Revisiting On-policy Adversarial Black-Box Distillation: Calibrating Groupwise Reward Geometry for Effective Advantage Construction

Xiao Cui[1] **Mo Zhu**[2*] **Yulei Qin**[3] **Wengang Zhou**[1] **Yuze Wu**[2] **Houqiang Li**[1]
[1] University of Science and Technology of China
[2] Zhejiang University [3] Independent Researcher
cuixiao2001@mail.ustc.edu.cn, {mozhu,wuyuze000}@zju.edu.cn
qinyulei@sjtu.edu.cn, {zhwg,lihq}@ustc.edu.cn

Abstract

Black-box distillation is a practical route for transferring capabilities from API-accessible large language models that expose only text outputs into smaller student models. Recent on-policy adversarial methods such as GAD improve over SeqKD by forming an adversarial loop between a critic and a student, where the critic provides rewards for GRPO-based student policy optimization over the student's sampled responses. However, GRPO computes advantages from the within-group relative rewards of student samples for the same prompt, whereas the critic is trained primarily to distinguish teacher responses from student responses. This objective mismatch can produce reward groups with collapsed scale or fragile margins, leading to brittle grouped optimization signals. We propose Groupwise Reward Geometry Conditioning (GRGC), a two-stage framework that improves advantage construction by shaping student-side reward groups during both critic training and policy optimization. To improve critic-side conditioning, Gaussian groupwise Optimal Transport calibration regularizes the critic during training to produce reward groups with non-collapsed spread and smooth rank-wise gaps by matching sorted prompt-wise rewards to group-centered Gaussian quantiles. Building on this conditioned reward geometry, policy-side group power modulation reshapes the prompt-wise reward groups before they are converted into advantages, preserving the critic-induced ordering while increasing optimization-relevant margin separability. Extensive experiments across diverse teachers, student model families and scales, and training datasets demonstrate the effectiveness of GRGC on both in-distribution and out-of-distribution evaluations, while introducing negligible overhead over GAD. The code is available at `https://github.com/2018cx/GRGC`.

1 Introduction

Large language models (LLMs) have achieved strong performance across NLP tasks [1, 2, 3], but their scale leads to high computational and deployment costs [4, 5, 6]. Knowledge distillation (KD) mitigates this issue by transferring capabilities from large teacher models to smaller students [7, 8, 9]. Most existing KD methods assume a *white-box* setting, where internal teacher signals such as logits and hidden states are accessible, enabling objectives based on distribution and representation alignment [10, 11, 12]. However, modern frontier LLMs are typically accessible only via text-generation APIs [13, 14, 15, 16], which restricts supervision to sampled outputs and defines the *black-*

*: Contributing equally with the first author.
Corresponding authors: Wengang Zhou and Houqiang Li.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

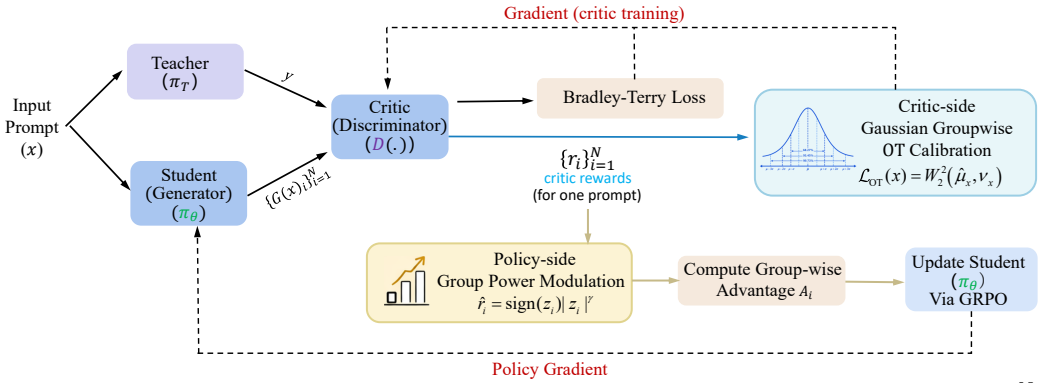

Figure 1: **Overview of GRGC.** For each prompt x, the student samples a response group $\{G(x)_i\}_{i=1}^N$, and the critic assigns raw sequence-level rewards $\{r_i\}_{i=1}^N$. CGC regularizes prompt-group reward geometry during critic training, while PGM transforms the conditioned rewards into $\{\hat{r}_i\}_{i=1}^N$ before the final prompt-wise advantages $\{A_i\}_{i=1}^N$ are computed for policy optimization.

box distillation setting. In this regime, standard white-box objectives become inapplicable, making distillation fundamentally more challenging and necessitating new learning paradigms [17, 18].

Early black-box distillation methods [19, 20, 21] typically adopt sequence-level knowledge distillation (SeqKD) [22], which treats teacher-generated responses as supervised targets for direct student training. SeqKD is simple and scalable, but it is fundamentally off-policy: the student is trained on teacher trajectories and then tested on its own rollouts, leading to exposure bias and a persistent train-test mismatch. This limitation has motivated recent moves toward on-policy black-box distillation [23, 18, 17]. Among these directions, on-policy adversarial distillation [18, 24] is particularly attractive because it replaces continual online teacher scoring with a learned critic, making training more cost-effective than approaches that repeatedly query the teacher during policy optimization. This practical advantage, however, exposes a sharper critic-to-advantage interface problem. The critic is trained to assign higher scores to teacher responses than to student responses, but GRPO updates the student from prompt-wise advantages computed only over the student's own sampled response group. Because the critic objective does not explicitly constrain the gaps, ordering robustness, or scale of this student-only reward group, these geometric properties can deteriorate even when teacher–student discrimination improves. Our analysis and propositions show that this deterioration directly affects advantage construction: low within-group dispersion makes the student-side ordering fragile under critic noise, while weak group scale amplifies noise sensitivity in the downstream advantage signal.

We call this groupwise reward geometry mismatch. To address it, we propose Groupwise Reward Geometry Conditioning (GRGC), which intervenes at two coupled stages: critic-side formation of student reward groups and their policy-side use in GRPO. Gaussian groupwise Optimal Transport (OT) calibration acts during critic training, at the source where critic scores form the raw student-side reward group. It uses an efficient OT-based objective to align the sorted rewards in each prompt group with group-centered Gaussian quantiles, chosen for their symmetric, moderate rank-gap profile, thereby inducing a non-degenerate spread, a stable scale, and a smooth ordered gap profile. This geometry is optimization-relevant: it makes within-group comparisons less prone to ranking flips and less sensitive to perturbations. Policy-side group modulation then acts on this conditioned reward geometry. After critic-side geometry calibration reduces collapse and scale drift, centering and scaling each reward group yields standardized scores with more reliable relative positions. In this regime, our analysis shows that the signed power transform preserves the critic-induced ordering and enlarges informative within-group margins under bounded perturbations, yielding more separable reward groups for GRPO-based student policy training. Our propositions further make this optimization link explicit: in a grouped softmax surrogate, advantage gaps exactly determine one-step pairwise log-odds shifts, while advantage errors induce bounded distortion in the clipped GRPO surrogate.

Our overall pipeline is illustrated in Figure 1, and our contributions are summarized as follows:

- We identify the groupwise reward-geometry mismatch in on-policy adversarial black-box distillation, where the critic can learn effective teacher–student discrimination while producing student-side reward groups that are poorly conditioned for GRPO advantage estimation.

- We propose GRGC, a lightweight framework for effective advantage construction: critic-side geometry calibration first stabilizes prompt-wise reward groups, and policy-side group modulation then amplifies informative margins on the conditioned signal.
- Experiments across diverse teachers, student model families and scales, and training datasets show consistent gains on in-distribution and out-of-distribution benchmarks, supported by different auto-judging models, evaluation metrics, and human preference evaluation.

2 Related Works

Teacher-based LLM distillation methods are broadly divided into white-box and black-box distillation.

2.1 White-box Knowledge Distillation of LLMs

White-box knowledge distillation leverages full access to the teacher's internal states, such as token-level probabilities and hidden representations, to guide student learning [25, 26]. Traditional methods mainly align output distributions with forward Kullback–Leibler (KL) divergence [10, 11], but some studies show that forward KL encourages mode-covering behavior, which can overburden limited-capacity students [11, 27]. To address this, MiniLLM [11] and GKD [12] use reverse KL divergence to promote mode-seeking behavior, ensuring the student focuses on the teacher's high-confidence reasoning paths. Furthermore, structural distillation through intermediate features, such as hidden states and attention maps in TinyBERT [28], has proven essential for deep architectural imitation. Beyond direct distribution matching, MMKD [29] formulates distillation as adversarial action-value moment matching. Recent studies have explored cross-tokenizer knowledge distillation to overcome conventional KD's reliance on identical teacher–student tokenizers and enable cross-family model transfer. ULD [30] addresses vocabulary misalignment by using the Wasserstein distance to align disparate token spaces. DSKD [31] unifies prediction spaces through bidirectional linear projections and an Exact Token Alignment algorithm. Complementary approaches like ALM [32] and DistillMoE [33] explore approximate likelihood matching and mixture-of-experts strategies to further stabilize cross-tokenizer knowledge transfer. However, white-box distillation is impractical for state-of-the-art models such as GPT-5 [13] and Doubao-Seed [34], whose internal states are typically inaccessible.

2.2 Black-box Knowledge Distillation of LLMs

Black-box distillation uses only observable teacher outputs rather than internal teacher signals. The dominant paradigm is SeqKD [22], which treats teacher responses as supervised labels and has supported many open-source instruction-tuned models [19, 20, 35, 36]. However, SeqKD is inherently off-policy and therefore vulnerable to exposure bias and distribution shift [37, 38]. To enrich supervision under this paradigm, later works have explored richer reasoning traces [39, 40, 41, 42, 21, 43]. Motivated by the success of white-box on-policy distillation [11, 12, 44], recent works have extended this idea to black-box distillation [23, 18]. OVD [23] obtains on-policy supervision by repeatedly querying the black-box teacher for verbal scores on student-generated trajectories. While effective, this teacher-in-the-loop design incurs substantial inference cost. GAD [18] instantiates adversarial distillation by introducing a critic model that is trained alternately with the student under a GAN-style minimax objective. The critic provides online rewards for student-generated responses, which are converted into prompt-wise advantages for GRPO-based policy optimization, while offline-generated teacher responses can be reused throughout training to avoid continual teacher querying. However, replacing online teacher supervision with an online critic shifts the key challenge to the critic-to-advantage interface. We address this bottleneck by calibrating groupwise reward geometry.

3 Methods

3.1 Problem Setup

We study black-box distillation from an API-accessible teacher policy π_T to a student policy π_θ. Given a prompt x, the teacher produces a response $y \sim \pi_T(\cdot \mid x)$. The objective is to optimize π_θ to approximate π_T using only teacher responses, without access to the teacher's internal signals.

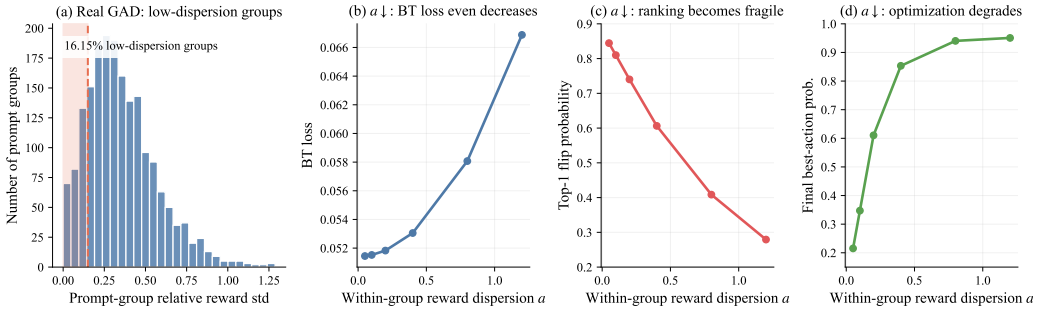

Figure 2: **Why BT-style adversarial critic training alone is insufficient for advantage construction.** (a) Real GAD training often yields low-dispersion prompt groups. The x-axis denotes the group reward standard deviation normalized by the global reward standard deviation at the same step. (b)–(d) Toy grouped-bandit analysis. We fix latent utility scores u_i to define the within-group preference order and set reward $r_i^{(a)} = a\,u_i$, so that varying a changes only the within-group reward dispersion. Smaller a means a more collapsed reward group. As a decreases, BT loss does not increase and may even decrease, while ranking becomes more noise-sensitive and grouped optimization deteriorates.

3.2 Revisiting On-policy Adversarial Black-Box Distillation

To reduce the risk of reward hacking against a fixed reward model, on-policy adversarial black-box distillation [18, 24] instantiates black-box distillation through an alternating student–critic training loop within a grouped policy optimization framework. For each prompt x, the student policy π_θ samples a response group $\{G(x)_i\}_{i=1}^N$, where N is the group size. The critic, serving as an online reward model, assigns sequence-level scores to both student responses and the teacher response y; these scores are used to form the critic's adversarial training objective. The student-side scores are then used as rewards to construct within-group advantages for GRPO-based optimization of π_θ.

GAD [18] trains the critic model with a Bradley–Terry-style objective:

$$\mathcal{L}_{\mathrm{BT}}(x) = \frac{1}{N}\sum_{i=1}^{N} -\log\sigma(D(y) - D(G(x)_i))\,, \tag{1}$$

where $D(\cdot)$ is the critic score and $\sigma(\cdot)$ is the logistic sigmoid. The core mismatch is especially relevant in adversarial black-box distillation, where the critic is updated online rather than calibrated as a general preference evaluator. Its BT objective enforces teacher–student discrimination, which can be achieved without within-prompt separability among student responses. GRPO, however, builds the student update from prompt-wise advantages derived from student-side rewards, making the update sensitive to their within-group reward geometry.

Proposition 1. Fix the teacher score $D(y)$ and let $\mu_x = \frac{1}{N}\sum_{i=1}^N D(G(x)_i)$ be the mean student score. Then strict convexity and Jensen's inequality [45] give

$$\mathcal{L}_{\mathrm{BT}}(x) \geq \log(1 + \exp(\mu_x - D(y)))\,, \tag{2}$$

with equality if and only if $D(G(x)_1) = \cdots = D(G(x)_N) = \mu_x$. Thus, on every fixed-mean slice, the BT objective uniquely favors zero within-group dispersion. The proof is given as Proposition 1 in Appendix C.

Figure 2 provides visual evidence for this mismatch. Figure 2(a) shows that real GAD training produces a non-negligible fraction of low-dispersion prompt groups, indicating that the emitted reward geometry can already be weak in practice. In a toy grouped-bandit construction, we adjust the within-group reward dispersion while leaving the teacher–student preference direction essentially unchanged. Figure 2(b) shows that this deterioration in geometry does not worsen the BT loss and can even reduce it,

Table 1: Low-dispersion incidence over the complete 1500-step Qwen2.5-3B/GPT-5/LMSYS trajectories. Each stage contains 64000 groups per method.

Stage	Rel. std < 0.15 GAD	Rel. std < 0.15 GRGC	Rel. std < 0.25 GAD	Rel. std < 0.25 GRGC
Early (500 steps)	8.57%	**1.28%**	15.73%	**1.28%**
Middle (500 steps)	8.69%	**1.90%**	16.04%	**1.90%**
Late (500 steps)	8.48%	**1.86%**	14.73%	**1.86%**

consistent with Proposition 1: the critic objective does not protect student-side dispersion and, at fixed mean, strictly favors collapse. However, once the reward group becomes too flat, the within-group ordering becomes much more fragile under noise, as shown in Figure 2(c). This increased fragility then directly degrades grouped optimization, as shown in Figure 2(d).

Table 1 extends Figure 2(a) to the complete run. GAD remains near $8.5\%/15\%$ at the $0.15/0.25$ thresholds throughout training, whereas GRGC never exceeds 1.90%. The equal GRGC columns are exact: every flagged group contains eight identical responses, while its minimum nonzero relative std is 0.8105. Thus, GAD's pathology persists beyond initialization or a selected interval.

Proposition 2. Let $\Delta_{\min}(r) = \min_{i \neq j, r_i \neq r_j} |r_i - r_j|$ denote the smallest nonzero group margin. Under bounded perturbations $|\eta_i| \leq \eta$, the ordering is guaranteed to be preserved if and only if $\Delta_{\min}(r) > 2\eta$. Thus low dispersion directly makes the grouped comparison structure fragile.

