{42}$$

To remove differences caused purely by the numerical scales of these discrepancies, we calibrate each substitute signal once on the initial calibration distribution before applying the controllers. Specifically, each signal is multiplied by a fixed positive coefficient so that its 80th percentile

Controller signal	Math Reasoning					Code Generation			
	AIME 2024	AIME 2025	HMMT Feb.	HMMT Nov.	Avg.	HumanEval$^+$	MBPP$^+$	LiveCodeBench	Avg.
Symmetric KL	64.0	57.5	35.4	40.6	49.4	89.0	72.2	30.5	63.9
Jensen–Shannon	63.9	57.8	35.7	40.7	49.5	89.1	72.5	30.8	64.1
Bhattacharyya/Fisher (ours)	**64.6**	**58.8**	**36.2**	**41.1**	**50.2**	**89.6**	**73.0**	**31.3**	**64.6**

Table 9: Signal-substitution ablation under the Qwen3-4B single-teacher textual-distillation settings. The complete IG-OPD architecture is retained, and only its controller signal is replaced.

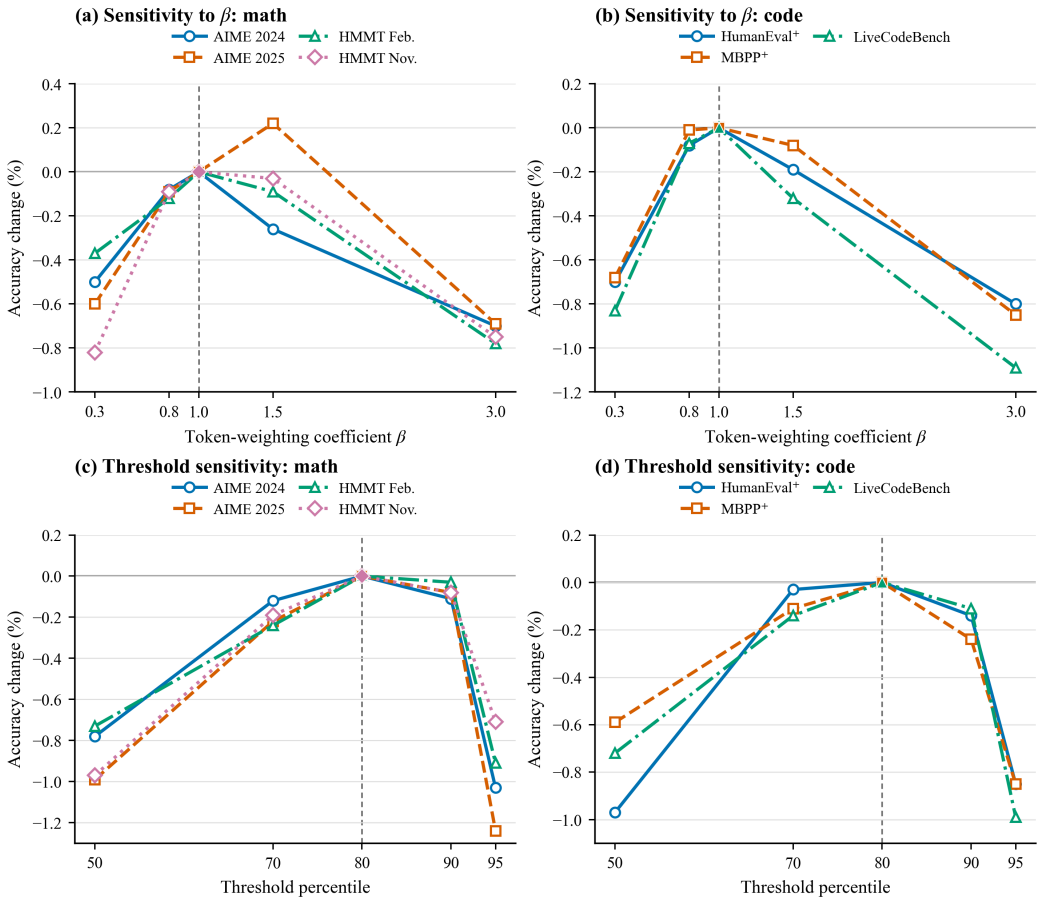


Figure 6: Sensitivity of IG-OPD to its revised hyperparameters. Panels (a) and (b) vary the geometric attenuation coefficient β, while panels (c) and (d) vary the percentile used to calibrate the fixed D_{thresh} on mathematical-reasoning and code-generation benchmarks, respectively. The KL controller is parameter-free and therefore requires no sensitivity sweep. Results are reported as changes in performance relative to the default configuration, which is marked by the vertical dashed lines.

