GENERALIZED PROXIMAL POLICY OPTIMIZATION: UNIFYING LLM RL ALGORITHMS FROM THE PERSPECTIVE OF DIVERGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning has become the foundational paradigm for enhancing the reasoning capabilities of LLMs. Yet the dominant approach, Proximal Policy Optimization (PPO), relies on a geometrically degenerate clipping proxy that cannot adapt its resistance to policy shifts, leading to premature clipping and degraded exploration. Subsequent variants alleviate these issues through fragmented and heuristic modifications. We propose *Generalized Proximal Policy Optimization (GPPO)*, a principled and unified framework for understanding and designing LLM RL algorithms through the lens of divergence. We establish a generalized f-divergence trust-region theory with a monotonic performance improvement guarantee, and show that different f-divergences induce distinct geometries and policy update dynamics. We further introduce novel estimator topologies that provide more faithful divergence estimates than the conventional single-sample Monte Carlo estimation. Under this framework, PPO and recent variants such as DAPO, DPPO, and RIPO emerge as special cases, establishing a principled two-dimensional design space of *Divergence* $\times$ *Estimator*. Extensive experiments on eight competition-level math and coding benchmarks show that both dimensions are critical. The best-performing GPPO instantiation surpasses standard GRPO by up to 50% and 14% on math and coding, delivering SOTA performance.

1 INTRODUCTION

Reinforcement learning (RL) has established itself as the primary paradigm for advancing complex reasoning in LLMs (Guo et al., 2025; Comanici et al., 2025; OpenAI, 2024), powering breakthroughs across long-horizon domains such as mathematical reasoning (Shao et al., 2024; Wang et al., 2024; Luo et al., 2025), code generation (Dou et al., 2024; Anthropic, 2024), and deep search (OpenAI, 2025; Jin et al., 2025).

Proximal Policy Optimization (PPO) (Schulman et al., 2017) has become a dominant approach for stable and effective RL training, with its "proximal" behavior enforced through clipping. Yet beneath this heuristic mechanism lies a geometrically degenerate proxy: PPO implicitly constrains a single-sample Monte Carlo (MC) estimate of the Total Variation (TV) divergence, whose generating function has zero curvature and strictly symmetric bounds. This linear geometry cannot adapt its resistance to the magnitude of policy shifts, treating small and large deviations alike. This rigidity can lead to premature clipping, distorted trust-region scaling, and exploration collapse (Yu et al., 2025; Cai et al., 2026; Qi et al., 2026).

To alleviate these pathologies, recent variants (e.g., DAPO (Yu et al., 2025), DPPO (Qi et al., 2026), RIPO (Cai et al., 2026)) refine trust-region constraints through relaxing clipping bounds or adjusting Riemannian metrics. Despite their empirical gains, these methods are largely motivated by specific empirical observations, leading to fragmented modifications of the trust-region constraint. Consequently, existing approaches offer no unified, first-principles account of how these seemingly different mechanisms are related, why they improve optimization, or how to systematically design trust-region constraints beyond these isolated heuristics.

We propose Generalized Proximal Policy Optimization (GPPO), a principled framework that unifies the fragmented landscape of LLM RL through the lens of divergence. Rather than introducing

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.

another isolated algorithm, GPPO establishes a principled two-dimensional design space of *Divergence* $\times$ *Estimator*, enabling systematic design of proximal policy optimization.

At its core, GPPO establishes a generalized f-divergence trust-region theory, showing that policy optimization under f-divergence level-set constraints admits a monotonic performance improvement guarantee. Crucially, we show that different f-divergences induce distinct geometries, providing a principled basis for understanding their different policy update dynamics.

Beyond the choice of divergence, GPPO further generalizes how divergences are estimated, introducing novel estimator topologies that reduce single-sample Monte Carlo variance and provide more faithful divergence estimates. Under this unified lens, PPO and recent variants such as DAPO, DPPO, and RIPO emerge naturally as special cases of GPPO.

Empirically, our experiments reveal that proximal policy optimization is highly sensitive to both divergence geometry and estimator topology. Compared with the linear and symmetric TV geometry, divergences with non-zero curvature generally yield stronger performance. Estimator topology is equally important for that Atomic estimation consistently outperforms single-sample MC across all evaluated divergences. Notably, KL with Atomic estimation achieves the strongest performance, improving the average score by 50% on math and 14% on coding relative to the standard TV-MC formulation. These results validate the proposed two-dimensional design space that divergence geometry and estimator topology are both critical to effective RL optimization.

In summary, our contributions are threefold:

- We propose Generalized Proximal Policy Optimization (GPPO), establishing a two-dimensional design space of *Divergence* $\times$ *Estimator*, which unifies PPO and recent proximal variants, and inspires principled exploration of new proximal policy optimization algorithms.
- We establish a generalized f-divergence trust-region theory with a monotonic performance improvement guarantee, and illustrate how divergence geometry shapes policy update dynamics.
- We introduce novel estimator topologies that provide more faithful divergence estimates than single-sample Monte Carlo estimation, achieving SOTA performance on math and coding.