Proposition 3. Let $z = (z_i)_{i=1}^N$ be the prompt-wise normalized reward vector. For a perturbed group $\tilde{r} = r + \zeta$, let $\tilde{z}$ be its normalized vector. Then $\|\tilde{z} - z\|_2 = \mathcal{O}\left(\frac{\|\zeta\|_2}{\sigma_x + \varepsilon}\right)$. Weak group scale amplifies critic noise after grouped normalization and makes the downstream advantage signal more sensitive.

They correspond to Propositions 2 and 3 in Appendix C, where the full statements and proofs are given. Together with Figure 2(b), these facts explain why teacher–student discrimination alone is insufficient: the BT objective can remain favorable while the student-side reward geometry becomes too flat, noise-sensitive, and poorly conditioned for advantage construction.

3.3 Calibrating Groupwise Reward Geometry for Effective Advantage Construction

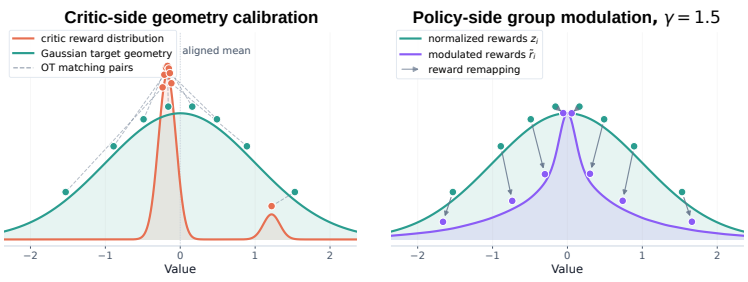

Figure 3: **Mechanism view of GRGC.** Critic-side Gaussian OT calibration aligns prompt-group rewards with mean-centered Gaussian quantiles, promoting non-collapsed scale and smooth rank-wise gaps. Policy-side group power modulation remaps group rewards to enlarge informative margins while preserving the critic-induced ordering.

The geometry mismatch above motivates an intervention at the critic-to-advantage interface. To address this, we explicitly condition prompt-wise reward group geometry for effective advantage construction. The preceding analysis shows that student-side critic rewards must avoid collapse, maintain stable within-group scale, and preserve sufficient separation for reliable advantage estimation. Additionally, prior studies [46] emphasize the importance of stable within-group variance for this estimation, both of which motivate the need for critic-side geometry calibration. While geometry calibration improves conditioning, it does not actively sharpen the informative margins needed for early grouped policy separation. To address this, we introduce policy-side group modulation, which amplifies useful grouped margins before deriving the final advantages. GRGC, therefore, consists of two complementary components: **(1) Critic-side geometry calibration (CGC):** During critic training, an OT objective regularizes reward groups toward well-separated, non-collapsed, and variance-stable geometry. **(2) Policy-side group modulation (PGM):** During policy optimization, a group power transform reshapes critic rewards before the final prompt-wise advantages are computed for GRPO.

Gaussian Groupwise OT Calibration for Critic Training During critic training, we regularize each prompt-wise student reward group toward a reference geometry. The goal is to impose an optimization-friendly within-group structure: the group should avoid collapse, maintain a controlled spread, and provide smoothly varying rank-wise gaps. We use a Gaussian quantile template to provide balanced geometry within a symmetric location-scale family, preventing rigid spacing and excessive tail emphasis, and producing reward groups that are suitable for subsequent advantage construction.

For each prompt x, let $\hat{\mu}_x = \frac{1}{N}\sum_{i=1}^N \delta_{D(G(x)_i)}$ be the empirical student reward distribution. We introduce a Gaussian reference distribution $\nu_x = \mathcal{N}(\mu_x, 1)$, where $\mu_x = \frac{1}{N}\sum_{j=1}^N D(G(x)_j)$. By

centering the reference at the current group mean, the calibration constrains only the centered shape and scale of the group. The critic-side calibration objective is defined as the squared 2-Wasserstein distance between the empirical distribution and the reference:

$$\mathcal{L}_{\mathrm{OT}}(x) = W_2^2(\hat{\mu}_x, \nu_x). \tag{3}$$

In the one-dimensional case, the exact W_2^2 distance between a discrete $\hat{\mu}_x$ and a continuous ν_x theoretically involves an integral over the Gaussian quantile function. To ensure computational efficiency, we employ a discrete approximation of ν_x by sampling N equally weighted quantiles. Let $r_{(1)} \le r_{(2)} \le \cdots \le r_{(N)}$ be the sorted rewards, and define the target Gaussian quantiles as:

$$q_j = \Phi^{-1}\left(\frac{j-0.5}{N}\right), \qquad t_j = \mu_x + q_j, \quad j = 1, \ldots, N. \tag{4}$$

In practice, we compute the Optimal Transport loss between $\hat{\mu}_x$ and the equally weighted discrete reference measure $\nu_x^{(N)} = \frac{1}{N}\sum_{j=1}^{N} \delta_{t_j}$, which admits the following closed-form expression [47]:

$$\mathcal{L}_{\mathrm{OT}}(x) = W_2^2(\hat{\mu}_x, \nu_x^{(N)}) = \frac{1}{N}\sum_{j=1}^{N} \big(r_{(j)} - t_j\big)^2. \tag{5}$$

The group mean is the unique OT-optimal translation of this centered template. Specifically, Proposition 4.3 in Appendix C proves that the cost for a translated template $a + q_j$ decomposes as

$$\mathcal{C}_x(a) = \mathcal{C}_x(\mu_x) + (a - \mu_x)^2, \tag{6}$$

and, writing the group loss as a function of its reward vector, that $\mathcal{L}_{\mathrm{OT}}(r + c\mathbf{1}) = \mathcal{L}_{\mathrm{OT}}(r)$. CGC therefore constrains within-group geometry while preserving common-score translation symmetry. Accordingly, the critic training objective becomes:

$$\mathcal{L}_{\mathrm{critic}}(x) = \mathcal{L}_{\mathrm{BT}}(x) + \lambda_{\mathrm{OT}}\mathcal{L}_{\mathrm{OT}}(x). \tag{7}$$

Proposition 4. Let σ_x denote the empirical standard deviation of the centered reward group and σ_q the standard deviation of the centered Gaussian target quantiles. Then

$$\mathcal{L}_{\mathrm{OT}}(x) \ge (\sigma_x - \sigma_q)^2. \tag{8}$$

Hence any collapsed or severely under-dispersed prompt group necessarily incurs a large OT penalty.

Proposition 4 corresponds to Proposition 4.1 in Appendix C and shows that CGC penalizes collapse and scale mismatch. Proposition 4.2 shows that OT penalizes ordered shape mismatch even when variance is matched, Proposition 4.3 establishes optimal mean anchoring and translation invariance, and Proposition 10 shows that Gaussian quantiles induce a more suitable interior-to-tail gap profile than linear or heavier-tailed alternatives. These analyses show that CGC is not only a variance regularizer, but a group-centered conditioner of the full ordered reward geometry.

Group Power Transform for Policy Optimization Critic-side calibration encourages the critic to emit stable, non-collapsed, and structurally regular reward groups, providing a well-conditioned geometry for grouped optimization. Building on this calibrated geometry, we apply a group power transform before final prompt-wise advantage computation to convert useful within-group differences into sufficiently decisive advantage gaps while preserving the critic-induced ordering. By retaining relative magnitude information and enlarging informative margins with a controlled polynomial gain, this transform makes the resulting reward groups more responsive to grouped optimization.

For each prompt x, let $\{r_i\}_{i=1}^N$ denote the corresponding critic reward group. The signed power transform is symmetric around zero; therefore, centering the reward group provides a natural reference point from which positive and negative deviations can be amplified in a balanced manner:

$$z_i = \frac{r_i - \mu_x}{\sigma_x + \varepsilon}, \tag{9}$$

where μ_x and σ_x are the mean and standard deviation of $\{r_i\}_{i=1}^N$. We then apply the group transform:

$$\hat{r}_i = \mathrm{sign}(z_i)\,|z_i|^{\gamma}, \quad \gamma > 1. \tag{10}$$

Table 2: Automatic evaluation results for models distilled from GPT-5-Chat and trained on the LMSYS-Chat training set. We report the averaged Qwen2.5-72B evaluation score and win rate.

Model	Method	LMSYS		Dolly		SelfInst		Vicuna	
		Score	Win	Score	Win	Score	Win	Score	Win
GPT-5-Chat	Teacher	51.21	53.2%	49.57	49.4%	49.71	49.2%	50.27	55.0%
Qwen2.5-3B-Instruct	Before Distill.	45.79	11.9%	44.93	4.0%	46.56	12.8%	47.85	3.8%
	SeqKD	47.06	18.4%	45.62	7.2%	46.83	16.9%	48.25	20.0%
	GAD	48.44	25.1%	46.19	8.2%	47.33	16.9%	48.76	28.8%
	GRGC	**50.19**	**45.3%**	**47.40**	**26.6%**	**48.99**	**30.6%**	**50.48**	**38.8%**
Qwen2.5-1.5B-Instruct	Before Distill.	41.93	4.8%	39.78	0.6%	41.06	5.0%	43.24	0.0%
	SeqKD	45.43	18.8%	42.91	4.6%	44.34	11.2%	47.41	11.3%
	GAD	46.05	15.7%	42.97	7.6%	45.27	9.9%	47.78	12.5%
	GRGC	**47.36**	**26.1%**	**45.16**	**18.6%**	**47.42**	**20.2%**	**48.81**	**27.5%**
Llama-3.2-3B-Instruct	Before Distill.	44.46	13.8%	45.72	8.0%	47.25	16.5%	47.98	5.0%
	SeqKD	45.99	17.3%	47.19	11.2%	47.38	16.5%	48.40	22.5%
	GAD	46.51	15.9%	47.67	11.6%	48.34	21.5%	48.52	21.3%
	GRGC	**49.07**	**40.1%**	**48.91**	**33.4%**	**49.28**	**30.2%**	**49.57**	**26.3%**
Llama-3.2-1B-Instruct	Before Distill.	39.50	4.4%	42.04	2.8%	41.85	7.0%	47.00	3.8%
	SeqKD	43.04	11.9%	42.30	5.2%	43.26	10.3%	46.89	7.5%
	GAD	42.57	8.8%	42.48	5.0%	42.93	9.9%	47.08	11.3%
	GRGC	**44.68**	**18.2%**	**43.84**	**14.8%**	**44.82**	**12.8%**	**47.54**	**20.0%**

GRPO subsequently derives prompt-wise advantages from the transformed rewards $\{\hat{r}_i\}_{i=1}^N$:

$$A_i = \frac{\hat{r}_i - \mathrm{mean}\big(\{\hat{r}_j\}_{j=1}^N\big)}{\mathrm{std}\big(\{\hat{r}_j\}_{j=1}^N\big) + \varepsilon}. \tag{11}$$

Proposition 5. Let $\hat{r}_i = f_\gamma(z_i)$ as defined in Eq. (10). The transform f_γ is strictly monotone and therefore preserves the critic-implied ordering. For any same-sign pair with $\min(|z_i|, |z_j|) \geq \rho > 0$,

$$|\hat{r}_i - \hat{r}_j| \geq \gamma \rho^{\gamma-1} |z_i - z_j|. \tag{12}$$

Hence, already-informative within-group gaps are enlarged before GRPO processing.

Proposition 5 shows that PGM preserves ordering while enlarging already informative gaps. The proof is given in Appendix C. Proposition 11 in the appendix formalizes the structural requirements for a pre-advantage transformation: it should preserve order, treat both sides symmetrically, apply a scale-consistent gain rule, and introduce no extra shape parameters. Under these requirements, the signed power form is uniquely determined among common transforms. Proposition 6 in the appendix further shows that the enlarged margins remain valid under bounded noise in the informative regime.

From advantage gaps to policy separation. The two components above define the mechanism chain of GRGC: CGC improves the geometry of rewards emitted by the critic, and PGM turns that geometry into more separable transformed rewards. In the grouped softmax surrogate analyzed in Appendix C, let $\pi_i(\theta)$ and $\pi_j(\theta)$ denote the policy probabilities assigned to candidates i and j within the same prompt group. Proposition 7 shows that one update changes their pairwise log-odds by

$$\log \frac{\pi_i(\theta^+)}{\pi_j(\theta^+)} - \log \frac{\pi_i(\theta)}{\pi_j(\theta)} = \eta_{\mathrm{pg}}(A_i - A_j). \tag{13}$$

Thus, reward geometry determines transformed reward separability, which determines the advantage gaps driving one-step policy separation in the grouped softmax surrogate. Proposition 8 shows that misranked constructed advantages can make this movement fail to improve the desired pairwise preference, or even move in the wrong direction. Proposition 9 connects this mechanism to clipped GRPO: errors in the constructed advantages induce bounded distortion of the clipped surrogate, so improving advantage conditioning and separability matter for the actual policy objective.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate our method on two complementary datasets for assistant response generation, covering distinct data distributions. (1) **LMSYS-Chat**: We primarily use a subset of the LMSYS-Chat-

Table 3: Automatic evaluation results for models distilled from Doubao-Seed-2.0 and trained on the LMSYS-Chat training set. We report the averaged Qwen2.5-72B evaluation score and win rate.

Model	Method	LMSYS		Dolly		SelfInst		Vicuna	
		Score	Win	Score	Win	Score	Win	Score	Win
Doubao-Seed-2.0	Teacher	53.23	82.5%	53.14	75.8%	52.70	74.0%	53.20	93.8%
Qwen2.5-3B-Instruct	Before Distill.	45.79	11.9%	44.93	4.0%	46.56	12.8%	47.85	3.8%
	SeqKD	46.95	25.3%	45.96	9.6%	46.87	16.9%	47.81	15.0%
	GAD	48.16	42.2%	46.86	37.0%	47.78	38.0%	49.84	60.0%
	GRGC	**49.49**	**49.9%**	**47.92**	**48.0%**	**48.75**	**43.0%**	**51.23**	**68.8%**
Qwen2.5-1.5B-Instruct	Before Distill.	41.93	4.8%	39.78	0.6%	41.06	5.0%	43.24	0.0%
	SeqKD	44.65	17.7%	43.53	6.0%	43.70	12.8%	47.48	15.0%
	GAD	45.13	27.1%	43.23	24.8 %	46.15	27.3%	47.98	37.5%
	GRGC	**47.28**	**30.9%**	**46.40**	**29.2%**	**46.94**	**32.6%**	**48.74**	**45.0%**
Llama-3.2-3B-Instruct	Before Distill.	44.46	13.8%	45.72	8.0%	47.25	16.5%	47.98	5.0%
	SeqKD	46.68	23.6%	46.28	9.4%	47.35	17.8%	48.49	18.8%
	GAD	47.74	45.3%	47.17	46.0%	48.69	49.6%	49.89	58.8%
	GRGC	**48.97**	**56.4%**	**49.57**	**64.6%**	**50.01**	**62.4%**	**51.49**	**78.8%**

1M dataset [48] preprocessed by GAD [18], which represents large-scale, diverse, and spontaneous user-chatbot interactions. (2) **Dolly-Train**: To further evaluate performance on human-authored instruction data, we adopt the MiniLLM-processed Dolly training data released on Hugging Face [11].

Models. We evaluate our method under black-box distillation from two API-based teachers, **GPT-5** [13] and **Doubao-Seed-2.0** [34]. Teacher responses are generated through their official APIs, with thinking mode enabled for Doubao-Seed-2.0. As students, we use two small-model families: **Qwen2.5-Instruct** at 1.5B and 3B scales [49], and **Llama-3.2-Instruct** at 1B and 3B scales [50].

Baselines and Evaluation. We compare GRGC with offline-teacher-response black-box distillation baselines, i.e., methods that learn from pre-generated teacher outputs without additional online teacher queries: SeqKD [22] and GAD [18]. Following GAD, evaluation covers both in-distribution and OOD benchmarks: a 479-sample LMSYS held-out test set, DollyEval [51] with 500 samples, SelfInst [52] with 242 user-oriented instructions, and VicunaEval [20] with 80 complex open-ended queries. We use Qwen2.5-72B [49] as the main LLM-as-a-judge, with details in Appendix F. Additional GPT-OSS-120B-based [53] and human evaluations are reported in Appendix G.