matches that of the Bhattacharyya/Fisher discrepancy; the Bhattacharyya/Fisher coefficient is one. The calibrated signal is then used in token weighting, prompt routing, and adaptive KL orientation, with the routing threshold recalibrated for that signal using the same protocol as the full method. All calibration coefficients are frozen throughout training. This preserves the ordering induced by each signal while aligning their controller operating scales. As shown in Table 9, both substitute signals benefit from the same IG-OPD architecture, but the Bhattacharyya/Fisher signal achieves the strongest result on every benchmark. The scale-controlled comparison supports the use of the proposed signal beyond the benefit of the surrounding controller architecture.

$w_{\min}$	Math Reasoning					Code Generation			
	AIME 2024	AIME 2025	HMMT Feb.	HMMT Nov.	Avg.	HumanEval$^+$	MBPP$^+$	LiveCodeBench	Avg.
0	64.1	58.0	35.7	40.6	49.6	89.1	72.5	30.6	64.1
0.05	64.4	58.6	36.1	40.9	50.0	89.5	72.7	31.2	64.5
0.10	**64.6**	**58.8**	**36.2**	**41.1**	**50.2**	**89.6**	**73.0**	**31.3**	**64.6**
0.20	64.6	58.7	36.0	40.8	50.0	89.2	72.9	31.1	64.4

Table 10: Sensitivity to the minimum token weight $w_{\min}$ on the textual-distillation benchmarks. All other control hyperparameters are held at their default values.

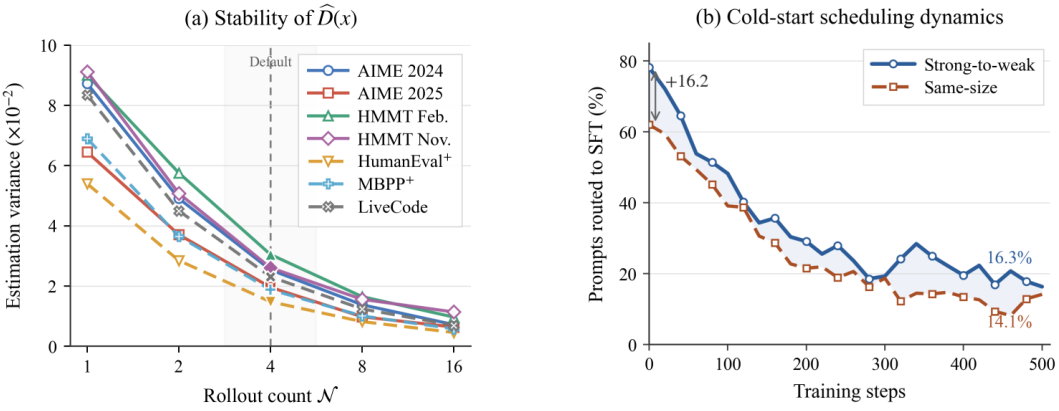


Figure 7: Empirical analysis of discrepancy-guided prompt routing. (a) Empirical variance of repeated prompt-level discrepancy estimates. (b) Shared-threshold diagnostic comparing the fraction of prompts above one common absolute cutoff in strong-to-weak and same-size distillation.

F.7 Hyperparameter Sensitivity Analysis

We examine the sensitivity of IG-OPD to its three hyperparameter choices: the geometric attenuation coefficient β, the minimum token weight $w_{\min}$, and the percentile used to calibrate the routing threshold D_{thresh}. We vary one while holding the others at their default values, namely $\beta = 1.0$, $w_{\min} = 0.1$, and the 80th percentile. Experiments are conducted with the Qwen3-4B student in the single-teacher setting.