2 PRELIMINARY

2.1 MONOTONIC IMPROVEMENT UNDER TRUST REGION CONSTRAINT

Schulman et al. (2015) establishes a trust-region lower bound on the performance improvement between a behavior policy μ and a target policy π. Specifically,

Theorem 1 (Trust-Region Performance Bound)**.** *For behavior policy μ and target policy π, the performance improvement satisfies*

$$\eta(\pi) - \eta(\mu) \geq \frac{1}{1-\gamma}\mathbb{E}_{s\sim\rho^{\mu},a\sim\mu(\cdot|s)}\left[r(s,a)A^{\mu}(s,a)\right] - \frac{4\xi\gamma}{(1-\gamma)^2}D_{\mathrm{TV}}^{\max}(\pi \parallel \mu)^2, \tag{1}$$

where $\eta(\pi)$ denotes the expected discounted return, $r(s,a) = \frac{\pi(a|s)}{\mu(a|s)}$ is the importance ratio, A^{μ} is the advantage function under μ, $\xi = \max_{s,a}|A^{\mu}(s,a)|$, and $D_{\mathrm{TV}}^{\max}(\pi \parallel \mu) = \max_s D_{\mathrm{TV}}(\pi(\cdot|s) \parallel \mu(\cdot|s))$ is the maximum state-wise Total Variation (TV) divergence.

This bound motivates maximizing the surrogate objective $L_{\mu}(\pi) = \mathbb{E}_{s\sim\rho^{\mu},\, a\sim\mu(\cdot|s)}\left[r(s,a)A^{\mu}(s,a)\right]$, while controlling the policy discrepancy measured by $D_{\mathrm{TV}}^{\max}(\pi \parallel \mu)$, providing a formal monotonic improvement guarantee.

This principle underlies Trust Region Policy Optimization (TRPO) (Schulman et al., 2015), which maximizes the surrogate objective $\mathcal{L}_{\mu}(\pi)$ subject to a constraint that keeps the new policy within a trust region around the current policy. Define $\delta > 0$ as the trust region radius, formally, a TV-constrained trust-region formulation can be written as :

$$\begin{aligned} \max_{\pi} \quad & \mathcal{L}_{\mu}(\pi) \\ \text{s.t.} \quad & D_{\mathrm{TV}}^{\max}(\pi \parallel \mu) \leq \delta. \end{aligned} \tag{2}$$

Alternatively, this trust-region constraint can be formulated using the KL divergence, as theoretically justified by Pinsker's inequality: $D_{\mathrm{TV}}(\pi \parallel \mu)^2 \leq \frac{1}{2}D_{\mathrm{KL}}(\pi \parallel \mu)$.

2.2 Proximal Policy Optimization as a Divergence Proxy

Despite the theoretical guarantees of TRPO, its constrained second-order optimization incurs substantial computational overhead, limiting its scalability in modern LLM RL. Proximal Policy Optimization (PPO) (Schulman et al., 2017) addresses this limitation by replacing the second-order optimization with a first-order clipped surrogate objective:

$$\mathcal{J}_{\mathrm{PPO}}(\pi) = \mathbb{E}_{s\sim\rho^{\mu},\, a\sim\mu(\cdot|s)}\left[\min\left(r(s,a)\hat{A}(s,a), \mathrm{clip}\left(r(s,a), 1-\epsilon, 1+\epsilon\right)\hat{A}(s,a)\right)\right], \tag{3}$$

where $\hat{A}$ denotes the estimated advantage, typically obtained using GAE (Schulman et al., 2016) in PPO and group-relative advantages in modern LLM RL methods such as GRPO (Guo et al., 2025) and its variants (Yu et al., 2025; Cai et al., 2026).

The connection between PPO-Clip and the trust-region constraint can be understood through the TV divergence, for each state s,

$$D_{\mathrm{TV}}(\pi(\cdot|s) \parallel \mu(\cdot|s)) = \frac{1}{2}\mathbb{E}_{a\sim\mu(\cdot|s)}\left[|r(s,a)-1|\right]. \tag{4}$$

Thus, $|r(s,a)-1|$ is a single-sample Monte Carlo estimator of $D_{\mathrm{TV}}(\pi(\cdot|s) \parallel \mu(\cdot|s))$ (Qi et al., 2026) (ignore the constant coefficiency). From this perspective, PPO-Clip imposes a per-sample bound (that is *proximal*) on a noisy single-sample proxy of the divergence.

While effective in practice, this proxy has two fundamental limitations. First, from a *geometric perspective*, the underlying TV geometry has zero curvature and symmetric penalties, providing no adaptive resistance to different magnitudes of policy shifts. Second, from an *estimation perspective*, PPO constrains only a single-sample Monte Carlo estimate rather than the full expectation defining the TV divergence, making the effective trust region sensitive to sampling noise. These limitations can manifest as premature clipping, distorted trust-region scaling, and exploration collapse (Yu et al., 2025; Cai et al., 2026; Qi et al., 2026).

3 Generalized Proximal Policy Optimization

The preceding analysis reveals that PPO is not merely a clipping heuristic, but a particular instantiation of trust-region policy optimization characterized by a specific divergence and its estimator. We introduce **Generalized Proximal Policy Optimization (GPPO)**, a principled framework that systematically generalizes proximal policy optimization along two complementary dimensions: the *divergence* being constrained and the *estimator* used to compute it.

3.1 Generalized f-Divergence Trust Region

We first generalize the trust-region constraint from Total Variation (TV) to arbitrary f-divergences. The key theoretical step is to establish a tight quantitative relationship between an arbitrary f-divergence and the quadratic TV divergence D_{TV}^2 appearing in the monotonic improvement bound. Specifically, Vajda's inequality provides the tightest possible lower bound of an f-divergence in terms of TV divergence:

Theorem 2 (Vajda's Tight Lower Bound)**.** *Let $D_f(P \parallel Q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right)dx$ be an f-divergence generated by a convex function f satisfying $f(1) = 0$. For $V = D_{\mathrm{TV}}(P \parallel Q) \in [0,1]$, we have*

$$D_f(P \parallel Q) \geq L_f(V) := \inf_{\beta\in[0,\,1-V]}\left[\beta f\left(1+\frac{V}{\beta}\right) + (1-\beta)f\left(1-\frac{V}{1-\beta}\right)\right]. \tag{5}$$

where $\beta \in [0, 1-V]$ is the auxiliary probability-mass parameter in the tight binary reduction.