Implementation Details. All main results are trained for a total of 2 epochs. For GAD and our approach, the process consists of 1 epoch of supervised warmup followed by 1 epoch of adversarial training. We use a global batch size of 128. Following GAD [18], the learning rate is set to 5×10^{-6} for SeqKD, and 1×10^{-6} for both stages in GAD and our method. The group size is $N = 8$, the KL weight is $\beta = 0.001$, and the training temperature is set to 0.8. The critic is initialized from the student model's parameters and extended with an additional head to serve as the reward model. The maximum context length is 2,048 tokens for prompts and 1,536 tokens for responses, with responses from the Doubao-Seed-2.0 teacher capped at 2,048 tokens. Our OT weight is $\lambda_{\text{OT}} = 0.01$, and the group power hyperparameter is $\gamma = 1.5$. Experiments are conducted on 8 NVIDIA A100 GPUs.

4.2 Results and Discussions

GPT-5-Teacher Results. Table 2 shows that GRGC achieves the best results across all reported students and test sets. This pattern appears in both averaged scores and reference-based win rates, indicating consistent gains across metrics and stronger preference relative to the reference answers. These results are consistent with our central claim that better reward geometry improves grouped optimization and adversarial black-box distillation. GPT-OSS-120B judging and human preference evaluation show the same trend as the Qwen2.5-72B evaluation. Appendix G reports these evaluations, along with judge-free IFEval, Math500, and long-form mathematical-reasoning tests, sensitivity analyses on λ_{OT} and γ, and token-length analysis. The confidence intervals in Tables 20 and 21 further demonstrate the robustness of our improvement.

Doubao-Seed-2.0 Teacher Results. With thinking mode enabled, Doubao-Seed-2.0 produces longer and more structured responses, making it a more challenging teacher to distill. Comparing the LMSYS-Chat results in Tables 2 and 3, Doubao-Seed-2.0 achieves higher teacher win rates than

Table 4: Automatic evaluation results for models distilled from Doubao-Seed-2.0 and trained on Dolly Train. We report the averaged Qwen2.5-72B evaluation score and win rate on the test datasets.

Model	Method	LMSYS		Dolly		SelfInst		Vicuna	
		Score	Win	Score	Win	Score	Win	Score	Win
Doubao-Seed-2.0	Teacher	53.23	82.5%	53.14	75.8%	52.70	74.0%	53.20	93.8%
Qwen2.5-3B-Instruct	Before Distill.	45.79	11.9%	44.93	4.0%	46.56	12.8%	47.85	3.8%
	SeqKD	46.65	31.9%	46.60	33.2%	46.63	28.9%	48.36	36.3%
	GAD	47.92	33.6%	46.91	35.2%	48.38	36.4%	50.16	55.0%
	GRGC	**48.93**	**45.1%**	**47.88**	**42.8%**	**49.14**	**43.4%**	**51.25**	**62.5%**
Qwen2.5-1.5B-Instruct	Before Distill.	41.93	4.8%	39.78	0.6%	41.06	5.0%	43.24	0.0%
	SeqKD	42.19	14.2%	43.30	19.6%	43.70	21.9%	45.50	15.0%
	GAD	44.51	19.4%	43.95	20.0%	45.83	30.6%	47.29	25.0%
	GRGC	**45.10**	**22.1%**	**44.82**	**29.2%**	**46.73**	**32.2%**	**48.30**	**33.8%**
Llama-3.2-3B-Instruct	Before Distill.	44.46	13.8%	45.72	8.0%	47.25	16.5%	47.98	5.0%
	SeqKD	45.43	28.6%	47.53	30.8%	47.96	33.1%	49.43	51.3%
	GAD	45.47	30.3%	49.04	41.4%	48.70	39.3%	50.15	56.3%
	GRGC	**47.95**	**45.5%**	**49.71**	**57.4%**	**49.86**	**56.6%**	**51.26**	**70.0%**

Table 5: Ablation study of models distilled from GPT-5-Chat and trained on the LMSYS-Chat across adversarial training epochs.

CGC	PGM	Adv. Epoch = 1		Adv. Epoch = 2	
		Qwen2.5-3B	Qwen2.5-1.5B	Qwen2.5-3B	Qwen2.5-1.5B
×	×	48.44	46.05	48.52	46.37
✓	×	47.52	46.40	49.60	47.38
×	✓	48.88	46.44	48.95	46.69
✓	✓	50.19	47.36	50.27	47.70

Table 6: Comparison with critic-side calibration alternatives.

Distribution	LMSYS	Others
Variance only	46.83	46.61
Linear	47.00	46.79
Laplace	46.52	46.32
Logistic	46.54	46.45
Gaussian	47.36	47.13

GPT-5-Chat. GRGC helps student models better approximate Doubao-Seed-2.0's response behavior, as reflected by improved reference-based win rates. Table 4 shows that GRGC maintains its advantage when trained on Dolly-Train, demonstrating generalization beyond the LMSYS dataset distribution.

Impact of the two components. Table 5 verifies the intended interaction between CGC and PGM. CGC alone gives strong results after two adversarial epochs, showing that critic-side geometry calibration improves optimization by preventing reward collapse and scale instability. Its 1-epoch gain is smaller, and on Qwen2.5-3B it is even below GAD, because early OT calibration mainly organizes the reward geometry rather than immediately making the top student candidates more separated. PGM addresses exactly this early-stage gap: after group centering and scaling, the signed power transform preserves the critic-induced order while enlarging informative standardized differences, producing more decisive advantage gaps for the policy update. Combining the two therefore gives both well-conditioned critic rewards from CGC and stronger early advantage separation from PGM.

Discussion of CGC. Table 6 shows that Gaussian CGC gives the strongest results among the tested conditioning priors. Variance-only calibration helps by reducing the most obvious collapse, but it does not constrain how gaps are allocated across ranks. Linear targets impose non-collapsed but overly rigid equal spacing, while Laplace and logistic targets introduce heavier tails that allocate too much separation to extreme-ranked samples. Gaussian OT gives the best trade-off by matching ordered rewards to a non-collapsed, symmetric, and smoothly varying reference geometry with moderate tail emphasis. Details of these alternatives and more mechanism analysis are provided in Appendix C.

Discussion of PGM. Table 7 shows that group power is the strongest reward-shaping choice among the tested alternatives. Its advantage is that it preserves the critic-implied ordering while placing larger gain on already separated grouped margins. This keeps the critic's magnitude in-

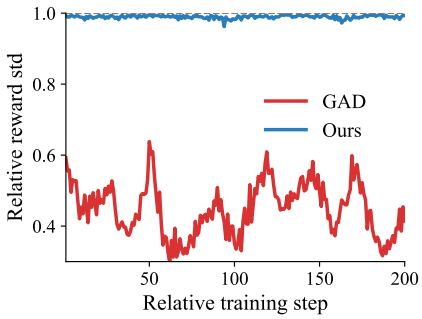

Figure 4: Mean relative prompt-group reward spread, computed as the mean group reward standard deviation normalized by the global reward standard deviation at each step. Our approach maintains a more stable reward scale.

Table 7: Ablation on the policy-side group modulation transformation.

Transformation	LMSYS	Others
CDF	46.63	46.74
Top-1	46.44	46.41
Rank	46.69	46.35
Group Power	47.36	47.13

Table 8: Average per-step training time, module-wise overhead, and combined overhead ratio across student model sizes.

Metric	Qwen2.5-1.5B-Instruct	Qwen2.5-3B-Instruct
Avg. step time	37.54 s	61.93 s
CGC overhead	0.03 s	0.03 s
PGM overhead	0.01 s	0.01 s
Overhead ratio	0.11%	0.06%

Table 9: Common-normalized final-advantage geometry over all 192,000 groups in the complete Qwen2.5-3B/GPT-5/LMSYS trajectory.

Metric	GAD	GRGC	Change
Mean top-2 gap	0.569	**0.835**	+46.7%
Median top-2 gap	0.431	**0.674**	+56.4%
Weak ($\Delta_{1,2} \leq 0.1$)	17.12%	**13.23%**	−22.7%
Top-1 (5% noise)	87.44%	**93.76%**	+6.32 pp
Pairwise (5% noise)	96.70%	**98.36%**	+1.66 pp
Top-1 (10% noise)	79.54%	**91.32%**	+11.78 pp

Table 10: Effect of doubling the prompt-batch or response-group size for Llama-3.2-1B. Low dispersion denotes relative std < 0.15.

Setting	Method	Low-disp.	LMSYS
Base	GAD	7.98%	42.57
Base	**GRGC**	**0.63%**	**44.68**
2× group size	GAD	6.83%	42.71
2× group size	**GRGC**	**0.03%**	**45.12**
2× batch size	GAD	7.92%	42.31
2× batch size	**GRGC**	**0.68%**	**44.40**

formation while avoiding the overly sparse update induced by Top-1 and the margin-information loss induced by CDF or Rank [54] remapping. More mechanism-level comparison is provided in Appendix C.

Runtime analysis. GRGC operates on the scalar reward groups already produced by the critic, adding only sorting, normalization, OT matching, and elementwise reward modulation without introducing extra model forward or backward passes. Table 8 confirms that this low-order overhead is negligible and largely insensitive to model size: the measured overhead is only 0.11% for Qwen2.5-1.5B and 0.06% for Qwen2.5-3B. A more detailed complexity analysis is provided in Appendix H.

Visualizations. The qualitative geometry diagnostics align closely with the quantitative gains. Fig. 4 shows that GRGC keeps the relative prompt-group reward spread within a more controlled range than GAD, suggesting improved critic-side conditioning for effective advantage construction. Figure 8 shows that GRGC reduces collapsed prompt groups, while Fig. 9 shows that it also reduces weak-margin groups. Figure 10 further reports the unnormalized prompt-group reward spread with uncertainty bands. These diagnostics support the proposed reward-geometry conditioning mechanism.

Direct advantage analysis. Table 9 reconstructs both methods' final advantages using their respective reward transformations under the same population z-score convention. The top-2 gap is the difference between the largest and second-largest advantages within each prompt group. GRGC yields larger gaps, fewer near ties, and stronger top-1 and pairwise retention, demonstrating better separation and perturbation stability of the final GRPO advantages. Table 10 further shows that additional sampling alone does not resolve low dispersion. Here batch size counts prompts per update, group size counts responses per prompt, and low dispersion denotes relative std < 0.15. Increasing either the number of prompts per update or the number of responses per prompt provides only limited improvement for GAD, whereas GRGC consistently maintains substantially lower low-dispersion frequency and higher LMSYS performance across all sampling settings.

5 Conclusion

We identify a previously overlooked critic-to-advantage mismatch in adversarial black-box distillation: the Bradley–Terry critic optimizes teacher–student discrimination, whereas GRPO depends on the within-group geometry of student rewards. This diagnosis yields a two-stage reward-geometry principle, instantiated by GRGC through source-mean-anchored groupwise OT during critic-side reward formation and order-preserving contrast reshaping before final advantage normalization. Across model pairs, datasets, and evaluation protocols, GRGC produces better separated and more perturbation-stable advantages, improves distillation quality, and adds negligible runtime overhead.

Acknowledgement This work was supported by the GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC, and the Supercomputing Center of the USTC.

Competing Interests The authors declare no competing interests.

References

[1] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, et al. Instruction tuning for large language models: A survey. *ACM Computing Surveys*, 58(7):1–36, 2026.

[2] Shangheng Du, Jiabao Zhao, Jinxin Shi, Zhentao Xie, Xin Jiang, Yanhong Bai, and Liang He. A survey on the optimization of large language model-based agents. *ACM Computing Surveys*, 58(9):1–37, 2026.

[3] Yulei Qin, Gang Li, Zongyi Li, Zihan Xu, Yuchen Shi, Zhekai Lin, Xiao Cui, Ke Li, and Xing Sun. Incentivizing reasoning for advanced instruction-following of large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:108337–108401, 2025.

[4] Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. Holistic evaluation of large language models for medical tasks with medhelm. *Nature Medicine*, 2026.

[5] Chenxu Niu, Wei Zhang, Jie Li, Yongjian Zhao, Tongyang Wang, Xi Wang, and Yong Chen. Tokenpowerbench: Benchmarking the power consumption of llm inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32582–32590, 2026.

[6] Luyang Fang, Xiaowei Yu, Jiazhang Cai, Yongkai Chen, Shushan Wu, Zhengliang Liu, Zhenyuan Yang, Haoran Lu, Xilin Gong, Yufang Liu, et al. Knowledge distillation and dataset distillation of large language models: Emerging trends, challenges, and future directions. *Artificial Intelligence Review*, 59(1):17, 2025.

[7] Mingjie Zhang, Xiaoling Zhou, Yuxiao Luo, Yiyu Liu, Shikun Zhang, and Wei Ye. Askd: Reinforcement learning-style knowledge distillation with quality-adaptive skewness. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pages 34781–34789, 2026.

[8] Huazheng Wang, Yongcheng Jing, Haifeng Sun, Jingyu Wang, Jianxin Liao, Leszek Rutkowski, and Dacheng Tao. Bridging the tokenizer gap: Semantics and distribution-aware knowledge transfer for unbiased cross-tokenizer distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pages 33494–33502, 2026.

[9] Hoang Tran Vuong, Tue Le, Quyen Tran, Linh Ngo Van, and Trung Le. Mcw-kd: Multi-cost wasserstein knowledge distillation for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33332–33340, 2026.

[10] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[11] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: On-Policy Distillation of Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[12] 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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[13] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card, 2025.

[14] 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.

[15] Xiao Cui, Qi Sun, Wengang Zhou, and Houqiang Li. Exploring gpt-4 vision for text-to-image synthesis evaluation. In *Tiny Papers@ ICLR*, 2024.

[16] Qi Sun, Xiao Cui, Wengang Zhou, and Houqiang Li. Exploiting gpt-4 vision for zero-shot point cloud understanding. *arXiv preprint arXiv:2401.07572*, 2024.

[17] Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. Lion: Adversarial distillation of proprietary large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3134–3154, 2023.

[18] Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025.

[19] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model, 2023.

[20] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

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

[22] Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[23] Jing Xiong, Hui Shen, Shansan Gong, Yuxin Cheng, Jianghan Shen, Chaofan Tao, Haochen Tan, Haoli Bai, Lifeng Shang, and Ngai Wong. Ovd: On-policy verbal distillation. *arXiv preprint arXiv:2601.21968*, 2026.

[24] Sudong Wang, Weiquan Huang, Xiaomin Yu, Zuhao Yang, Hehai Lin, Keming Wu, Chaojun Xiao, Chen Chen, Wenxuan Wang, Beier Zhu, et al. Beyond sft-to-rl: Pre-alignment via black-box on-policy distillation for multimodal rl. *arXiv preprint arXiv:2604.28123*, 2026.

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

[26] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge Distillation: A Survey. *International Journal of Computer Vision (IJCV)*, 2021.

[27] 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 International Conference on Computational Linguistics (COLING)*, pages 5737–5755, 2025.

[28] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Yeqiang Fang, and Qun Liu. TinyBERT: Distilling BERT for Natural Language Understanding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[29] Chen Jia. Adversarial moment-matching distillation of large language models. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 37:112184–112216, 2024.

[30] Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. Towards Cross-Tokenizer Distillation: the Universal Logit Distillation Loss for LLMs. In *arXiv preprint arXiv:2402.12030*, 2024.

[31] Xue Zhang, Songming Zhang, Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. A dual-space framework for general knowledge distillation of large language models. *arXiv preprint arXiv:2504.11426*, 2025.

[32] Benjamin Minixhofer, Ivan Vulić, and Edoardo Maria Ponti. Universal Cross-Tokenizer Distillation via Approximate Likelihood Matching. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[33] Minh-Phuc Truong, Hai An Vu, Tu Vu, Nguyen Thi Ngoc Diep, Linh Ngo Van, and Trung Le. Distillmoe: Multi-faceted knowledge distillation for cross-tokenizer embedding models. *OpenReview submission to ICLR 2026*, 2025. https://openreview.net/forum?id=VIYNWGb3TL.

[34] ByteDance Seed Team. Seed2.0 model card: Towards intelligence frontier for real-world complexity. Official Model Card Document, 2026.

[35] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[36] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:55006–55021, 2023.

[37] Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*, 2023.

[38] Hendra Setiawan. Accurate Knowledge Distillation via n-best Reranking. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

[39] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive Learning from Complex Explanation Traces of GPT-4. *arXiv preprint arXiv:2306.02707*, 2023.

[40] Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[41] Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

[42] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.

[43] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[44] Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards Streamlined Distillation for Large Language Models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[45] Edward James McShane. Jensen's inequality. 1937.