Effect of β (geometric attenuation coefficient). Figs. 6(a) and (b) report performance as β varies from 0.3 to 3.0. Smaller values weakly attenuate geometrically large corrections, whereas larger values more aggressively suppress supervision at high-discrepancy positions and may also down-weight informative feedback at moderately discrepant positions. Performance is strongest overall at the intermediate setting $\beta = 1.0$, which we use by default.

Effect of D_{thresh} (prompt-routing threshold). Figs. 6(c) and (d) vary the percentile used to calibrate D_{thresh} from 50 to 95. Lower percentiles route a larger fraction of prompts to teacher-forced SFT, reducing the amount of on-policy training. In contrast, higher percentiles allow more highly discrepant prompts to enter the OPD branch, where supervision on student-generated contexts may be less stable. The 80th percentile achieves the strongest overall performance across various benchmarks, providing a favorable balance between teacher-forced supervision and OPD.

Effect of $w_{\min}$ (minimum token weight). We evaluate $w_{\min} \in \{0, 0.05, 0.1, 0.2\}$, where $w_{\min} = 0$ recovers the unclipped geometric weight. Increasing the floor preserves more supervision from geometrically extreme but potentially informative tokens, whereas an excessively large floor weakens the intended attenuation of large-discrepancy corrections. Table 10 uses the same mathematical-reasoning and code-generation benchmarks as Table 1. The results demonstrate that the default minimum weight of 0.1 is consistently above the other settings across benchmarks.

Component	Time (s)
Shared with standard OPD	
Student rollout generation	38.20
Teacher/student KL evaluation & optimization	7.10
IG-OPD additions	
BC discrepancy computation ($\widehat{d}_t$)	0.04
Detached w_t and λ_t controllers	0.02
Prompt score and routing decision	0.01
Total (IG-OPD)	**45.37**
Total (standard OPD)	45.30
Relative overhead	**+0.07**

Table 11: OPD-branch wall-clock microbenchmark under identical hardware and batch configurations. Both methods use $R = 16$ student rollouts per prompt; IG-OPD uses the first $\mathcal{N} = 4$ as pilot rollouts and retains them in the subsequent OPD loss.

F.8 Effect of Pilot Rollout Count $\mathcal{N}$ on Discrepancy Estimation

IG-OPD estimates the prompt-level discrepancy $\widehat{D}(x)$ by averaging token-level discrepancies over $\mathcal{N}$ student pilot rollouts. Increasing $\mathcal{N}$ reduces the sampling variability of this Monte Carlo estimate but requires additional rollout computation. For each prompt, we compute the empirical variance across repeated independent estimates of $\widehat{D}(x)$ and then average it over the calibration set. As shown in Fig. 7(a), this estimation variance decreases with $\mathcal{N}$. Considering the trade-off between estimation stability and routing overhead, we set $\mathcal{N} = 4$ as the default value in our experiments.

F.9 Shared-Threshold Routing Diagnostic

We conduct a controlled comparison between distilling Qwen3-30B-A3B-Instruct into Qwen3-4B and the same-size setting with a Qwen3-4B-Math-RL teacher and Qwen3-4B student. For this diagnostic only, we select a single absolute cutoff D_{shared} and apply it unchanged to both configurations. We therefore do not recompute an 80th-percentile threshold within either plotted distribution. This differs from the default algorithm, which calibrates a fixed threshold separately for each training configuration using Eq. (12). Fig. 7(b) reports the fraction of prompts satisfying $\widehat{D}^{(k)}(x) \geq D_{\text{shared}}$. Because D_{shared} is not defined as the 80th percentile of either initial distribution, these fractions are not expected to equal 20%. At initialization, approximately 78% of prompts in the strong-to-weak setting and 62% in the same-size setting exceed the shared cutoff, indicating higher initial teacher–student discrepancy in the stronger-teacher configuration under the same reference threshold. As training proceeds, the fraction in the strong-to-weak setting decreases to approximately 16%, showing that the prompt-level discrepancy is substantially reduced over training.