Consequently, inverting the tight lower bound yields $D_{\mathrm{TV}}(P \parallel Q) \leq L_f^{-1}(D_f(P \parallel Q))$.
To establish a tractable quadratic trust-region bound, we seek the largest constant $C_f > 0$ such that $D_f(P \parallel Q) \geq C_f D_{\mathrm{TV}}(P \parallel Q)^2$ holds uniformly for all P, Q. Using Vajda's tight lower bound, this optimal constant is:

$$C_f^{\mathrm{Vajda}} := \inf_{0<V\leq 1}\frac{L_f(V)}{V^2}, \tag{6}$$

which establishes the universal quadratic lower bound [1]:

$$D_{\mathrm{TV}}(P \parallel Q)^2 \leq \frac{1}{C_f^{\mathrm{Vajda}}} D_f(P \parallel Q). \tag{7}$$

We can now generalize the TV trust-region constraint in Theorem 1 to an arbitrary f-divergence:

Theorem 3 (Generalized f-Divergence Trust-Region Performance Bound)**.** *Define* $D_f^{\max}(\mu \| \pi) := \max_s D_f(\pi(\cdot|s) \parallel \mu(\cdot|s))$. *Then, for any* f*-divergence with* $C_f^{\mathrm{Vajda}} > 0$*,*

$$\begin{aligned} \eta(\pi) - \eta(\mu) &\geq \frac{1}{1-\gamma}\mathcal{L}_\mu(\pi) - \frac{4\xi\gamma}{(1-\gamma)^2} D_{\mathrm{TV}}^{\max}(\pi \parallel \mu)^2 \\ &\geq \frac{1}{1-\gamma}\mathcal{L}_\mu(\pi) - \frac{4\xi\gamma}{C_f^{\mathrm{Vajda}}(1-\gamma)^2} D_f^{\max}(\pi \parallel \mu). \end{aligned} \tag{8}$$

Thus, constraining any f-divergence to a sufficiently small level set guarantees monotonic performance improvement, establishing a unified trust-region framework for TV, Pearson χ^2, KL, Hellinger, and other f-divergences. This naturally motivates the generalized trust-region optimization problem:

$$\begin{aligned} \max_{\pi} \quad & \mathcal{L}_\mu(\pi) \\ \text{s.t.} \quad & D_f^{\max}(\pi \parallel \mu) \leq \delta. \end{aligned} \tag{9}$$

From f-Divergence Trust Regions to Generalized Proximal Policy Optimization. Directly solving this generalized constrained optimization problem is computationally prohibitive for large-scale language model policy optimization. We therefore retain the first-order proximal paradigm of PPO, while replacing its fixed clipping rule with a clipped proximal region induced by the chosen f-divergence. Let $\widehat{D}_f$ denote an estimator of the underlying f-divergence. The corresponding proximal region is defined as the δ-sublevel set of $\widehat{D}_f$:

$$\widehat{\mathcal{R}}_f(\delta) = \left\{ r > 0 : \widehat{D}_f(r) \leq \delta \right\}. \tag{10}$$

This induces the objective of generalized proximal policy optimization:

$$\mathcal{J}_{\mathrm{GPPO}}(\pi) = \mathbb{E}_{s\sim\rho^\mu,\, a\sim\mu(\cdot|s)} \left[\min\left(r(s,a)\hat{A}(s,a), \mathrm{clip}_{\widehat{\mathcal{R}}_f(\delta)}(r(s,a))\hat{A}(s,a) \right) \right]. \tag{11}$$

Thus the policy ratio is clipped whenever it moves outside the *divergence-induced proximal region* $\widehat{\mathcal{R}}_f(\delta)$, thereby suppressing the gradient associated with the corresponding samples.

3.2 Divergence Geometry and Policy Update Dynamics

GPPO in Eq. 11 reduces generalized f-divergence trust regions to practical updates. We now characterize how the geometry of f affects the proximal region and policy update dynamics. Under our convention in Eq. 9, we constrain $D_f(\pi(\cdot|s) \parallel \mu(\cdot|s)) = \mathbb{E}_{a\sim\mu(\cdot|s)}\left[f(r(s,a))\right] \leq \delta$. Thus the generating function $f(r)$ determines the shape of the divergence level set and hence the admissible proximal region of policy-ratio, for which its first derivative $f'(r)$ characterizes the marginal divergence cost while its curvature $f''(r)$ governs how that cost grows with the magnitude of a finite ratio shift. This distinction becomes particularly critical beyond the local regime $r \approx 1$.

Local Fisher–Rao Equivalence and Finite-Shift Geometry. For smooth f-divergences, a Taylor expansion around the reference point $r = 1$ gives

$$f(r) = f(1) + f'(1)(r-1) + \frac{1}{2}f''(1)(r-1)^2 + o\big((r-1)^2\big). \tag{12}$$

Since $\mathbb{E}_{a\sim\mu}[r-1] = 0$ and $f(1) = 0$, the linear term vanishes after taking expectation, yielding

$$D_f(\pi \parallel \mu) = \frac{f''(1)}{2}\mathbb{E}_{a\sim\mu}\big[(r-1)^2\big] + o\big(\|r-1\|^2\big). \tag{13}$$

Hence, all twice-differentiable f-divergences with $f''(1) > 0$ share the same local Fisher–Rao geometry up to a scaling factor. Their distinction emerges in the finite-shift regime ($r \not\approx 1$), where higher-order and asymmetric terms determine how resistance varies with the magnitude and direction of the policy-ratio deviation.