[46] Hu Wang, Congbo Ma, Ian Reid, and Mohammad Yaqub. Kalman filter enhanced grpo for reinforcement learning-based language model reasoning. *arXiv preprint arXiv:2505.07527*, 2025.

[47] Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

[48] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. In *International Conference on Learning Representations (ICLR)*, 2024.

[49] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[50] 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.

[51] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. Available at https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

[52] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 13484–13508, 2023.

[53] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

[54] Kyuseong Choi, Dwaipayan Saha, Woojeong Kim, Anish Agarwal, and Raaz Dwivedi. Gopo: Policy optimization using ranked rewards. *arXiv preprint arXiv:2602.03876*, 2026.

[55] Cédric Villani. The wasserstein distances. pages 93–111, 2009.

[56] Jingwei Zhang, Tongliang Liu, and Dacheng Tao. An optimal transport analysis on generalization in deep learning. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 34(6):2842–2853, 2023.

[57] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pages 214–223, 2017.

[58] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. *Conference on Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[59] Artan Sheshmani, Yi-Zhuang You, Baturalp Buyukates, Amir Ziashahabi, and Salman Avestimehr. Renormalization group flow, optimal transport, and diffusion-based generative model. *Physical Review E*, 111(1):015304, 2025.

[60] Weilin Chen, Jie Qiao, Ruichu Cai, and Zhifeng Hao. On the role of entropy-based loss for learning causal structure with continuous optimization. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 36(1):1594–1608, 2025.

[61] Dylan Wheeler and Balasubramaniam Natarajan. Conceptual learning and causal reasoning for semantic communication. *IEEE Transactions on Cognitive Communications and Networking (TCCN)*, 2025.

[62] Tung Le, Khai Nguyen, Shanlin Sun, Nhat Ho, and Xiaohui Xie. Integrating efficient optimal transport and functional maps for unsupervised shape correspondence learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23188–23198, 2024.

[63] Yonghao Liu, Fausto Giunchiglia, Ximing Li, Lan Huang, Xiaoyue Feng, and Renchu Guan. Enhancing unsupervised graph few-shot learning via set functions and optimal transport. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 871–882, 2025.

[64] Eduardo Fernandes Montesuma, Adel El Habazi, and Fred Ngole Mboula. Unsupervised anomaly detection through mass repulsing optimal transport. *arXiv preprint arXiv:2502.12793*, 2025.

[65] Ali Baheri, Zahra Sharooei, and Chirayu Salgarkar. Wasserstein adaptive value estimation for actor-critic reinforcement learning. *arXiv preprint arXiv:2501.10605*, 2025.

[66] Pascal Klink, Carlo D'Eramo, Jan Peters, and Joni Pajarinen. On the benefit of optimal transport for curriculum reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024.

[67] David Millard and Ali Baheri. Can optimal transport improve federated inverse reinforcement learning? *arXiv preprint arXiv:2601.00309*, 2026.

[68] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Conference on Neural Information Processing Systems (NeurIPS)*, 26, 2013.

[69] Jiafei Lyu, Mengbei Yan, Zhongjian Qiao, Runze Liu, Xiaoteng Ma, Deheng Ye, Jing-Wen Yang, Zongqing Lu, and Xiu Li. Cross-domain offline policy adaptation with optimal transport and dataset constraint. In *International Conference on Learning Representations (ICLR)*, 2025.

[70] Zhichen Zeng, Boxin Du, Si Zhang, Yinglong Xia, Zhining Liu, and Hanghang Tong. Hierarchical multi-marginal optimal transport for network alignment. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 16660–16668, 2024.

[71] Okan Koç, Alexander Soen, Chao-Kai Chiang, and Masashi Sugiyama. Domain adaptation and entanglement: an optimal transport perspective. *arXiv preprint arXiv:2503.08155*, 2025.

[72] Xiao Cui, Yulei Qin, Mo Zhu, Wengang Zhou, Yu Zhu, Hongsheng Li, Ming-Hsuan Yang, and Houqiang Li. Optimal transport-based difficulty-aware contribution allocation for dataset distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2026.

[73] Xiao Cui, Yulei Qin, Wengang Zhou, Hongsheng Li, and Houqiang Li. Optimizing distributional geometry alignment with optimal transport for generative dataset distillation. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:121368–121408, 2025.

[74] Xiao Cui, Yulei Qin, Wengang Zhou, Hongsheng Li, and Houqiang Li. Optical: Leveraging optimal transport for contribution allocation in dataset distillation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15245–15254. IEEE, 2025.

[75] Xiao Cui, Yulei Qin, Mo Zhu, Wengang Zhou, Hongsheng Li, and Houqiang Li. Geometry-aware dataset condensation for diffusion model training. *arXiv preprint arXiv:2606.05883*, 2026.

[76] Xiao Cui, Yulei Qin, Yuting Gao, Enwei Zhang, Zihan Xu, Tong Wu, Ke Li, Xing Sun, Wengang Zhou, and Houqiang Li. Sinkhorn distance minimization for knowledge distillation. In *International Joint Conference on Language Resources and Evaluation and International Conference on Computational Linguistics (LREC-COLING)*, pages 14846–14858, 2024.

[77] Xiao Cui, Yulei Qin, Yuting Gao, Enwei Zhang, Zihan Xu, Tong Wu, Ke Li, Xing Sun, Wengang Zhou, and Houqiang Li. Sinkd: Sinkhorn distance minimization for knowledge distillation. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 36(7):11887–11901, 2025.

[78] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16296–16305, 2021.

[79] Rishabh Bhardwaj, Tushar Vaidya, and Soujanya Poria. KNOT: Knowledge distillation using optimal transport for solving NLP tasks. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 4801–4820, 2022.

[80] 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 (AAAI)*, pages 23724–23732, 2025.

[81] Xiao Cui, Mo Zhu, Yulei Qin, Binbin Lin, Wengang Zhou, Hongsheng Li, and Houqiang Li. Flexible multi-level optimal transport for universal cross-tokenizer knowledge distillation on llms and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2026.

[82] Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 278–287, 1999.

[83] Eric Wiewiora. Potential-based shaping and q-value initialization are equivalent. *Journal of Artificial Intelligence Research*, 19:205–208, 2003.

[84] Sam Devlin and Daniel Kudenko. Dynamic potential-based reward shaping. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 433–440, 2012.

[85] Hado van Hasselt, Arthur Guez, Matteo Hessel, Volodymyr Mnih, and David Silver. Learning values across many orders of magnitude. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 4287–4295, 2016.

[86] Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. RUDDER: Return decomposition for delayed rewards. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 13544–13555, 2019.

[87] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 4299–4307, 2017.

[88] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

[89] 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. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 27730–27744, 2022.

[90] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 10835–10866, 2023.

[91] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[92] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling, 2025.

[93] Yanjun Chen, Dawei Zhu, Yirong Sun, Xinghao Chen, Wei Zhang, and Xiaoyu Shen. The accuracy paradox in RLHF: When better reward models do not yield better language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2980–2989, 2024.

[94] Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in LLMs: Reward calibration in RLHF. *arXiv preprint arXiv:2410.09724*, 2024.

[95] Zeyu Huang, Zihan Qiu, Zili Wang, Edoardo M. Ponti, and Ivan Titov. Post-hoc reward calibration: A case study on length bias. *arXiv preprint arXiv:2409.17407*, 2024.

[96] Igor Melnyk, Youssef Mroueh, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jarret Ross. Distributional preference alignment of LLMs via optimal transport. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[97] 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.

[98] Hugging Face Open R1 Team. OpenR1-Math-220k. Hugging Face dataset, 2025.

[99] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

[100] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, 2024.

[101] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Appendix

A Overview

This appendix provides supplementary materials that further elaborate on the related work, mechanism analysis, implementation details, evaluation protocol, and empirical findings of GRGC. It includes the following sections:

- **Section B: More Related Work.** Detailed discussions on prior studies, with an emphasis on optimal transport and related reward-shaping perspectives.
- **Section C: Mechanism Analysis for OT Calibration and Group Power.** Formal analyses of the BT objective's fixed-mean collapse bias, the downstream effects of poor reward geometry, the scale, shape, and anchoring properties of Gaussian OT calibration, the signed power transform, and the effect of advantage errors on the clipped GRPO surrogate.
- **Section D: Symbol Description.** A complete summary of key mathematical notations, hyperparameters, and definitions referenced throughout the paper.
- **Section E: Training Pseudocode.** Step-by-step pseudocode for the GRGC training pipeline, including the warmup phase, critic-side OT calibration, group-power reward shaping, and GRPO-based student updates.
- **Section F: Automatic Evaluation Details.** Full descriptions of the evaluation prompts, normalized score definition, and win-rate calculation used in automatic evaluation.
- **Section G: Further Experimental Analyses.** Additional experiments and analyses beyond the main paper, including sensitivity analysis, experiments with step-by-step teacher prompting, human evaluation, visualization analyses, GPT-OSS evaluation, token-length analysis, rule-based IFEval, Math500 and long-form mathematical reasoning, and multi-seed evaluation robustness.
- **Section H: Complexity Analysis.** A detailed comparison of the training complexity of GAD and GRGC, including the costs of BT loss, GRPO normalization, OT calibration, and group power shaping.
- **Section I: Broader Impact.** Reflections on the broader societal, ethical, and practical implications of more effective black-box distillation.
- **Section J: Limitations.** Critical discussion of the limitations of our framework.

Together, these supplementary materials provide a transparent view of the method, support reproducibility, and offer additional evidence that complements the main paper.

B More Related Work

Optimal Transport Optimal Transport (OT) provides a principled geometric framework for comparing probability distributions by computing the minimal cost required to transform one distribution into another. Compared to divergences such as KL and Jensen–Shannon, OT remains meaningful even when supports do not overlap and thus yields a more faithful notion of distributional discrepancy [55, 56]. The resulting Wasserstein distance has been widely adopted in image generation [57, 58, 59], causal structure learning and reasoning [60, 61], unsupervised learning [62, 63, 64], and reinforcement learning [65, 66, 67]. Entropy-regularized OT, or Sinkhorn distance [68], further enables scalable OT-based methods in domain adaptation [69, 70, 71] and dataset distillation [72, 73, 74, 75].

In knowledge distillation, OT has been used to transfer richer structural information than pointwise KL matching [76, 77]. WCoRD aligns teacher and student representation distributions with Wasserstein contrastive objectives [78]; KNOT distills teacher label distributions for NLP tasks by minimizing OT cost in label space [79]; and recent LLM distillation methods use Wasserstein, OT, or cross-tokenizer distribution-matching objectives to handle vocabulary mismatch and multi-level token/sequence alignment [30, 80, 81, 32, 9]. These works use distribution matching as a teacher–student knowledge-transfer loss, typically aligning logits, label distributions, hidden representations, or token spaces. Our use of OT is different: GRGC does not align student outputs to teacher logits or representations.

Instead, it applies a one-dimensional groupwise OT regularizer to the adversarial critic's student-side reward group, so that the reward geometry consumed by GRPO has controlled spread, ordered shape, and stable pre-normalization conditioning.

Reward Shaping and Reward Calibration Reward shaping has long been used to improve reinforcement-learning optimization by modifying the reward signal while preserving or clarifying the target behavior. Potential-based reward shaping gives policy-invariance guarantees for MDPs [82], with later extensions connecting shaping to value initialization [83] and dynamic shaping functions [84]. Other classical methods address reward scale and credit assignment: PopArt normalizes value targets across changing reward magnitudes [85], while RUDDER redistributes delayed returns to reduce temporal credit-assignment difficulty [86]. In language-model alignment, preference-based reward learning trains scalar reward models from comparisons [87, 88, 89], with later work studying reward overoptimization [90], implicit reward modeling through DPO [91], and reward-model evaluation through RewardBench [92]. Recent calibration-oriented work further shows that reward-model accuracy and reward-model confidence biases can be misaligned with downstream policy quality [93, 94], proposes post-hoc correction of reward-model biases such as length bias [95], and introduces OT-based distributional preference alignment [96]. These methods mainly study environment-reward transformations, value-target normalization, delayed-return redistribution, fixed/implicit reward models, or distributional alignment between preferred and dispreferred samples. GRGC addresses a different interface: in adversarial black-box distillation, the critic is trained online as a teacher–student discriminator and its student-side scores are immediately consumed by grouped advantage construction. Our calibration therefore targets the prompt-wise geometry of student-only reward groups, including their spread, ordered shape, and margin separability, rather than global reward scale, reward-model accuracy, or preference-distribution dominance alone.

C Mechanism Analysis for OT Calibration and Group Power

Here we formalize the mechanism behind GRGC across critic training, reward geometry, advantage construction, and grouped policy optimization. To keep the notation aligned with the main method, we distinguish four quantities throughout this section: raw critic rewards r_i, intermediate standardized grouped scores z_i, transformed rewards $\hat{r}_i$, and final GRPO advantages A_i. The mechanism we analyze is therefore

$$\begin{aligned}&\text{BT critic objective} \longrightarrow \text{reward geometry of } r \longrightarrow \text{stability of } z \longrightarrow \text{separability of } \hat{r} \\ &\longrightarrow \text{quality of the final advantages } A \longrightarrow \text{policy separation and clipped-surrogate stability.}\end{aligned} \tag{14}$$

This section establishes the complete mechanism chain. We first show that the BT critic objective has an intrinsic fixed-mean bias toward collapsed student reward groups, and then quantify how low dispersion and weak scale make grouped signals fragile. We next show that Gaussian OT calibration controls scale and ordered shape while using the unique optimal translation anchor, after which the signed power transform enlarges already informative margins. Finally, we connect the resulting advantage gaps to one-step policy separation and bound the effect of advantage errors on the clipped GRPO surrogate.

Setup. Consider a prompt group with raw critic rewards $\mathcal{G}(x) = \{r_i\}_{i=1}^N$ and group mean

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} r_i. \tag{15}$$

Let the ordered centered rewards be

$$u_{(j)} = r_{(j)} - \mu_x, \qquad j = 1, \dots, N, \tag{16}$$

and denote their empirical standard deviation by

$$\sigma_x^2 = \frac{1}{N}\sum_{j=1}^{N} u_{(j)}^2. \tag{17}$$

The intermediate standardized grouped scores used by the power transform are

$$z_i(r) = \frac{r_i - \mu_x}{\sigma_x + \varepsilon}. \tag{18}$$

The transformed rewards are then

$$\hat{r}_i = f_\gamma(z_i) = \mathrm{sign}(z_i)|z_i|^\gamma, \tag{19}$$

and the final GRPO advantages constructed from $\hat{r}$ are

$$A_i(\hat{r}) = \frac{\hat{r}_i - \frac{1}{N}\sum_{j=1}^N \hat{r}_j}{\sqrt{\frac{1}{N}\sum_{j=1}^N \left(\hat{r}_j - \frac{1}{N}\sum_{\ell=1}^N \hat{r}_\ell\right)^2 + \varepsilon}}. \tag{20}$$

This makes three aspects of reward geometry immediately relevant:

- *ranking stability*: whether the within-group ordering is robust to perturbations;
- *signal sensitivity*: how strongly perturbations in raw rewards are transmitted into the constructed grouped signal.
- *margin separability*: whether informative candidates remain sufficiently separated after transformation.

Proposition 1: the BT objective has a fixed-mean group-collapse bias. Fix a prompt x, its teacher critic score $D(y)$, and a student group mean μ_x. For any student score vector $r = (r_1, \ldots, r_N)$ satisfying $\frac{1}{N}\sum_i r_i = \mu_x$, define its prompt-wise BT loss by

$$\mathcal{B}_y(r) = \frac{1}{N}\sum_{i=1}^N \log(1 + \exp(r_i - D(y)))\,. \tag{21}$$

Then

$$\mathcal{B}_y(r) \geq \log(1 + \exp(\mu_x - D(y)))\,, \tag{22}$$

with equality if and only if

$$r_1 = \cdots = r_N = \mu_x. \tag{23}$$

Thus, for fixed teacher score and fixed student group mean, the BT objective uniquely minimizes its loss at zero within-group dispersion.

Proof. Let $\phi_y(s) = \log(1 + \exp(s - D(y)))$. Its second derivative is

$$\phi_y''(s) = \frac{\exp(s - D(y))}{(1 + \exp(s - D(y)))^2} > 0, \tag{24}$$

so ϕ_y is strictly convex. Jensen's inequality gives

$$\frac{1}{N}\sum_{i=1}^N \phi_y(r_i) \geq \phi_y\left(\frac{1}{N}\sum_{i=1}^N r_i\right) = \phi_y(\mu_x). \tag{25}$$

Strict convexity makes equality possible only when all r_i are identical, and the mean constraint then forces their common value to be μ_x.