F.10 Additional Reference-Direction Conflict Diagnostics

We provide domain-specific and cross-domain diagnostics of the empirical relationship between geometric discrepancy and reference-direction conflict. Following Section 5.4, we use the mathematical-reasoning and code-generation controlled banks. For each valid paired position, we compute the Fisher alignment ρ_t between the idealized correction direction toward the deviated-prefix teacher target and the direction toward the corresponding verified reference target. Positions satisfying $\rho_t < 0$ are counted as reference-direction conflicts. The results in Fig. 2(d) demonstrate that IG-OPD has the lowest conflict fraction, highlighting that its controllers more effectively preserve geometric compatibility between teacher supervision on student-visited contexts and the correction associated with task-valid reference states.

We additionally examine the relationship between the tractable discrepancy $\widehat{d}_t$ and the directional alignment ρ_t at the Step-500 vanilla-OPD checkpoint. On the combined cross-domain bank containing valid mathematical-reasoning and code-generation positions, we obtain $\text{Spearman}(\widehat{d}_t, -\rho_t) = 0.38$, $p < 0.001$. This cross-domain result indicates a moderate positive association between

Vanilla OPD (incorrect)	**IG-OPD (correct)**
Problem Let $\alpha(n)$ be the number of distinct prime divisors of n. Find $\lim_{n\to\infty} \alpha(n)/n$. The verifier-provided answer is 0.	
Reasoning excerpt. Every distinct prime divisor is at least 2. Hence the sum of the distinct prime divisors is at least $2\alpha(n)$ and cannot exceed n, so $\alpha(n) \le n/2$. **First invalid step.** The bound is attained asymptotically by even integers; therefore $\alpha(n)/n \to 1/2$. The inequality is only a loose upper bound: even integers do not have $n/2$ distinct prime factors. **Final answer.** $\boxed{1/2}$ **Diagnostic.** Error: $\widehat{d}_t = 0.94$, $\rho_t = -0.22$; next step: $\widehat{d}_t = 1.21$, $\rho_t = -0.41$.	**Reasoning excerpt.** Let $k = \alpha(n)$, and let $p_1, \dots, p_k$ be the distinct prime divisors of n. Their product divides n, so **Key valid step.** $$n \ge \prod_{i=1}^{k} p_i \ge 2^k, \qquad k \le \log_2 n.$$ Consequently, $0 \le \alpha(n)/n \le (\log_2 n)/n \to 0$. **Final answer.** $\boxed{0}$ **Diagnostic.** Decision point: $\widehat{d}_t = 0.91$, $\rho_t = -0.18$, $w_t = 0.40$, $\lambda_t = 0.60$; recovery: $\widehat{d}_t = 0.31$, $\rho_t = +0.27$.

Table 12: Mathematical-reasoning case study. Red highlights the first invalid inference and incorrect answer produced by vanilla OPD, while blue highlights the valid IG-OPD argument. All geometric quantities are computed offline using the controlled reference-pairing diagnostic; in particular, ρ_t is not observed by IG-OPD during training or inference.

teacher–student discrepancy and more negative reference alignment. Furthermore, using the actual-update protocol, we evaluate the negatively aligned positions from the combined mathematical-reasoning and code-generation banks. The pre-update discrepancy is positively associated with the signed reference-distance change in Eq. (40), with $\mathrm{Spearman}(\widehat{d}_t, \Delta_t^{\mathrm{ref}}) = 0.56$ and $p = 0.0002$. Thus, the observable correction-magnitude signal remains informative under actual parameter updates across both controlled domains.

F.11 Computational Efficiency Analysis

We evaluate the wall-clock time of the OPD branch for IG-OPD. Table 11 reports an OPD-branch microbenchmark under identical hardware and batch configurations. Both methods use $R = 16$ student rollouts per prompt, and IG-OPD reuses the first $\mathcal{N} = 4$ rollouts for discrepancy estimation. Its controllers add 0.07 seconds per step, increasing the measured OPD-branch cost from 45.30 to 45.37 seconds. Moreover, prompt routing reduces rollout-side computation: prompts assigned to the SFT branch use their teacher-generated trajectories, reuse the $\mathcal{N} = 4$ pilot rollouts for routing, and require no additional student generation or online teacher generation, whereas prompts assigned to OPD use $R = 16$ rollouts. Therefore, the SFT-routing fraction can reduce the average number of student rollouts relative to applying OPD to every prompt.