[1] The detailed derivation is provided in Appendix.

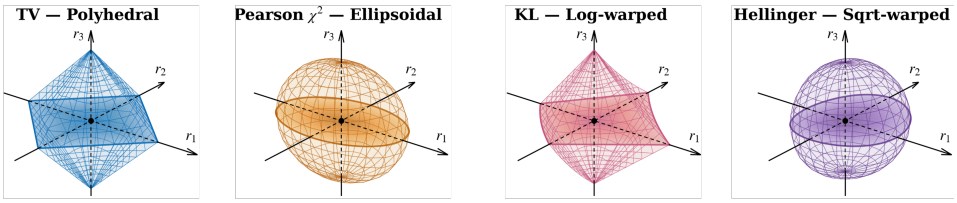

Figure 1: Schematic of four f-divergence geometries and the corresponding proximal regions they induce, shown for three sampled policy ratios $r = (r_1, r_2, r_3)$ centred at $r = 1$.

Zero Curvature of TV and Rigid Clipping. For TV divergence, the generating function gives

$$f_{\mathrm{TV}}(r) = \frac{1}{2}|r-1|, \qquad f'_{\mathrm{TV}}(r) = \frac{1}{2}\,\mathrm{sign}(r-1), \qquad f''_{\mathrm{TV}}(r) = 0 \quad (r \neq 1). \tag{14}$$

Thus, TV's polyhedral geometry is piecewise linear with zero curvature almost everywhere, yielding a marginal penalty independent of the magnitude of the ratio deviation $|r-1|$. PPO-Clip reflects this linear penalty structure through a fixed pointwise clipping threshold, beyond which the corresponding surrogate gradient is suppressed. Hence, the distinctive limitation of the TV-induced geometry is not the existence of a fixed clipping boundary itself, but the absence of curvature that would otherwise provide magnitude-dependent resistance. This rigidity provides a geometric explanation for the premature clipping and exploration-related issues discussed in Section 2.2 and explored by subsequent variants (Yu et al., 2025; Cai et al., 2026; Qi et al., 2026).

Adaptive Resistance of Smooth Divergences. In contrast, smooth f-divergences provide adaptive resistance through non-zero curvature. For Pearson χ^2 divergence,

$$f_{\chi^2}(r) = (r-1)^2, \qquad f'_{\chi^2}(r) = 2(r-1), \qquad f''_{\chi^2}(r) = 2, \tag{15}$$

induces an ellipsoidal (Mahalanobis) geometry where the penalty grows quadratically with the ratio deviation, yielding progressively stronger resistance as $|r-1|$ increases. For KL divergence,

$$f_{\mathrm{KL}}(r) = -\log r, \qquad f'_{\mathrm{KL}}(r) = -\frac{1}{r}, \qquad f''_{\mathrm{KL}}(r) = \frac{1}{r^2}, \tag{16}$$

introduces an asymmetric and logarithmically warped geometry where contraction $r \to 0$ is increasingly penalized, while expansion toward large r faces diminishing marginal resistance, thus facilitating probability expansion of low-probability actions and thereby favor exploration. For Hellinger,

$$f_{\mathrm{H}}(r) = (\sqrt{r}-1)^2, \qquad f'_{\mathrm{H}}(r) = 1 - \frac{1}{\sqrt{r}}, \qquad f''_{\mathrm{H}}(r) = \frac{1}{2r^{3/2}}. \tag{17}$$

It induces a spherical geometry under the square-root embedding whose curvature likewise increases toward $r \to 0$ and decreases as r grows, resulting in asymmetric resistance.

Fig. 1 visualizes the geometry of above four f-divergence. To conclude, *Divergence Geometry* determines the shape of the admissible policy-ratio region and its finite-shift resistance profile, thereby shaping where and how policy updates are constrained.

3.3 Divergence Estimator Topologies

The proximal region in Eq. 10 depends on how the underlying f-divergence is estimated. Beyond the classical *Single-Sample Monte Carlo* estimator, we propose two alternative topologies, *Atomic* and *Binary*, that more faithfully preserve the distribution-level divergence structure.

Single-Sample Monte Carlo Estimation. Single-Sample Monte Carlo estimation directly evaluates the divergence integrand at the sampled action:

$$\widehat{D}_f^{\mathrm{MC}}(r(s,a)) = f(r(s,a)), \qquad a \sim \mu(\cdot|s). \tag{18}$$

This induces the proximal region $\widehat{\mathcal{R}}_f^{\mathrm{MC}}(\delta) = \{r > 0 : f(r) \leq \delta\}$, suppressing sample gradients whenever $f(r(s,a)) > \delta$. Although unbiased, $\widehat{D}_f^{\mathrm{MC}}$ relies entirely on single-sample realizations, making its proximal clipping decisions highly susceptible to MC sampling variance.