Interpretation. The mismatch is stronger than a missing regularizer: on every fixed-mean slice of student scores, the BT term strictly penalizes mean-preserving dispersion and uniquely favors collapse. This does not claim that unconstrained adversarial training must always collapse, because the teacher score and group mean also evolve. It establishes that the critic objective contains no countervailing within-group preference and instead exerts a structural pressure toward the failure mode diagnosed in Figure 2.

Proposition 2: low dispersion makes grouped ranking fragile. Assume perturbed rewards $\tilde{r}_i = r_i + \eta_i$ with $|\eta_i| \leq \eta$ for all i. Let

$$\Delta_{\min}(r) = \min_{i \neq j,\ r_i \neq r_j} |r_i - r_j| \tag{26}$$

be the smallest nonzero within-group margin, with the convention $\Delta_{\min}(r) = 0$ if all rewards are tied. If

$$\Delta_{\min}(r) > 2\eta, \tag{27}$$

then every originally strict pairwise comparison is preserved under perturbation. Exact ties are not covered by this guarantee and may be broken by perturbations. Equivalently, any change to an originally strict comparison must satisfy

$$\Delta_{\min}(r) \leq 2\eta. \tag{28}$$

Proof sketch. For any pair i, j with $r_i > r_j$,

$$\tilde{r}_i - \tilde{r}_j = (r_i - r_j) + (\eta_i - \eta_j) \geq (r_i - r_j) - 2\eta. \tag{29}$$

Hence if every nonzero pairwise gap exceeds 2η, the sign of every originally nonzero pairwise difference is preserved. Conversely, changing any originally strict pairwise order is possible only if at least one original nonzero gap is at most 2η.

Interpretation. Proposition 2 formalizes the most direct downstream consequence of low dispersion and weak margins. When a reward group becomes too flat, the smallest useful within-group gap becomes comparable to ordinary critic noise. The result is not merely a cosmetic change in summary statistics: the identity of the better student response can itself become unstable.

Proposition 3: low scale makes the standardized grouped signal more sensitive. Let $z(r) = (z_1(r), \ldots, z_N(r))^\top$ and $\tilde{z} = z(\tilde{r})$ with $\tilde{r} = r + \zeta$. Write $\tilde{\mu}_x = \frac{1}{N}\sum_i \tilde{r}_i$ and $\tilde{\sigma}_x^2 = \frac{1}{N}\sum_i (\tilde{r}_i - \tilde{\mu}_x)^2$ for the perturbed group mean and perturbed group scale. Assume the perturbation is local in the sense that

$$\|\zeta\|_2 \leq \frac{\sqrt{N}}{2}(\sigma_x + \varepsilon). \tag{30}$$

Then

$$\|\tilde{z} - z\|_2 \leq \frac{2\|\zeta\|_2}{\sigma_x + \varepsilon}\left(1 + \frac{\sigma_x}{\sigma_x + \varepsilon}\right) \leq \frac{4\|\zeta\|_2}{\sigma_x + \varepsilon}. \tag{31}$$

In particular, for fixed perturbation magnitude $\|\zeta\|_2$, the perturbation in the standardized grouped signal grows as the pre-normalization group scale σ_x becomes small.

Proof sketch. Write

$$z(r) = \frac{Pr}{\sigma_x + \varepsilon}, \qquad P = I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top. \tag{32}$$

Then

$$\tilde{z} - z = \frac{P(r + \zeta)}{\tilde{\sigma}_x + \varepsilon} - \frac{Pr}{\sigma_x + \varepsilon}. \tag{33}$$

Add and subtract $\frac{Pr}{\tilde{\sigma}_x + \varepsilon}$ to obtain

$$\tilde{z} - z = \frac{P\zeta}{\tilde{\sigma}_x + \varepsilon} + Pr\left(\frac{1}{\tilde{\sigma}_x + \varepsilon} - \frac{1}{\sigma_x + \varepsilon}\right). \tag{34}$$

The empirical standard deviation is $1/\sqrt{N}$-Lipschitz with respect to the Euclidean norm, so

$$|\tilde{\sigma}_x - \sigma_x| \leq \frac{\|\zeta\|_2}{\sqrt{N}}. \tag{35}$$

Under the local perturbation assumption, this implies

$$\tilde{\sigma}_x + \varepsilon \geq \frac{1}{2}(\sigma_x + \varepsilon). \tag{36}$$

Using $\|P\zeta\|_2 \leq \|\zeta\|_2$, $\|Pr\|_2 = \|r - \mu_x\mathbf{1}\|_2$, and

$$\left|\frac{1}{\tilde{\sigma}_x + \varepsilon} - \frac{1}{\sigma_x + \varepsilon}\right| \leq \frac{|\tilde{\sigma}_x - \sigma_x|}{(\sigma_x + \varepsilon)(\tilde{\sigma}_x + \varepsilon)}, \tag{37}$$

we obtain

$$\left|\frac{1}{\tilde{\sigma}_x + \varepsilon} - \frac{1}{\sigma_x + \varepsilon}\right| \leq \frac{2\|\zeta\|_2}{\sqrt{N}(\sigma_x + \varepsilon)^2}. \tag{38}$$

Therefore

$$\|\tilde{z} - z\|_2 \leq \frac{2\|\zeta\|_2}{\sigma_x + \varepsilon} + \frac{2\|r - \mu_x\mathbf{1}\|_2\,\|\zeta\|_2}{\sqrt{N}(\sigma_x + \varepsilon)^2}, \tag{39}$$

and the identity $\|r - \mu_x\mathbf{1}\|_2 = \sqrt{N}\sigma_x$ gives

$$\|\tilde{z} - z\|_2 \leq \frac{2\|\zeta\|_2}{\sigma_x + \varepsilon}\left(1 + \frac{\sigma_x}{\sigma_x + \varepsilon}\right) \leq \frac{4\|\zeta\|_2}{\sigma_x + \varepsilon}. \tag{40}$$

Interpretation. Proposition 3 formalizes why unstable or collapsed reward scale directly hurts grouped signal construction. The issue is not simply that some groups have lower variance than others. Rather, once the pre-normalization scale becomes too small, the same raw perturbation produces a much larger change in the standardized grouped signal z, which is exactly the quantity later consumed by the power transform.

Proposition 4.1: OT calibration lower-bounds scale mismatch. Define the centered Gaussian target quantiles

$$q_j = \Phi^{-1}\left(\frac{j-0.5}{N}\right), \qquad t_j = \mu_x + q_j, \tag{41}$$

and let $\nu_x^{(N)} = \frac{1}{N}\sum_{j=1}^{N}\delta_{t_j}$ be the discrete Gaussian-quantile reference used in training. The critic-side OT loss is the exact one-dimensional Wasserstein distance from the empirical reward distribution to this discrete reference:

$$\mathcal{L}_{\mathrm{OT}}(x) = W_2^2(\hat{\mu}_x, \nu_x^{(N)}) = \frac{1}{N}\sum_{j=1}^{N}(r_{(j)} - t_j)^2 = \frac{1}{N}\sum_{j=1}^{N}(u_{(j)} - q_j)^2. \tag{42}$$

Let

$$\sigma_q^2 = \frac{1}{N}\sum_{j=1}^{N} q_j^2. \tag{43}$$

Assume $N \geq 2$, so $\sigma_q > 0$. Then

$$\mathcal{L}_{\mathrm{OT}}(x) \geq (\sigma_x - \sigma_q)^2. \tag{44}$$

In particular, if the reward group collapses to a constant, so that $\sigma_x = 0$, then

$$\mathcal{L}_{\mathrm{OT}}(x) \geq \sigma_q^2 > 0. \tag{45}$$

Proof sketch. Expanding the centered form gives

$$\mathcal{L}_{\mathrm{OT}}(x) = \sigma_x^2 + \sigma_q^2 - 2\langle u, q\rangle_N, \qquad \langle u, q\rangle_N = \frac{1}{N}\sum_{j=1}^{N} u_{(j)}q_j. \tag{46}$$

By Cauchy–Schwarz,

$$\langle u, q\rangle_N \leq \sigma_x\sigma_q, \tag{47}$$

which yields the stated lower bound.

Interpretation. Proposition 4.1 shows that CGC does not merely encourage a larger average variance. It penalizes collapse against a full ordered target geometry with strictly positive spread. This matters because Proposition 3 depends on the actual scale of the reward group consumed by normalization. OT calibration therefore improves the pre-normalization geometry that determines the stability of the downstream grouped signal.

Proposition 4.2: OT calibration also penalizes ordered shape mismatch. Assume $\sigma_x > 0$ and $\sigma_q > 0$. Define the normalized alignment between the ordered centered reward group and the centered Gaussian target quantiles by

$$\rho_x = \left\langle \frac{u}{\sigma_x}, \frac{q}{\sigma_q} \right\rangle_N = \frac{1}{N}\sum_{j=1}^{N} \frac{u_{(j)}}{\sigma_x}\frac{q_j}{\sigma_q}, \tag{48}$$

Then

$$\mathcal{L}_{\mathrm{OT}}(x) = (\sigma_x - \sigma_q)^2 + 2\sigma_x\sigma_q(1 - \rho_x). \tag{49}$$

In particular, even when $\sigma_x = \sigma_q$,

$$\mathcal{L}_{\mathrm{OT}}(x) = 2\sigma_x^2(1 - \rho_x), \tag{50}$$

so the OT loss still penalizes deviations in ordered within-group geometry unless the reward group is perfectly aligned with the target quantile shape.

Proof sketch. From the expansion

$$\mathcal{L}_{\mathrm{OT}}(x) = \sigma_x^2 + \sigma_q^2 - 2\langle u, q\rangle_N, \tag{51}$$

we write

$$\langle u, q\rangle_N = \sigma_x \sigma_q \left\langle \frac{u}{\sigma_x}, \frac{q}{\sigma_q} \right\rangle_N = \sigma_x \sigma_q \rho_x. \tag{52}$$

Substituting this identity gives the stated decomposition.

Interpretation. Proposition 4.2 directly answers why CGC is more than variance control. The first term measures scale mismatch, but the second term measures ordered shape mismatch: how the empirical within-group gaps align with the target quantile progression across ranks. A variance-only regularizer can constrain the first term, but it leaves the second completely unconstrained. OT calibration therefore regularizes the full one-dimensional prompt-group geometry rather than only its overall spread.

Proposition 4.3: group-mean anchoring is OT-optimal and translation invariant. Let $q_1 \le \cdots \le q_N$ be any centered template with $\frac{1}{N}\sum_j q_j = 0$; the Gaussian quantiles used by CGC satisfy this condition. Among all translations of this template, define

$$\mathcal{C}_x(a) = \frac{1}{N}\sum_{j=1}^{N} \big(r_{(j)} - (a + q_j)\big)^2, \qquad a \in \mathbb{R}. \tag{53}$$

Then

$$\mathcal{C}_x(a) = \mathcal{C}_x(\mu_x) + (a - \mu_x)^2. \tag{54}$$

Consequently, $a^* = \mu_x$ is the unique minimizer. Moreover, writing $\mathcal{L}_{\mathrm{OT}}(r) = \mathcal{C}_x(\mu_x)$, for every common shift $c \in \mathbb{R}$,

$$\mathcal{L}_{\mathrm{OT}}(r + c\mathbf{1}) = \mathcal{L}_{\mathrm{OT}}(r). \tag{55}$$

Proof. Using $u_{(j)} = r_{(j)} - \mu_x$ and the centeredness of both u and q,

$$\mathcal{C}_x(a) = \frac{1}{N}\sum_{j=1}^{N} \big(u_{(j)} - q_j + \mu_x - a\big)^2 \tag{56}$$

$$= \frac{1}{N}\sum_{j=1}^{N} \big(u_{(j)} - q_j\big)^2 + (\mu_x - a)^2, \tag{57}$$

because $\frac{1}{N}\sum_j (u_{(j)} - q_j) = 0$. The decomposition and uniqueness follow immediately. For the invariance result, adding c shifts every order statistic and the group mean by c, so

$$\big(r_{(j)} + c\big) - (\mu_x + c + q_j) = r_{(j)} - (\mu_x + q_j) \tag{58}$$

for every j, leaving the OT loss unchanged.

Interpretation. Centering the reference at the current student group mean is therefore the unique least-cost translation of any prescribed centered geometry, rather than a heuristic anchor. It also makes CGC insensitive to an arbitrary common offset in critic scores. If the teacher and student critic scores are shifted together, the BT score differences are unchanged and Proposition 4.3 shows that the CGC term is unchanged as well; the combined critic objective can therefore regulate within-group geometry without imposing an absolute score origin.

Why Gaussian is a canonical reference rather than an arbitrary one. The Gaussian target is best understood as a canonical member of a broader ordered location-scale family. More generally, one could use any ordered location-scale family

$$\tilde{t}_j = \mu_x + s\,\psi_j, \qquad \psi_1 \le \cdots \le \psi_N. \tag{59}$$

The relevant question is which family best satisfies the structural requirements imposed by grouped advantage construction. We require four properties:

- non-degenerate spread, so collapse is penalized;
- symmetry, so neither tail is privileged a priori;
- smooth rank progression, so adjacent target gaps change gradually rather than abruptly;
- single-parameter scale control, so the overall conditioning strength is easy to tune.

These desiderata are necessary but not sufficient to identify a unique target: other symmetric location-scale families, including logistic quantiles, also provide closed-form, non-degenerate, single-scale templates. The key distinction is therefore the induced rank-gap profile. We use the Gaussian target because it provides a moderate interior-to-tail progression: less rigid than constant linear spacing, but less tail-aggressive than heavier-tailed alternatives such as Laplace or logistic. In this sense, the Gaussian target is used as a canonical conditioning prior for grouped optimization geometry, not as a claim about the true semantic distribution of teacher preference.

Proposition 5: signed power preserves ordering and enlarges already informative gaps. Let

$$z_i = z_i(r) = \frac{r_i - \mu_x}{\sigma_x + \varepsilon}, \qquad f_\gamma(z) = \operatorname{sign}(z)|z|^\gamma, \qquad \gamma > 1. \tag{60}$$

Then f_γ is strictly monotone on $\mathbb{R}$ and therefore preserves within-group ordering. Moreover, if z_i and z_j have the same sign and satisfy

$$\min(|z_i|, |z_j|) \geq \rho, \tag{61}$$

for some $\rho > 0$, then

$$|f_\gamma(z_i) - f_\gamma(z_j)| \geq \gamma \rho^{\gamma-1} |z_i - z_j|. \tag{62}$$

Proof sketch. For $z \neq 0$,

$$f'_\gamma(z) = \gamma |z|^{\gamma-1} > 0, \tag{63}$$

so the map is strictly monotone. If z_i and z_j share the same sign, the mean value theorem gives

$$|f_\gamma(z_i) - f_\gamma(z_j)| = f'_\gamma(\xi)|z_i - z_j| \tag{64}$$

for some ξ between z_i and z_j. The lower bound follows because $|\xi| \geq \rho$ on that interval.

Interpretation. Proposition 5 moves the power-transform analysis beyond simple monotonicity. Once critic-side calibration has produced a group whose standardized scores have already exited the near-tie regime, group power increases the usable separation of those scores for advantage construction. This is exactly the quantity grouped optimization needs: not arbitrary rescaling, but stronger separation among already informative candidates.

Proposition 6: bounded-noise margin preservation in the informative regime. Assume latent standardized scores $z_i^\star$ and observed scores $z_i = z_i^\star + \xi_i$ with $|\xi_i| \leq \eta$. Consider two candidates i, j such that

$$z_i^\star > z_j^\star, \qquad z_j^\star \geq 1 + \eta, \tag{65}$$

and define the latent margin

$$\Delta_{ij}^\star = z_i^\star - z_j^\star. \tag{66}$$

If

$$\Delta_{ij}^\star > 2\eta, \tag{67}$$

then the observed ordering is preserved and the transformed margin satisfies

$$|f_\gamma(z_i) - f_\gamma(z_j)| \geq \gamma(\Delta_{ij}^\star - 2\eta). \tag{68}$$

Proof sketch. As in Proposition 2,

$$z_i - z_j \geq \Delta_{ij}^\star - 2\eta > 0, \tag{69}$$

so the observed ordering is unchanged. The condition $z_j^\star \geq 1 + \eta$ implies $z_j \geq 1$ and therefore $z_i \geq z_j \geq 1$, so the interval between them lies entirely in a regime where $f'_\gamma(z) \geq \gamma$. Applying Proposition 5 gives the bound. The corresponding lower-tail case follows by applying the same argument to $-z_i$ and using the odd symmetry of f_γ.