F.12 Case Studies

We present one mathematical-reasoning example and one code-generation example. As for mathematical reasoning, the selected problem asks for the asymptotic ratio between the number of distinct prime divisors of n and n. As shown in Table 12, vanilla OPD derives the loose upper bound $\alpha(n) \le n/2$, but incorrectly treats it as asymptotically tight and returns $1/2$. IG-OPD instead derives the logarithmic bound $\alpha(n) \le \log_2 n$ and obtains the verified answer 0. Offline geometric diagnostics reveal negative Fisher alignment following the invalid transition in the vanilla-OPD trace. At a high-discrepancy context along the IG-OPD trace, the corresponding controller values are $w_t = 0.40$ and $\lambda_t = 0.60$, illustrating that such a state would receive attenuated supervision and a larger FKL contribution during training. At a subsequent diagnostic position, the discrepancy is lower and the Fisher alignment is positive. These observations are consistent with the proposed training mechanism.

Vanilla OPD (fails tests)

Task. Implement `count_zigzags(a)`, which returns the number of index quadruples (i, j, k, l) satisfying $1 \le i < j < k < l \le n$, $a_i = a_k$, and $a_j = a_l$, for $n \le 3000$.

Implementation excerpt.

```
def count_zigzags(a):
  n = len(a)
  suffix = [[0] * (n + 1)
        for q in range(n + 1)]
  for p in range(n - 1, -1, -1):
    suffix[p] = suffix[p +
1].copy()
    suffix[p][a[p]] += 1
  ans = 0; left = Counter()
  for j in range(1, n - 2):
    left[a[j - 1]] += 1
    for k in range(j + 1, n - 1):
      r = suffix[k][a[j]]
      ans += left[a[k]] * r
  return ans
```

Execution result.
`count_zigzags([2,2,2,2,2])`:
expected 5, observed 9.
Deterministic-test pass rate. $4/6$.
Diagnostic.
First overcount: $\widehat{d}_t = 1.17$, $\rho_t = -0.32$; next line: $\widehat{d}_t = 1.34$, $\rho_t = -0.45$.

IG-OPD (passes tests)

Implementation excerpt.

```
from collections import Counter
def count_zigzags(a):
  n = len(a)
  ans = 0
  left = Counter([a[0]])
  for j in range(1, n - 2):
    right = Counter(a[j + 1:])
    for k in range(j + 1, n - 1):
      right[a[k]] -= 1
      ans += left[a[k]] *
right[a[j]]
    left[a[j]] += 1
  return ans
```

Execution result.
`count_zigzags([2,2,2,2,2])`:
expected 5, observed 5.
Deterministic-test pass rate. $6/6$.
Diagnostic.
Decision point: $\widehat{d}_t = 1.05$, $\rho_t = -0.18$, $w_t = 0.35$, $\lambda_t = 0.65$; recovery: $\widehat{d}_t = 0.32$, $\rho_t = +0.29$.

Table 13: Code-generation case study.

For the code-generation task in Table 13, vanilla OPD precomputes suffix-frequency tables but queries `suffix[k]` rather than `suffix[k + 1]`. The resulting count includes position k itself as a candidate for l whenever $a_j = a_k$, violating the required ordering $k < l$. It therefore returns 9 instead of 5 on the all-equal example and passes only four of the six deterministic tests. The controlled code-prefix diagnostic identifies negative Fisher alignment at the first overcount in the vanilla-OPD trace ($\widehat{d}_t = 1.17$, $\rho_t = -0.32$); at the following diagnostic position, the discrepancy increases to 1.34 and the alignment decreases to -0.45. At the corresponding IG-OPD decision point, $\widehat{d}_t = 1.05$ gives $w_t = 0.35$ and $\lambda_t = 0.65$, attenuating the high-discrepancy supervision while assigning a larger FKL contribution. After the suffix update is correctly applied, the subsequent diagnostic shows lower discrepancy ($\widehat{d}_t = 0.32$) and positive alignment ($\rho_t = +0.29$).