Atomic Estimation. To mitigate the sensitivity of MC-estimated constraints to individual sampled outliers, we propose Atomic Estimation, which provides a probability-aware alternative. Recall that

the exact divergence expands over discrete actions as $D_f(\pi \parallel \mu) = \sum_a \mu(a|s) f(r(s,a))$. Atomic estimation imposes the proximal constraint directly on each atomic mass contribution:

$$\widehat{D}_f^{\text{Atom}}(r(s,a)) = \mu(a|s) f(r(s,a)), \qquad a \sim \mu(\cdot|s). \tag{19}$$

This induces the probability-aware proximal region $\widehat{\mathcal{R}}_f^{\text{Atom}}(\delta) = \{r > 0 : \mu(a|s) f(r) \leq \delta\}$. Crucially [2], Atomic estimation distinguishes the magnitude of a ratio deviation from its probability-weighted contribution: a large shift on a low-probability action does not necessarily constitute a large divergence contribution, thereby avoiding premature clipping of rare exploratory actions. Furthermore, we prove in Appendix that atomic estimation offers an improved bias-variance trade-off than MC, enabling stable and efficient training.

Binary Estimation. To capture distribution-level discrepancies beyond the randomly-sampled action, we consider Binary Estimation (Qi et al., 2026), which collapses the action space into the sampled action and its complement, yielding the binary f-divergence estimation:

$$\widehat{D}_f^{\text{Bin}}(r(s,a)) = \mu(a|s) f\left(r(s,a)\right) + (1-\mu(a|s))\; f\left(\frac{1-\mu(a|s) r(s,a)}{1-\mu(a|s)}\right), \qquad a \sim \mu(\cdot|s). \tag{20}$$

This induces the proximal region $\widehat{\mathcal{R}}_f^{\text{Bin}}(\delta) = \{r(s,a) : \widehat{D}_f^{\text{Bin}}(r(s,a)) \leq \delta\}$. Unlike MC and Atomic estimation, which evaluate only the sampled action contribution, Binary estimation accounts for the complementary probability mass and therefore captures the resulting distribution-level shift. [3].

To conclude, these estimators provide distinct ways of translating the underlying f-divergence geometry into a practical proximal constraint.

3.4 Unified View of Existing Proximal Policy Optimization Methods

GPPO provides a principled framework for interpreting existing proximal policy optimization methods within a two-dimensional design space of *Divergence* f × *Estimator* $\mathcal{T}$. Specifically,

PPO (Schulman et al., 2017), as previously stated, essentially *constrains the TV divergence under single-sample MC estimation*, follower by GRPO (Shao et al., 2024; Guo et al., 2025). GSPO (Zheng et al., 2025) and GMPO (Zhao et al., 2026) retain the TV divergence and MC estimator as PPO does, while changing the sampling unit from individual tokens to complete rollouts.

DAPO (Yu et al., 2025) adopts an asymmetric clipping mechanism:

$$\mathcal{J}_{\text{DAPO}}(\pi) = \mathbb{E}_{s,a}\left[\min\left(r(s,a)\hat{A}(s,a), \text{clip}\left(r(s,a), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\right)\hat{A}(s,a)\right)\right],$$

which corresponds to *constraining the KL divergence under single-sample MC estimation*. According to Eq. 18, $\widehat{D}_{\text{KL}}^{\text{MC}}(r(s,a)) = -\log r$, constraining $-\delta \leq \log r \leq \delta$ gives $\widehat{\mathcal{R}}_{\text{KL}}^{\text{MC}}(\delta) = \{e^{-\delta} \leq r \leq e^{\delta}\}$. Let $\delta = 0.23$, thus $0.8 \leq r \leq 1.26$, matching DAPO's asymmetric clipping bounds of ϵ_{low} and ϵ_{high}.

DPPO (Qi et al., 2026) constrains $|\pi(a|s) - \mu(a|s)| \leq \delta$, which corresponds exactly to *constraining TV divergence under binary estimation*. According to Eq. 20 we obtain,

$$\widehat{D}_{\text{TV}}^{\text{Bin}}(r(s,a)) = \frac{\mu(a|s)}{2}|r(s,a)-1| + \frac{(1-\mu(a|s))}{2}\left|\frac{1-\mu(a|s)r(s,a)}{1-\mu(a|s)} - 1\right| = |\pi(a|s) - \mu(a|s)| \leq \delta,$$

which yields the actual proximal region $\widehat{\mathcal{R}}_{\text{TV}}^{\text{Bin}}(\delta) = \{1 - \frac{\delta}{\mu(a|s)} \leq r(s,a) \leq 1 + \frac{\delta}{\mu(a|s)}\}$. Thus, DPPO can be equivalently written as:

$$\mathcal{J}_{\text{DPPO}}(\pi) = \mathbb{E}_{s,a}\left[\min\left(r(s,a)\hat{A}(s,a), \text{clip}\left(r(s,a), 1-\frac{\delta}{\mu(a|s)}, 1+\frac{\delta}{\mu(a|s)}\right)\hat{A}(s,a)\right)\right].$$

[2] Detailed illustration and complete theoretical analysis are provided in Appendix.

[3] We also explored Top-K estimation since both distributions of μ, π are known. However, it incurs prohibitively high memory overhead ($\mathcal{O}(K)$) and exhibits elevated variance across candidate actions, making the proximal threshold δ excessively difficult to tune. These empirical findings suggest that overly fine-grained estimation does not necessarily benefit, and can even compromise overall system robustness.