Interpretation. This proposition clarifies the role of PGM. Once CGC has moved a prompt group into a regime where standardized margins are already informative, PGM preserves those rankings and enlarges the usable separation further. This makes the transformed reward signal $\hat{r}$ more decisive for final advantage construction; the resulting optimization benefit is then supported by the grouped-surrogate analysis below and the main experiments.

Proposition 7: advantage gaps control one-step policy separation in a grouped softmax surrogate. Consider a grouped softmax policy

$$\pi_i(\theta) = \frac{e^{\theta_i}}{\sum_{k=1}^{N} e^{\theta_k}}, \tag{70}$$

and a one-step grouped update driven by constructed advantages,

$$\theta_i^+ = \theta_i + \eta_{\mathrm{pg}} A_i, \qquad \eta_{\mathrm{pg}} > 0. \tag{71}$$

Then for every pair i, j,

$$\log \frac{\pi_i(\theta^+)}{\pi_j(\theta^+)} - \log \frac{\pi_i(\theta)}{\pi_j(\theta)} = \eta_{\mathrm{pg}}(A_i - A_j). \tag{72}$$

Proof sketch. For the softmax parameterization,

$$\log \frac{\pi_i(\theta)}{\pi_j(\theta)} = \theta_i - \theta_j. \tag{73}$$

Applying the update gives

$$\log \frac{\pi_i(\theta^+)}{\pi_j(\theta^+)} = (\theta_i + \eta_{\mathrm{pg}} A_i) - (\theta_j + \eta_{\mathrm{pg}} A_j), \tag{74}$$

and subtracting the original log-odds yields the claim.

Interpretation. Proposition 7 links final advantage quality to a policy-level quantity in the grouped softmax surrogate. Pairwise policy separation after one grouped update is controlled by the corresponding pairwise advantage gap. Larger and more reliable final advantage gaps therefore produce larger and more reliable one-step separation toward the candidate assigned higher advantage in this surrogate model.

Proposition 8: ranking errors in the constructed advantages can induce misaligned one-step movement. Let $i^\star$ denote the best candidate under the underlying latent preference in a prompt group. Under the same one-step grouped update, for any competitor j,

$$\begin{aligned} &\log \frac{\pi_{i^\star}(\theta^+)}{\pi_j(\theta^+)} - \log \frac{\pi_{i^\star}(\theta)}{\pi_j(\theta)} \\ &\qquad = \eta_{\mathrm{pg}}(A_{i^\star} - A_j). \end{aligned} \tag{75}$$

Hence:

- if $A_{i^\star} > A_j$, the update increases the policy odds of the best candidate against j;
- if $A_{i^\star} \le A_j$, the update fails to improve this pairwise preference and may move in the wrong direction.

Proof sketch. This is the special case of Proposition 7 obtained by setting $i = i^\star$. The sign of the one-step log-odds change is therefore exactly the sign of $A_{i^\star} - A_j$.

Interpretation. Proposition 8 links advantage quality to update direction in the grouped surrogate model. Once the constructed advantages become unstable or misranked, the grouped policy update itself can become misaligned. Combined with Propositions 1–7, this yields the mechanism-level chain

$$\begin{aligned} \text{poor reward geometry} &\implies \text{fragile or distorted advantages} \\ &\implies \text{weaker or wrong policy separation.} \end{aligned} \tag{76}$$

This is exactly the failure pattern visualized in Figure 2.

Proposition 9: advantage errors bound clipped GRPO surrogate distortion. Consider a fixed prompt group sampled from the old policy $\pi_{\theta_{\text{old}}}$, with positive old-policy probability for every sampled candidate. Assume $0 < \epsilon_{\text{clip}} < 1$. For a candidate policy π_θ, define the likelihood ratio

$$w_i(\theta) = \frac{\pi_\theta(G(x)_i \mid x)}{\pi_{\theta_{\text{old}}}(G(x)_i \mid x)} \tag{77}$$

and the clipped ratio

$$\bar{w}_i(\theta) = \text{clip}(w_i(\theta), 1 - \epsilon_{\text{clip}}, 1 + \epsilon_{\text{clip}}) . \tag{78}$$

For an advantage vector $A \in \mathbb{R}^N$, the clipped grouped surrogate is

$$\mathcal{J}_{\text{clip}}(\theta; A) = \frac{1}{N} \sum_{i=1}^{N} \min(w_i(\theta)A_i, \bar{w}_i(\theta)A_i) . \tag{79}$$

Assume the local update region satisfies $0 \le w_i(\theta) \le 1 + \kappa$ for all i and some $\kappa \ge \epsilon_{\text{clip}}$. Then for any two advantage vectors A and $\tilde{A}$,

$$\left|\mathcal{J}_{\text{clip}}(\theta; A) - \mathcal{J}_{\text{clip}}(\theta; \tilde{A})\right| \le \frac{1+\kappa}{N}\|A - \tilde{A}\|_1 \le \frac{1+\kappa}{\sqrt{N}}\|A - \tilde{A}\|_2. \tag{80}$$

Define the clean and distorted surrogate improvements by

$$\Delta_A(\theta) = \mathcal{J}_{\text{clip}}(\theta; A) - \mathcal{J}_{\text{clip}}(\theta_{\text{old}}; A), \qquad \Delta_{\tilde{A}}(\theta) = \mathcal{J}_{\text{clip}}(\theta; \tilde{A}) - \mathcal{J}_{\text{clip}}(\theta_{\text{old}}; \tilde{A}). \tag{81}$$

Consequently, if $\Delta_A(\theta) \ge m$ and the constructed advantages satisfy $\|A - \tilde{A}\|_2 \le \epsilon_A$, then $\Delta_{\tilde{A}}(\theta) > 0$ whenever

$$m > \frac{2(1+\kappa)}{\sqrt{N}}\epsilon_A. \tag{82}$$

If both advantage vectors are group-centered, i.e., $\sum_i A_i = \sum_i \tilde{A}_i = 0$, then the old-policy baseline terms vanish and the sufficient condition improves to

$$m > \frac{1+\kappa}{\sqrt{N}}\epsilon_A. \tag{83}$$

Proof sketch. For fixed θ, both $w_i(\theta)$ and $\bar{w}_i(\theta)$ are nonnegative constants with

$$0 \le w_i(\theta) \le 1 + \kappa, \qquad 0 \le \bar{w}_i(\theta) \le 1 + \epsilon_{\text{clip}} \le 1 + \kappa. \tag{84}$$

For any nonnegative constants $a, b \le 1 + \kappa$, the function

$$h(c) = \min(ac, bc) \tag{85}$$

is $(1 + \kappa)$-Lipschitz in c: for $c \ge 0$ it equals $\min(a, b)c$, and for $c < 0$ it equals $\max(a, b)c$, so its slope always lies in $[0, 1 + \kappa]$. Therefore,

$$\left|\min(w_iA_i, \bar{w}_iA_i) - \min(w_i\tilde{A}_i, \bar{w}_i\tilde{A}_i)\right| \le (1+\kappa)|A_i - \tilde{A}_i|. \tag{86}$$

Averaging over i gives the ℓ_1 bound, and Cauchy–Schwarz gives the ℓ_2 bound. For the preservation statement, apply this bound once to the candidate update and once to the old policy; the surrogate-improvement error is therefore at most $2(1 + \kappa)\epsilon_A/\sqrt{N}$. If A and $\tilde{A}$ are group-centered, then $w_i(\theta_{\text{old}}) = \bar{w}_i(\theta_{\text{old}}) = 1$ for all i, so

$$\mathcal{J}_{\text{clip}}(\theta_{\text{old}}; A) = \frac{1}{N}\sum_i A_i = 0, \qquad \mathcal{J}_{\text{clip}}(\theta_{\text{old}}; \tilde{A}) = \frac{1}{N}\sum_i \tilde{A}_i = 0. \tag{87}$$

Only the candidate-policy surrogate term needs to be controlled, giving the improved factor.

Interpretation. Proposition 9 gives an objective-level link between advantage quality and the actual clipped GRPO surrogate used for policy optimization. It shows that distorted advantages do not merely affect a ranking diagnostic: they perturb the clipped surrogate itself by an amount proportional to the advantage error. Because GRPO constructs group-centered advantages, the sharper preservation bound applies directly to our setting, and the same argument extends to minibatches by averaging over prompt groups. Thus, the preceding geometry results matter because CGC and PGM reduce the sources of advantage distortion that enter the real policy objective.

Why the two components are complementary. The two modules optimize successive links in the same mechanism. Proposition 1 identifies the BT objective's fixed-mean collapse bias, while Propositions 2 and 3 quantify the resulting ranking fragility and normalization sensitivity. CGC counteracts this source-side failure: Propositions 4.1–4.3 establish non-collapse, ordered-shape control, optimal anchoring, and common-shift invariance. PGM then acts on the conditioned grouped signal: Propositions 5–8 show that it preserves ordering, enlarges informative separation, and translates advantage gaps into correctly directed one-step policy movement. Proposition 9 further shows that advantage errors induce bounded distortion of the clipped GRPO surrogate. Together, these results formalize the full chain from critic-objective bias to reward geometry, advantage quality, and the policy objective; Figure 2 and the main ablations corroborate its optimization-level consequences.

C.1 Why a Gaussian Target Among Alternative Conditioning Priors

The OT construction in GRGC does not require the target geometry to be Gaussian in a literal semantic sense. More generally, one could choose any ordered reference quantiles

$$\tilde{t}_j = \mu_x + s\,\psi_j, \qquad j = 1, \dots, N, \tag{88}$$

where the template $\{\psi_j\}$ is centered, ordered, and non-degenerate. The question is therefore not whether other priors are possible, but why the Gaussian prior is a particularly suitable one for our setting.

Concrete target forms. To make the comparison explicit, let

$$p_j = \frac{j - 0.5}{N}, \qquad j = 1, \dots, N. \tag{89}$$

In our implementation, the alternative target families are written under a common nominal scale convention. For Gaussian, Laplace, and logistic targets this corresponds to unit population standard deviation, while the linear baseline is the bounded linearly spaced template used in the ablation. Therefore, the Linear variant should be read as an implemented bounded-spacing baseline, not as a unit-variance shape-matched prior. The concrete choices are

$$t_j^{\mathrm{lin}} = \mu_x + \ell_j, \qquad \ell_j \in \mathrm{linspace}(-1, 1), \tag{90}$$

$$t_j^{\mathrm{gau}} = \mu_x + \Phi^{-1}(p_j), \tag{91}$$

$$t_j^{\mathrm{lap}} = \mu_x + \begin{cases} b\log(2p_j), & p_j < 0.5, \\ -b\log(2(1-p_j)), & p_j \geq 0.5, \end{cases} \qquad b = 1/\sqrt{2}, \tag{92}$$

$$t_j^{\log} = \mu_x + s\log\frac{p_j}{1-p_j}, \qquad s = \sqrt{3}/\pi. \tag{93}$$

Thus, linear spacing uses a bounded constant-gap template, Gaussian uses standard-normal quantiles, and Laplace/logistic use heavier-tailed quantile families under the same nominal scale convention.

What properties are needed. The diagnosis in Section 3.2 suggests that the reference geometry should first satisfy several basic desiderata:

- *anti-collapse*: it must assign a nonzero cost to degenerate or nearly constant groups;
- *symmetry*: it should not a priori favor the top or bottom tail of a prompt group;
- *smooth rank progression*: adjacent quantiles should vary gradually rather than producing artificial hard edges;
- *fixed-scale control*: the target spread should be controlled without introducing additional shape hyperparameters.

Proposition 10: Gaussian induces an intermediate quantile-gap profile. For the purpose of comparing rank-gap profiles, we view each discrete template as the sampling of a continuous centered quantile function $Q_\psi(p)$ on $p \in (0, 1)$, and define its local gap profile by

$$g_\psi(p) = \frac{d}{dp} Q_\psi(p), \qquad p \in (0, 1). \tag{94}$$

Under the natural continuous interpolation $Q_{\mathrm{lin}}(p) = 2p - 1$ for the linear template used in GRGC,

$$g_{\mathrm{lin}}(p) = 2, \tag{95}$$

$$g_{\mathrm{gau}}(p) = \frac{1}{\phi(\Phi^{-1}(p))}, \tag{96}$$

$$g_{\mathrm{lap}}(p) = \begin{cases} \dfrac{b}{p}, & p < 0.5, \\ \dfrac{b}{1-p}, & p > 0.5, \end{cases} \qquad b = 1/\sqrt{2}, \tag{97}$$

$$g_{\log}(p) = \frac{s}{p(1-p)}, \qquad s = \sqrt{3}/\pi. \tag{98}$$

Hence linear spacing is constant across all ranks, while Laplace and logistic allocate increasingly large spacing to extreme ranks at rate $\Theta((1-p)^{-1})$ as $p \to 1$ (and symmetrically as $p \to 0$). By contrast, the Gaussian gap profile also increases toward the tails but more moderately, with

$$g_{\mathrm{gau}}(p) = \Theta\left(\frac{1}{(1-p)\sqrt{\log(1/(1-p))}}\right) \qquad \text{as } p \to 1. \tag{99}$$

Thus Gaussian induces a nonconstant and smooth rank progression, but is less tail-aggressive than heavier-tailed alternatives.

Proof sketch. The formulas follow by differentiating the corresponding centered quantile functions. The asymptotic statement for the Gaussian case uses the standard tail expansion $1 - p \asymp \phi(z)/z$ for $z = \Phi^{-1}(p) \to \infty$, which gives $\phi(\Phi^{-1}(p)) \asymp (1-p)\sqrt{\log(1/(1-p))}$ up to constants. Substituting this into $g_{\mathrm{gau}}(p) = 1/\phi(\Phi^{-1}(p))$ yields the stated tail order.

Interpretation. Proposition 10 makes precise why Gaussian is a better conditioning prior than either linear spacing or heavier-tailed alternatives. Linear targets are too rigid: they enforce the same gap everywhere and cannot distinguish central ambiguity from extreme ranks. Laplace and logistic are too tail-aggressive: they allocate an increasingly large share of the total separation budget to the extremes. Gaussian lies in the middle. It preserves nontrivial rank variation and smooth interior-to-tail progression without over-concentrating separation in the tails. This is exactly the geometry needed for conditioning rather than aggressive sharpening.

Why Gaussian fits these desiderata well. The Gaussian quantile template satisfies these basic requirements, but these requirements alone do not make it unique. Logistic and Laplace quantiles are also symmetric, non-degenerate, closed-form, and single-scale. The deciding factor is the rank-gap profile formalized in Proposition 10. Linear spacing is too flat because it assigns the same gap to every rank pair; Laplace and logistic are too tail-aggressive because their quantile gaps grow at rate $\Theta((1-p)^{-1})$ near the extremes. Gaussian quantiles change smoothly with rank while using a milder tail expansion, so they provide a balanced conditioning geometry: central candidates remain moderately separated, tails receive additional but not excessive spacing, and the template keeps a controlled scale.

Why simpler or sharper alternatives are less attractive. A uniform-quantile target would also prevent collapse, but it imposes constant spacing across the entire group. This creates a geometry with no distinction between central ambiguity and extreme ranks. In contrast, our diagnosis suggests that many groups contain several near-tied middle candidates together with a few more separated extremes; a Gaussian template accommodates this more naturally. Heavy-tailed priors such as Laplace or logistic are also possible and satisfy many of the same basic desiderata, but they allocate relatively more of the separation budget to extreme quantiles. Since our critic-side objective is meant to improve conditioning rather than aggressively enlarge extremes, the Gaussian prior is a better default because its tail progression sits between the overly rigid linear template and the heavier-tailed alternatives.

Why this is not generic moment regularization. The critic-side OT term constrains the *ordered geometry* of each prompt group, not just a few aggregate statistics. Simpler penalties such as variance regularization, moment matching, or entropy-style smoothing could encourage a broader spread on average, but they would still leave many rank configurations admissible. In contrast, the OT loss ties each ordered score $r_{(j)}$ to an ordered reference quantile t_j, so it simultaneously constrains collapse, tail imbalance, and the progression of gaps across ranks. This is the key distinction from generic reward normalization or scalar variance control: CGC regularizes the *full one-dimensional prompt-group shape* that will later be consumed by grouped optimization.

Summary of the Gaussian choice. The Gaussian target is therefore not used as a distributional assumption on semantic rewards. It is used as a group-centered ordered template whose gap profile sits between constant linear spacing and heavy-tailed alternatives. The important distinction is not that Gaussian is the only closed-form symmetric single-scale option, but that it provides the most suitable conditioning profile among the tested options: non-degenerate spread, symmetric treatment of ranks, smooth gap progression, and moderate rather than excessive tail expansion.

C.2 Why `group_power` Among Alternative Reward Transforms

Our implementation contains several reward-transform alternatives, including global transforms (`linear`, `linear_fixed`, `beta`, `beta_logit`, `temperature`), group-aware standardization-only transforms (`group_zscore`), stronger nonlinear transforms (`group_sinh`), sparse winner-focused transforms (`top1_margin_boost`), and rank-replacement transforms (`rank_gaussian`). The comparison is organized around the structural requirements of pre-GRPO advantage construction: the transform should be prompt-group aware, preserve critic ordering, retain magnitude information, and provide controlled gain to informative margins.

Concrete forms of the compared transforms. For clarity, the alternatives in Table 7 correspond to the following reward-shaping rules. The proposed transform is

$$z_i = \frac{r_i - \mu_x}{\sigma_x + \varepsilon}, \qquad \hat{r}_i^{\text{power}} = \text{sign}(z_i)\,|z_i|^{\gamma}. \tag{100}$$

The *Top-1* baseline corresponds to the code-level transform `top1_margin_boost`: if $i^\star = \arg\max_i r_i$ and $\Delta_x = r_{(N)} - r_{(N-1)}$ for increasingly sorted rewards, then only the winner is modified,

$$\hat{r}_i^{\text{top1}} = \begin{cases} r_i + \lambda_{\text{top1}}\Delta_x, & i = i^\star, \\ r_i, & \text{otherwise.} \end{cases} \tag{101}$$

The *Rank* baseline corresponds to the code-level transform `rank_gaussian`: responses are sorted within each prompt group and then reassigned Gaussian quantiles according to rank,

$$\hat{r}_{(j)}^{\text{rank}} = \tau\,\Phi^{-1}(p_j), \qquad j = 1, \ldots, N, \tag{102}$$

where τ is a scale constant. Finally, *CDF* denotes a quantile-remapping baseline. Let $\hat{F}_x$ be the empirical CDF of the standardized scores within prompt group x, and let F^{-1} be a fixed target inverse CDF, such as the inverse Gaussian CDF. The transform is

$$\hat{r}_i^{\text{cdf}} = F^{-1}\left(\hat{F}_x(z_i)\right). \tag{103}$$

This baseline preserves ordering and maps candidates to a smooth target quantile scale, but it relies mainly on empirical CDF position and therefore weakens the critic's original magnitude margins.

Global transforms are not prompt-group aware. The failures we target are groupwise: low dispersion, weak top-group margins, and unstable within-group scale. Global transforms apply the same scalar map to all samples regardless of prompt-group membership. They can rescale the critic output globally, but they do not directly repair a nearly tied group or strengthen margins relative to that group's own scale. As a result, they are mismatched to the grouped nature of the optimization problem.

Standardization alone is not enough. `group_zscore` removes within-group shift and scale, but it does not further separate ambiguous candidates. After critic-side OT calibration, this can still leave grouped margins too weak for efficient policy updates. In other words, `group_zscore` improves comparability but not decisiveness.

More aggressive transforms distort the signal more strongly. `group_sinh` is group-aware and monotone, but its tails grow exponentially, which makes it more sensitive to moderately large spurious standardized scores. `top1_margin_boost` is highly local: it only changes the group winner and tends to create a winner-take-all geometry rather than improving the whole reward group. CDF remapping and `rank_gaussian` both replace much of the original magnitude structure with quantile position; this preserves rank, but it weakens the critic's own margin information.

Why `group_power` is the right compromise. The transform

$$\hat{r}_i = \mathrm{sign}(z_i)|z_i|^\gamma \tag{104}$$

has four properties that are jointly important in our setting:

- it is *group-aware*, because it operates on per-group standardized scores;
- it is *order-preserving*, because it is strictly monotone in z_i;
- it is *margin-amplifying*, because deviations farther from the group center receive larger polynomial local gain;
- it is *less destructive* than rank replacement and typically less brittle than exponential tails.

Thus, among the available alternatives, `group_power` provides the most suitable compromise between preserving critic-implied ordering, sharpening already informative grouped margins, and avoiding the magnitude loss induced by quantile remapping or the overly sharp tails induced by stronger nonlinear transforms.

Proposition 11: signed power is the unique normalized homogeneous transform. The structural requirements above can be made precise. A pre-GRPO transform on standardized grouped scores should satisfy four conditions. First, it should be odd, $f(-z) = -f(z)$, because $z = 0$ is the group mean after standardization and positive and negative deviations should be treated symmetrically. Second, it should be order-preserving, because the critic-implied ranking should not be reversed during advantage construction. Third, it should be positively homogeneous of degree $\gamma > 1$,

$$f(cz) = c^\gamma f(z), \qquad c > 0, \tag{105}$$

so the transform applies one consistent magnitude-dependent gain rule without introducing additional shape parameters. Fourth, it should be normalized by $f(1) = 1$, which fixes the arbitrary output scale and leaves γ as the only shaping parameter.

Under these requirements, the transform is uniquely determined:

$$f(z) = \mathrm{sign}(z)|z|^\gamma. \tag{106}$$

Moreover, for $z \neq 0$,

$$f'(z) = \gamma|z|^{\gamma-1}, \tag{107}$$

so the local gain is magnitude-dependent and grows polynomially rather than exponentially.

Proof sketch. For $z > 0$, positive homogeneity and $f(1) = 1$ give

$$f(z) = f(z \cdot 1) = z^\gamma f(1) = z^\gamma. \tag{108}$$

For $z < 0$, odd symmetry gives

$$f(z) = -f(-z) = -(-z)^\gamma = -|z|^\gamma. \tag{109}$$

Thus $f(z) = \mathrm{sign}(z)|z|^\gamma$. Differentiating on $z \neq 0$ gives $f'(z) = \gamma|z|^{\gamma-1}$.

Interpretation. Proposition 11 explains why group power is not an arbitrary monotone map. If we require the transform to preserve critic ordering, treat positive and negative standardized deviations symmetrically, avoid extra shape parameters, and apply a scale-consistent power-law gain to informative margins, then the signed power form is forced. The compared alternatives violate at least one of these requirements: rank-Gaussian and CDF remapping replace magnitude information with quantile position, top-1 boosting is not a pointwise transform on the whole group, and $\sinh$ is not homogeneous and has exponentially growing local gain. This is why group power is the appropriate PGM transform for advantage construction.

Algorithm 1 GRGC: Groupwise Reward Geometry Conditioning

Input: Distillation data $\mathcal{T} = \{(x, y)\}$; student policy π_θ; critic D; OT weight λ_{OT}; power exponent γ; group size N
Output: Trained student policy π_θ

Warmup Stage
for each batch $(x, y) \sim \mathcal{T}$ **do**
  Update student π_θ on teacher responses y with cross-entropy loss
  Sample a prompt-wise student group $\{G(x)_i\}_{i=1}^N$ from π_θ
  Compute critic scores $D(y)$ and $\{r_i = D(G(x)_i)\}_{i=1}^N$
  Compute teacher-student BT loss $\mathcal{L}_{\mathrm{BT}}(x)$
  Construct unit-scale Gaussian target quantiles $\{t_j\}_{j=1}^N$ with group mean μ_x
  Compute OT loss $\mathcal{L}_{\mathrm{OT}}(x) = \frac{1}{N}\sum_{j=1}^N (r_{(j)} - t_j)^2$
  Update critic D with $\mathcal{L}_{\mathrm{critic}}(x) = \mathcal{L}_{\mathrm{BT}}(x) + \lambda_{\mathrm{OT}}\mathcal{L}_{\mathrm{OT}}(x)$
end for

GRGC Training Stage
repeat
  for each batch $(x, y) \sim \mathcal{T}$ **do**
    Sample a prompt-wise student group $\{G(x)_i\}_{i=1}^N$ from π_θ
    Compute critic scores $D(y)$ and $\{r_i = D(G(x)_i)\}_{i=1}^N$
    Compute teacher-student BT loss $\mathcal{L}_{\mathrm{BT}}(x)$
    Construct unit-scale Gaussian target quantiles $\{t_j\}_{j=1}^N$ with group mean μ_x
    Compute OT loss $\mathcal{L}_{\mathrm{OT}}(x) = \frac{1}{N}\sum_{j=1}^N (r_{(j)} - t_j)^2$
    Update critic D with $\mathcal{L}_{\mathrm{critic}}(x) = \mathcal{L}_{\mathrm{BT}}(x) + \lambda_{\mathrm{OT}}\mathcal{L}_{\mathrm{OT}}(x)$
    Standardize raw rewards within each prompt group to obtain $\{z_i\}_{i=1}^N$
    Apply group power shaping $\hat{r}_i = \mathrm{sign}(z_i)|z_i|^\gamma$
    Construct GRPO advantages $\{A_i\}_{i=1}^N$ from transformed rewards $\{\hat{r}_i\}_{i=1}^N$
    Update student π_θ with the GRPO objective using $\{A_i\}_{i=1}^N$
  end for
until convergence
return π_θ

How the two components interact. The two modules act on different failure modes and therefore should not be viewed as interchangeable reward shaping steps. CGC first reduces under-dispersion and scale inconsistency in the raw critic rewards $\{r_i\}$, which makes grouped normalization better conditioned but does not by itself impose a magnitude-dependent gain on already informative standardized gaps. PGM then acts on the standardized grouped structure that remains: it preserves ordering while enlarging non-near-zero gaps in the signal used to construct final advantages. Put differently, CGC improves the *input conditioning* of grouped optimization, whereas PGM improves the *separability* of the signal consumed after that conditioning step. This division of labor is why the full method is stronger than either component alone in Table 5.

D Symbol Description

To enhance clarity, a detailed description of mathematical symbols used in the present study is provided in Table 11.

E Training Pseudocode

Algorithm 1 summarizes the full GRGC training pipeline in the same spirit as the GAD pseudocode. We separate the warmup phase from the on-policy GRGC phase, and explicitly show where critic-side OT calibration and policy-side group power modulation are applied.

Table 11: Descriptions of symbols used in the main text and appendix.

Symbol	Definition
x	Input prompt.
y	Teacher response.
π_T	Teacher policy.
π_θ	Student policy.
θ	Student-policy parameters.
$G(x)_i$	i-th student response.
N	Prompt-group size.
$D(\cdot)$	Critic / discriminator score.
r_i	Raw student sequence reward.
$\hat{r}_i$	Transformed reward.
A_i	GRPO advantage.
μ_x	Group mean of rewards.
σ_x	Group standard deviation of rewards.
$r_{(j)}$	j-th ordered reward.
$\hat{\mu}_x$	Empirical reward distribution.
ν_x	Continuous Gaussian reference distribution.
$\nu_x^{(N)}$	Discrete Gaussian-quantile reference distribution.
$W_2^2(\hat{\mu}_x, \nu_x^{(N)})$	Squared 2-Wasserstein distance to the discrete reference.
t_j	j-th Gaussian target quantile.
p_j	Rank quantile level $(j-0.5)/N$.
$\Phi^{-1}(\cdot)$	Inverse standard normal CDF.
$\mathcal{L}_{\mathrm{BT}}(x)$	Groupwise BT loss.
$\mathcal{L}_{\mathrm{OT}}(x)$	Groupwise OT loss.
$\mathcal{L}_{\mathrm{critic}}(x)$	Groupwise critic loss.
λ_{OT}	OT loss weight.
z_i	Standardized group score.
ε	Numerical stability constant.
γ	Power exponent.
$f_\gamma(z)$	Signed power map.
u_i	Centered reward.
$u_{(j)}$	j-th ordered centered reward.
q_j	Centered Gaussian quantile.
σ_q	Target quantile standard deviation.
ρ	Minimum standardized magnitude in the PGM margin bound.
ρ_x	Empirical correlation between centered rewards and target quantiles.
$z_i^\star$	Latent standardized score.
ζ	Additive reward perturbation vector.
ξ_i	Additive score perturbation.
η	Perturbation bound.
η_{pg}	Step size in the grouped softmax surrogate update.
$\Delta_{ij}^\star$	Latent margin $z_i^\star - z_j^\star$.
ψ_j	Generic target-quantile template.
$\tilde{t}_j$	Generic target quantile.
s	Generic target scale.
F	Smooth CDF used by the CDF transform.
Δ_x	Top-1 reward gap used by the Top-1 transform.
λ_{top1}	Top-1 margin-boost coefficient.
τ	Scale of the rank-Gaussian transform.
$\langle u, q \rangle_N$	Empirical inner product.
B	Responses per minibatch.
G	Number of prompt groups in a minibatch.
L	Average response length.
$\mathcal{F}_\pi^{\mathrm{roll}}(L)$	Actor rollout cost.
$\mathcal{F}_\pi^{\mathrm{upd}}(L)$	Actor update cost.
$\mathcal{F}_D(L)$	Critic pass cost.
$\mathcal{M}_\pi$	Actor memory cost.
$\mathcal{M}_D$	Critic memory cost.

F Automatic Evaluation Details

We use greedy decoding and set the maximum response length to 1536 tokens, except for the Doubao-Seed-2.0 LMSYS setting where responses are capped at 2048 tokens as described in the implementation details. Prompts are constructed using the same prompt wrapper as MiniLLM [11] and GAD [18], which is illustrated in Figure 5 and Figure 6. For evaluation with Qwen2.5-72B feedback, we first use Qwen2.5-72B to generate a reference answer for each instruction using the prompt in Figure 5. We then present the student response and the corresponding Qwen2.5-72B reference answer to the judge model, which evaluates both responses using the prompt shown in Figure 6.

The judge assigns one scalar score to the student response and one scalar score to the reference response. The reported automatic evaluation score is computed as

$$\mathrm{Score} = \frac{\mathrm{StudentScore}}{\mathrm{StudentScore} + \mathrm{RefScore}}, \tag{110}$$

where $\mathrm{StudentScore}$ denotes the judge score assigned to the student output and $\mathrm{RefScore}$ denotes the judge score assigned to the paired reference answer. This normalized score lies in $(0, 1)$ and measures the quality of the student response relative to the reference answer. In the reported experimental metrics, we report $100 \times \mathrm{Score}$, so the automatic evaluation results are presented on a 0 to 100 scale.

In addition to the averaged score, we also report *win rate* when appropriate. For a given evaluation set, a sample is counted as a *win* if the judge assigns a strictly higher scalar score to the student response than to the reference response, a *loss* if the student score is strictly lower, and a *tie* otherwise. The win rate is then defined as

$$\mathrm{WinRate} = \frac{\#\{\mathrm{StudentScore} > \mathrm{RefScore}\}}{\#}\{\text{all evaluated samples}\}. \tag{111}$$

Thus, win rate measures the fraction of evaluation prompts on which the student is preferred to the reference answer under the same judge, while the averaged score reflects the relative quality margin after normalization.

```
Below is an instruction that describes a task.
Write a response that appropriately completes the request.

### Instruction:
{instruction}

### Response:
```

Figure 5: The prompt wrapper for training and evaluation.

```
We would like to request your feedback on the performance of two AI assistants in response
to the user instruction and input displayed above.
Please rate the helpfulness, relevance, accuracy, and level of detail of their responses. Each
assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better
overall performance.
Please first output a single line containing only two values indicating the scores for Assistant
1 and 2, respectively. The two scores are separated by a space.
In the subsequent line, please provide a comprehensive explanation of your evaluation,
avoiding any potential bias and ensuring that the order in which the responses were presented
does not affect your judgment.
```

Figure 6: Automatic evaluation prompt.

Table 12: Sensitivity to the critic-side OT weight λ_{OT}. Results are averaged Qwen2.5-72B evaluation scores. Very small λ_{OT} weakens geometry conditioning, while overly large λ_{OT} makes the critic too conservative. A broad middle range remains stable, and we use $\lambda_{\mathrm{OT}} = 0.01$ in the main experiments.

λ_{OT}	0.001	0.005	0.01	0.02	0.05	0.1
Qwen2.5-3B-Instruct	48.92	50.10	50.19	50.22	50.16	49.94
Qwen2.5-1.5B-Instruct	46.79	47.27	47.36	47.31	47.32	47.17

Table 13: Sensitivity to the power exponent γ. Results are averaged Qwen2.5-72B evaluation scores. Values too close to 1 reduce the shaping effect, whereas overly large values lead to overly sharp grouped rewards. A broad intermediate region is stable, and we use $\gamma = 1.5$ in the main experiments.