RIPO (Cai et al., 2026) guarantees isometric policy updates through the clipped objective:

$$\mathcal{J}_{\mathrm{RIPO}}(\pi) = \mathbb{E}_{s,a}\left[\min\left(r(s,a)\hat{A}(s,a), \mathrm{clip}\left(r(s,a), 1-\sqrt{\frac{\delta}{\mu(a|s)}}, 1+\sqrt{\frac{\delta}{\mu(a|s)}}\right)\hat{A}(s,a)\right)\right].$$

which essentially *constrains* χ^2 *divergence under atomic estimation*, since $\widehat{D}^{\mathrm{Atom}}_{\chi^2}(r(s,a)) = \mu(a|s)(r(s,a)-1)^2 \le \delta$ induces exactly the proximal region $\widehat{\mathcal{R}}^{\mathrm{Atom}}_{\chi^2}(\delta) = \{1-\sqrt{\frac{\delta}{\mu(a|s)}} \le r(s,a) \le 1+\sqrt{\frac{\delta}{\mu(a|s)}}\}$ used by RIPO.

In summary, existing proximal methods correspond to different $(f, \mathcal{T})$ choices, inducing distinct proximal regions for policy updates. More broadly, GPPO provides a systematic principle for designing new methods: f specifies the desired resistance profile, while $\mathcal{T}$ determines how the resulting divergence constraint is estimated and enforced from sampled data.

4 Experiments

To validate GPPO as a principled framework for proximal policy optimization and to characterize how the constrained divergence and its estimator shape policy update, we conduct extensive experiments across eight competition-level benchmarks spanning both mathematical reasoning and code generation, and benchmark instantiations drawn from the GPPO design space against nine representative RL algorithms.

4.1 Experimental Setup

Baselines. Since GPPO unifies proximal policy optimization along two complementary axes, namely, the constrained *divergence* f and its *estimator* $\mathcal{T}$, we instantiate it across four divergences (TV, Pearson χ^2, Hellinger, and KL) and three estimators (MC, Atomic, and Binary), forming a 4×3 design space. Within this space, several existing methods are recovered as special cases occupying specific $(f, \mathcal{T})$ cells, including GRPO (Shao et al., 2024) $(f_{\mathrm{TV}}, \mathcal{T}_{\mathrm{MC}})$, DAPO (Yu et al., 2025) $(f_{\mathrm{KL}}, \mathcal{T}_{\mathrm{MC}})$, DPPO (Qi et al., 2026) $(f_{\mathrm{TV}}, \mathcal{T}_{\mathrm{Bin}})$, and RIPO (Cai et al., 2026) $(f_{\chi^2}, \mathcal{T}_{\mathrm{Atom}})$.

Rather than tuning the trust-region radius δ directly, we calibrate it per divergence by matching the contraction-side clip of its MC proximal region $\widehat{\mathcal{R}}^{\mathrm{MC}}_{f}(\delta) = \{r > 0 : f(r) \le \delta\}$ to standard PPO-Clip, isolating divergence geometry from trust-region scale: fixing the lower bound to $r_{\min} = 0.8$ directly gives $\delta = f(0.8)$. Accordingly, $\delta_{\mathrm{TV}} = 0.2$, $\delta_{\chi^2} = 0.04$, $\delta_{\mathrm{H}} \approx 0.01$, and $\delta_{\mathrm{KL}} \approx 0.22$. While the upper bound $r_{\max}$ follows from $f(r_{\max}) = \delta$. The Atomic and Binary estimators directly reuse this δ, so that all estimators of a divergence share a common radius while their realized bounds adapt to the reference probability $\mu(a \mid s)$.

We further compare against other RL algorithms, including GSPO (Zheng et al., 2025) and GMPO (Zhao et al., 2026), which apply the TV–MC constraint at the full-rollout rather than per-token level; SAPO (Gao et al., 2025) and CISPO (Chen et al., 2025), which preserve otherwise-clipped gradients via soft gating and importance-weight clipping; and Clip-Cov (Cui et al., 2025b), which clips high-covariance tokens. All following their default hyper-parameter settings.

Following common practice (Yang et al., 2025b; Ye et al., 2020), we apply dual clipping to the importance ratio with lower and upper bounds of 0.5 and 10, and remove the KL penalty term for all methods (Yu et al., 2025; Yang et al., 2025b). All algorithms adopt the GRPO-style group-relative advantage estimator, and trained on Qwen3-8B-Base (Yang et al., 2025a).

Training. We evaluate the improvement induced by each RL algorithm on two complementary domains, mathematical reasoning and code generation, both of which admit verifiable rewards and are widely regarded as strong indicators of general reasoning ability. For mathematical reasoning we train on DAPO-Math-17k (Yu et al., 2025), and for code generation on Eurus-Code-25k (Cui et al., 2025a). For each query we generate 8 rollouts (namely, the group size is 8) with a maximum response length of 20k tokens for thorough reasoning. In each RL iteration, the behavior policy $\pi_{\theta_{\mathrm{old}}}$ totally generates 1024 rollouts with a train batch size of 128, and the current policy π_θ is updated 8 times with a mini-batch size of 16. Rewards are set to 1 for correct responses and 0 otherwise. We uniformly adopt token-mean loss aggregation and optimize all models to convergence using AdamW (Loshchilov & Hutter, 2017) with a constant learning rate of 1×10^{-6}, training for 400

Table 1: Results of the different divergences and estimators across various competition-level mathematical reasoning and code generation benchmarks (evaluated by Avg@8). We also compare the performance of other baselines. We highlight each cell as best, second best, and third best.

Mathematical Reasoning						
Divergence	Estimator	AIME24	AIME25	BRUMO25	SMT25	Avg.
TV	*MC* (GRPO)	34.6	20.8	33.3	21.9	27.7 (+0.00%)
	Binary (DPPO)	46.7	32.5	42.5	34.2	39.0 (+40.79%)
	Atomic	45.0	28.3	40.8	30.9	36.3 (+31.05%)
Hellinger	*MC*	37.9	30.8	35.4	23.1	31.8 (+14.80%)
	Binary	42.1	32.9	37.5	29.0	35.4 (+27.80%)
	Atomic	42.9	32.5	43.3	29.0	36.9 (+33.21%)
χ^2	*MC*	34.6	20.8	33.3	21.9	27.7 (+0.00%)
	Binary	45.0	29.2	36.7	28.1	34.8 (+25.63%)
	Atomic (RIPO)	47.9	32.5	47.5	34.2	40.5 (+46.21%)
KL	*MC* (DAPO)	45.0	27.5	37.5	28.5	34.6 (+24.91%)
	Binary	47.1	29.6	42.9	29.0	37.2 (+34.30%)
	Atomic	48.9	33.9	48.7	34.4	41.5 (+49.82%)
Other Baselines	GSPO	42.9	28.7	36.3	32.5	35.1 (+26.71%)
	GMPO	41.7	30.0	35.8	33.0	35.1 (+26.71%)
	SAPO	37.3	28.7	41.7	32.3	35.0 (+26.35%)
	CISPO	31.7	23.8	38.3	28.0	30.5 (+10.11%)
	Clip-Cov	36.3	22.9	36.6	30.2	31.5 (+13.72%)
	Base Model	3.8	3.3	15.0	10.8	8.2 (-70.40%)

Code Generation						
Divergence	Estimator	Codeforces	CodeContests	TACO	APPS	Avg.
TV	*MC* (GRPO)	61.9	57.4	36.2	54.2	52.4 (+0.00%)
	Binary (DPPO)	63.5	58.6	37.6	55.6	53.8 (+2.67%)
	Atomic	64.2	60.5	38.1	56.7	54.9 (+4.77%)
Hellinger	*MC*	66.4	58.2	36.3	56.8	54.4 (+3.82%)
	Binary	66.3	60.2	39.2	57.7	55.9 (+6.68%)
	Atomic	68.2	63.7	40.3	60.3	58.1 (+10.88%)
χ^2	*MC*	61.9	57.4	36.2	54.2	52.4 (+0.00%)
	Binary	64.8	59.3	38.6	56.3	54.8 (+4.58%)
	Atomic (RIPO)	68.2	62.3	39.3	58.9	57.2 (+9.16%)
KL	*MC* (DAPO)	64.0	59.1	37.4	56.8	54.3 (+3.63%)
	Binary	66.4	61.4	38.3	57.3	55.9 (+6.68%)
	Atomic	70.7	65.1	42.3	60.8	59.7 (+13.93%)
Other Baselines	GSPO	63.8	57.9	36.6	56.3	53.7 (+2.48%)
	GMPO	64.0	59.2	37.4	56.8	54.4 (+3.82%)
	SAPO	63.7	57.0	37.5	55.9	53.5 (+2.10%)
	CISPO	61.6	54.6	35.1	53.9	51.3 (-2.10%)
	Clip-Cov	62.9	57.4	36.0	55.5	53.0 (+1.15%)
	Base Model	32.3	32.1	16.5	32.1	28.3 (-45.99%)

steps on mathematical reasoning and 600 steps on code generation. All experiments are conducted on the VeRL (Sheng et al., 2025) framework using 8×H20 GPUs.

Evaluation. We comprehensively evaluate the effectiveness of these algorithms on eight competition-level benchmarks without contamination, spanning the two domains. For mathematical reasoning we use AIME24, AIME25, BRUMO25, and SMT25 (Balunović et al., 2025); for code generation we use Codeforces, CodeContests, TACO, and APPS (Cui et al., 2025a). We repeat the evaluation set for 8 times and report avg@8 metric for results stability.

4.2 Main Results

Table 1 summarizes the main experimental results across eight competition-level benchmarks. Fig 2 visualized the training dynamics of different GPPO instantiations on math and coding.

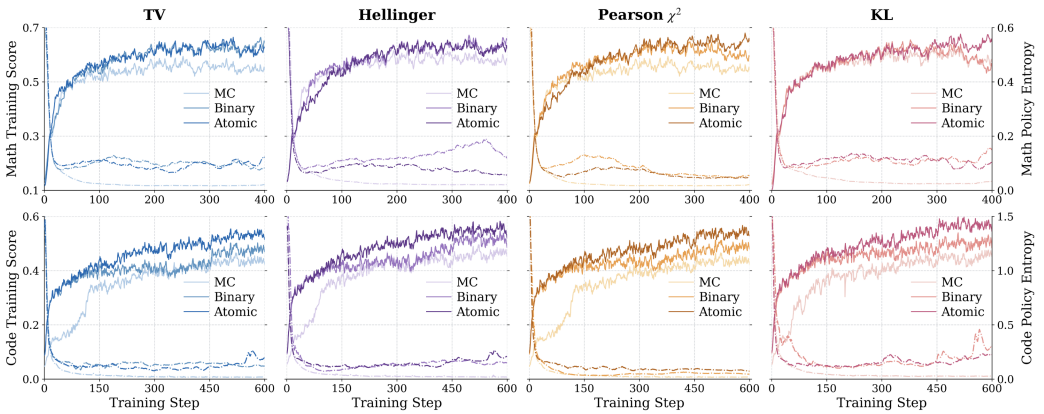

Figure 2: Training dynamics of different GPPO instantiations on Math and Code. Solid lines represent the training score, while dash-dotted lines represent the policy entropy.

GPPO establishes a new state-of-the-art across reasoning tasks. As highlighted in Table 1, our top-performing GPPO instantiation (KL divergence under Atomic estimation) achieves an average of 41.5 on mathematical reasoning and 59.7 on code generation, drastically improving upon the standard TV–MC baseline (GRPO) by $+49.8\%$ and $+13.9\%$, respectively, outperforming all existing RL algorithms. Beyond this optimal configuration, GPPO instantiations consistently dominate the design space, every divergence equipped with an Atomic or Binary estimator outperforms the strongest baseline (GMPO). These gains are uniform rather than dataset-specific, confirming that systematic exploration of the *Divergence* $\times$ *Estimator* interplay provides a fundamentally superior foundation for policy optimization than ad-hoc clipping heuristics.

Divergence geometry shapes policy optimization dynamics. Under single-sample MC, KL improves over TV by 24.9% on math and 3.6% on code, while Hellinger yields gains of 14.8% and 3.8%, both enhancing the entropy level of TV. TV–MC and χ^2–MC share the same results for that our matched calibration of single-sample MC coincidentally assigns them the same symmetric interval $r \in [0.8, 1.2]$, their geometric differences become contributing when estimated more faithfully with Atomic or Binary. Under Atomic estimation, KL surpasses TV by $+14.3\%$ and $+8.7\%$ on math and code. Binary reverses this on math but yields only modest gains for non-TV divergences on code, showing that divergence geometry is filtered through estimator topology rather than universally ranked. One possible explanation is that TV–Binary directly constrains the absolute mass change $|\pi(a \mid s) - \mu(a \mid s)| \leq \delta$, whereas curved divergences convert the same radius into a $\mu(a \mid s)$-dependent ratio penalty that can be tighter for rare actions. This could favor TV when useful mathematical updates require reallocating mass among a few low-probability reasoning branches, while coding's larger set of valid implementations may attenuate the difference.

Estimator topology determines how this geometry is enforced under sampling. Across four divergences, both proposed probability-aware estimators improve substantially over MC, while Atomic ranks first in every column of the code table and achieves the best math average. Averaged over the four divergences, Binary improves over MC by 21.3% on math and 3.2% on code, whereas Atomic yields larger gains of 28.3% and 7.7%, respectively. As Figure 2 shows, the entropy of MC collapse rapidly toward zero with the score curves plateau early at comparatively low levels. In contrast, Atomic and Binary maintain moderate range of entropy and continue improving over a longer horizon. This phenomenon follows directly from the estimator: Atomic avoids clipping large but low-probability exploratory updates by weighting with reference probability, and Binary instead accounts for the complementary probability mass.

5 Conclusion

This paper introduces GPPO, a principled framework that unifies proximal policy optimization through a two-dimensional design space of divergence geometry and estimator topology, supported by a generalized f-divergence trust-region theorem. Across eight mathematical reasoning and code generation benchmarks, GPPO substantially outperforms standard PPO, demonstrating that both dimensions are critical for effective LLM reinforcement learning. Ultimately, GPPO provides a systematic foundation for understanding and designing future LLM RL algorithms.

AI USE STATEMENT

In this work, we used generative AI tools for proofreading and grammatical refinement of the manuscript. We have not used generative AI tools for research ideation, theoretical derivation, experimental design, or code development, and automated decision-making or synthetic data generation tasks are not applicable to this work. Additionally, we used generative AI tools for minor stylistic and readability enhancements. We have reviewed all AI-assisted work. Specifically, all AI-suggested text modifications were manually reviewed and verified by the authors to ensure that the original research meaning and technical accuracy remained intact. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, the full source code and instructions for reproducing all experimental results are included in the supplementary material. The codebase includes environment setup files, training and evaluation scripts, and default hyperparameters. We will also open-source the codebase upon publication.

REFERENCES

Anthropic. Claude for coding. `https://www.anthropic.com/claude`, 2024.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmark*, 2025.

Zhicheng Cai, Xinyuan Guo, Hanlin Wu, Mingxuan Wang, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Beyond euclidean clipping: Overcoming exploration collapse in llm rl via riemannian isometric policy optimization. *International conference on machine learning*, 2026.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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

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

Shihan Dou, Yan Liu, Haoxiang Jia, Limao Xiong, Enyu Zhou, Wei Shen, Junjie Shan, Caishuang Huang, Xiao Wang, Xiaoran Fan, et al. Stepcoder: Improve code generation with reinforcement learning from compiler feedback. *arXiv preprint arXiv:2402.01391*, 2024.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

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

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *Proceedings of the Conference on Language Modeling (COLM)*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *International Conference on Learning Representations*, 2025.

OpenAI. Learning to reason with large language models. `https://openai.com/research/learning-to-reason-with-llms`, 2024.

OpenAI. Introducing deep research. `https://openai.com/index/introducing-deep-research/`, 2025.

Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the trust region in llm reinforcement learning. *International conference on machine learning*, 2026.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *International Conference on Learning Representations*, 2016.

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

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

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*, 2025a.

Shihui Yang, Chengfeng Dou, Peidong Guo, Kai Lu, Qiang Ju, Fei Deng, and Rihui Xin. Dcpo: Dynamic clipping policy optimization. *arXiv preprint arXiv:2509.02333*, 2025b.

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, et al. Mastering complex control in moba games with deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6672–6679, 2020.

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. *Proceedings of the Neural Information Processing Systems*, 2025.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, et al. Geometric-mean policy optimization. In *International Conference on Learning Representations*, volume 2026, pp. 41354–41374, 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.