γ	1.0	1.2	1.4	1.5	1.6	2.0
Qwen2.5-3B-Instruct	47.52	50.20	50.15	50.19	50.14	49.72
Qwen2.5-1.5B-Instruct	46.40	47.29	47.34	47.36	47.30	47.02

G Further Experimental Analyses

G.1 Sensitivity Analysis

GRGC requires tuning only two method-specific hyperparameters beyond the underlying GAD training recipe: the critic-side OT weight λ_{OT} and the policy-side power exponent γ. The Gaussian reference uses unit-scale quantiles throughout our experiments. We therefore focus the sensitivity study on λ_{OT} and γ. As shown in Table 12 and Table 13, the method is stable across a reasonably broad range, which is important in practice because it means GRGC does not require delicate tuning to outperform baseline GAD.

Table 12 studies the OT weight λ_{OT}. When λ_{OT} is too small, the critic-side geometry constraint becomes weak and the method moves toward a lightly regularized GAD regime, so under-dispersion and unstable reward scale are only partially corrected. When λ_{OT} is too large, the critic becomes overly conservative and policy improvement slows down because the reward geometry is over-regularized. Between these two extremes, however, performance is very stable: values from roughly 0.005 to 0.05 give nearly identical results, and we therefore adopt $\lambda_{\mathrm{OT}} = 0.01$ as a simple default.

Table 13 studies the power exponent γ. When γ is too close to 1, the transform degenerates toward a near-identity mapping, so the GRPO-side shaping effect becomes weak and only part of the grouped-margin problem is recovered. When γ is too large, the transformed rewards become overly sharp and the grouped optimization signal becomes less robust. Again, a broad intermediate region remains stable, with values from about 1.2 to 1.6 producing very similar results. For convenience, we use $\gamma = 1.5$ throughout the paper.

Overall, these results show that GRGC is easy to tune in practice: only two intuitive parameters are tuned, both admit stable operating ranges rather than narrow optima, and the default choices $(\lambda_{\mathrm{OT}} = 0.01, \gamma = 1.5)$ work well across model scales. This makes the method convenient to deploy in practical black-box distillation settings.

G.2 Distillation with an Explicit Step-by-Step Teacher Instruction

We additionally evaluate our framework in a more challenging Doubao-Seed-2.0 teacher setting by adding an explicit "think step by step" instruction to the input prompt. Doubao-Seed-2.0 uses thinking mode in both this setting and the standard setting; the difference here is the additional prompt instruction. This modification encourages the teacher to produce longer and more structured responses, making the resulting supervision more demanding than standard concise instruction-following outputs. Table 14 summarizes the results. Compared with the standard Doubao-Seed-2.0 teacher responses in Table 4, the step-by-step teacher achieves higher win rates against the reference answers on LMSYS, Dolly, and SelfInst, increasing from $82.5\%/75.8\%/74.0\%$ to $83.1\%/82.4\%/86.4\%$, while remaining equally strong on Vicuna at 93.8%.

Table 14: Automatic evaluation results for models distilled from Doubao-Seed-2.0 with an explicit step-by-step instruction and trained on the Dolly training set. We report the averaged Qwen2.5-72B evaluation score on the test datasets.

Model	Method	LMSYS		Dolly		SelfInst		Vicuna	
		Score	Win	Score	Win	Score	Win	Score	Win
Doubao-Seed-2.0	Teacher	53.58	83.1%	53.47	82.4%	53.78	86.4%	53.63	93.8%
Qwen2.5-3B-Instruct	Before Distill.	45.79	11.9%	44.93	4.0%	46.56	12.8%	47.85	3.8%
	SeqKD	46.09	41.8%	46.03	48.8%	46.81	46.3%	48.39	66.3%
	GAD	47.85	51.4%	47.61	52.0%	48.24	51.7%	50.32	73.8%
	GRGC	**48.84**	**55.1%**	**48.98**	**61.4%**	**49.17**	**53.7%**	**51.85**	**82.5%**
Qwen2.5-1.5B-Instruct	Before Distill.	41.93	4.8%	39.78	0.6%	41.06	5.0%	43.24	0.0%
	SeqKD	39.91	19.2%	41.81	36.0%	42.07	29.3%	45.79	38.8%
	GAD	39.86	20.3%	43.06	31.0%	44.78	31.8%	47.54	41.3%
	GRGC	**43.51**	**35.3%**	**44.08**	**44.2%**	**45.88**	**39.3%**	**48.43**	**53.8%**

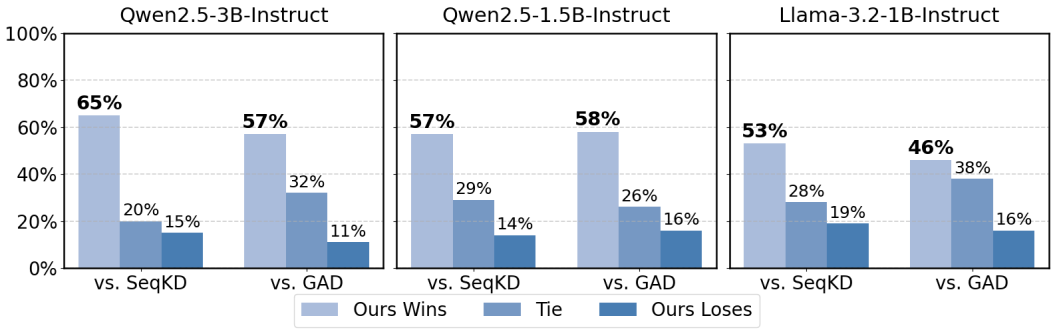

Figure 7: Human evaluation results on test sets. We compare GRGC to the models fine-tuned with SeqKD [22] and GAD [18].

The student results show that this stronger teacher behavior is also transferred more effectively. This indicates that when the teacher provides higher-quality and more structured answers, GRGC can better translate the richer supervision into stronger student responses by improving the grouped reward signal used for policy optimization.

Methodologically, this setting is useful because it enlarges the gap between simple sequence imitation and on-policy optimization from grouped rewards. When teacher answers become longer and structurally richer, merely copying surface form becomes less sufficient, and the quality of grouped reward signals becomes more important. The step-by-step teacher experiments therefore provide an additional stress test showing that GRGC remains effective when the teacher is encouraged to produce richer reasoning-style supervision.

G.3 Human Evaluation Results

Automatic metrics are useful but incomplete, so we additionally conduct a human study. We engage 10 professional volunteers for annotation. For each evaluated model, each volunteer judges 50 prompt-response groups, ensuring a balanced evaluation set that covers all four test sources used in the paper: 20 groups from the LMSYS test set, 10 from Dolly, 10 from SelfInst, and 10 from Vicuna. All annotators evaluate the same set of prompt-response groups and compare the candidate responses under the same prompt, selecting the better answer based on overall helpfulness, relevance, correctness, and completeness.

The annotators are shown the prompt together with the candidate responses and are asked to make a preference judgment using the same criteria throughout the study. The human study only involves expert evaluation of model outputs and does not collect personal or sensitive user information; no separate crowdsourcing platform is used.

Figure 7 shows that GRGC is preferred over both SeqKD and GAD in human comparison. This result is important because it confirms that the gains of GRGC are not limited to judge-model scores or

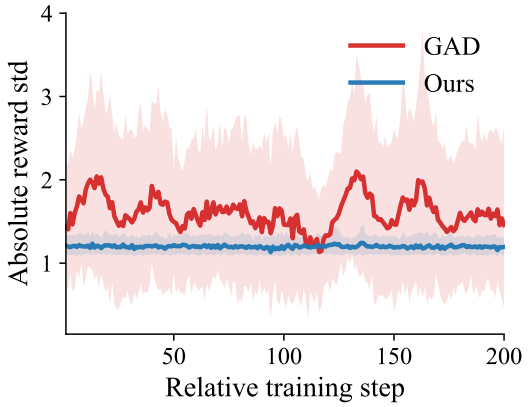

Figure 10: Unnormalized prompt-group reward spread . The solid lines show the mean prompt-group reward standard deviation, and shaded regions show the 10th–90th percentile range across prompt groups at each step. GRGC keeps the absolute reward spread in a more controlled range than GAD, confirming that the relative-spread improvement is not caused only by normalization with the global reward scale.

response-format differences. Instead, the improvements induced by reward-geometry conditioning are visible to human evaluators as better overall answer quality.

G.4 More Visualization Analysis

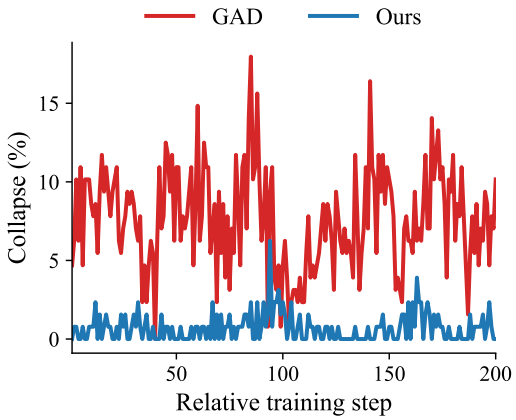

Figure 8: Collapsed prompt groups during training. GRGC keeps the fraction with relative reward std below 0.15 consistently lower than GAD.

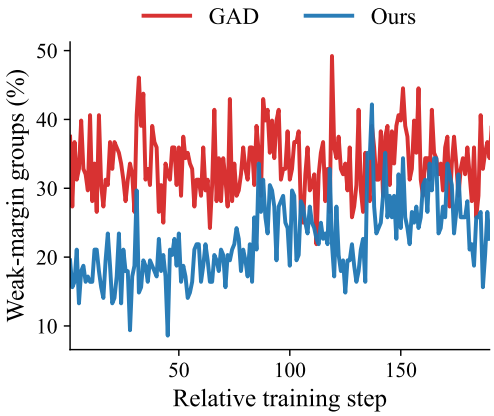

Figure 9: Weak-margin groups during training. GRGC reduces the fraction whose normalized top-2 gap is below 0.08.

Figure 8 and Figure 9 provide a more direct visualization of the geometric changes induced by GRGC. Here, a *collapsed prompt group* means that the reward spread within a prompt-wise group is too small, so the candidate responses become poorly separated by the critic; concretely, we mark a group as collapsed when its relative reward standard deviation satisfies $\mathrm{rel_std} < 0.15$, and the reported collapse ratio is the fraction of such groups in a training window. A *weak-margin group* means that the top candidates remain too close even after grouping. We define the top-2 margin as the reward difference between the highest-scored and second-highest-scored student responses within the same prompt group, and normalize it by the full within-group reward span to obtain $\mathrm{top1_top2_gap_ratio}$. Concretely, we mark a group as weak-margin when $\mathrm{top1_top2_gap_ratio} < 0.08$, and the reported weak-margin ratio is the fraction of prompt groups with such insufficient top-2 separation.

Under these definitions, Figure 8 shows that our method suppresses collapsed prompt groups throughout the training window, rather than only improving a single summary statistic at the end of training.

Table 15: GPT-OSS-120B evaluation results on the LMSYS test set. We report the averaged GPT-OSS-120B evaluation score and win rate.

Method	Qwen2.5-3B-Instruct		Qwen2.5-1.5B-Instruct	
	Score	Win Rate	Score	Win Rate
GPT-5-Chat Teacher	66.13	76.8%	66.13	76.8%
Before Distill	54.75	49.5%	52.81	43.4%
SeqKD [22]	55.65	54.7%	54.20	46.8%
GAD [18]	56.41	50.1%	55.56	47.6%
Ours	**60.52**	**60.8%**	**59.77**	**59.9%**

Table 16: Average response length (token length) of different methods and models on the LMSYS test set. The models are trained using the LMSYS-Chat training set with the GPT-5-Chat teacher.

Model	Teacher	Before Distill.	SeqKD	GAD	Ours
Qwen2.5-3B-Instruct	329.1	338.9	318.2	438.0	328.1
Qwen2.5-1.5B-Instruct	329.1	317.8	310.4	396.1	335.7

Figure 10 complements the relative-spread plot by showing that the unnormalized prompt-group reward standard deviation is also kept in a more controlled range under GRGC, so the effect is not simply produced by dividing by the global reward scale. Figure 9 further shows that this improvement is accompanied by a lower proportion of weak-margin groups in the reward signal eventually consumed by grouped optimization. Taken together, these visualizations support the intended division of labor between the two components of GRGC: critic-side OT calibration reduces group collapse, while policy-side reward shaping makes the remaining grouped margins more usable for GRPO.

These figures also clarify why our method should not be interpreted as merely making rewards "larger" or "more extreme." The goal is not to maximize separation indiscriminately, but to keep prompt groups neither collapsed nor dominated by uninformative near-ties. In this regime, grouped optimization can exploit reward differences more reliably and with fewer pathological updates.

G.5 GPT-OSS Evaluation Results

We further evaluate the LMSYS-trained models with GPT-OSS-120B as an additional automatic judge, in order to check whether the observed gains depend on the specific Qwen2.5-72B evaluation model used in the main experiments. The evaluation follows the same automatic protocol described in Appendix F: the judge assigns scalar scores to the student response and the reference response, from which we compute the normalized score and win rate; the only change is that GPT-OSS is used as the judge model. Table 15 shows the same trend under this alternative evaluator. For Qwen2.5-3B-Instruct, GRGC improves the score from 56.41 with GAD to 60.52, and raises the win rate from 50.1% to 60.8%. The improvement is even larger for Qwen2.5-1.5B-Instruct, where GRGC improves the score from 55.56 to 59.77 and the win rate from 47.6% to 59.9%. These results indicate that the advantage of GRGC is not tied to a single judge model: the better reward geometry learned during adversarial distillation translates into responses that are also preferred by an independent GPT-OSS evaluator.

G.6 Analysis of Token Length

Table 16 reports average response length as a behavioral diagnostic rather than as a direct quality metric. A good distilled student should avoid both severe under-generation, which often indicates missing content or truncated reasoning, and unnecessary verbosity, which often reflects repetitive or weakly controlled generation. From this perspective, GRGC is attractive because its response length remains much closer to the teacher than GAD while avoiding the strong length inflation that appears in standard adversarial distillation. SeqKD, by contrast, tends to produce shorter responses, which is consistent with its reliance on static teacher trajectories and its tendency to under-explore during on-policy rollout.

This result fits the broader interpretation of GRGC. Reward geometry conditioning does not simply encourage the student to speak more; rather, it provides a better-conditioned optimization signal so that the student can allocate response length more appropriately. In our experiments, this translates into outputs whose lengths remain closer to the teacher's response budget while still achieving higher automatic and human-evaluation performance.

G.7 Rule-based Instruction Following on IFEval

We evaluate instruction following on IFEval [97], whose verifiable constraints are scored by deterministic rule-based checkers rather than an LLM judge. Table 17 reports strict and loose accuracy. GRGC achieves the best result at both student scales: relative to GAD, it improves strict/loose accuracy by $1.80/2.28$ points for Qwen2.5-1.5B and $2.63/3.48$ points for Qwen2.5-3B. These results provide judge-free evidence that the gains extend to verifiable instruction compliance.

Table 17: Instruction-level IFEval accuracy (%) under strict and loose verification (834 instructions).

Student	Method	Strict	Loose
Qwen2.5-1.5B	SeqKD	54.56	57.79
	GAD	54.32	58.51
	GRGC	**56.12**	**60.79**
Qwen2.5-3B	SeqKD	66.91	70.14
	GAD	68.35	72.06
	GRGC	**70.98**	**75.54**

G.8 Long-form Mathematical Reasoning

We further test GRGC outside open-ended chat by sampling 30,000 training problems from OpenR1-Math-220k [98], which provides DeepSeek-R1-generated reasoning trajectories [21]. We train Qwen2.5-1.5B with a 10,000-token maximum sequence length for teacher trajectories, student rollouts, and critic scoring; many teacher and student outputs exceed 8,000 tokens. We evaluate benchmark accuracy on Minerva Math [99] and 340 retained Gaokao-MathQA questions from AGIEval [100], both of which have deterministic reference answers.

Table 18 shows that GRGC achieves the highest accuracy on both benchmarks. It improves over GAD by 2.58 points on Minerva Math and 3.53 points on Gaokao-MathQA, while exceeding the undistilled student by 6.25 and 5.59 points, respectively. The gains therefore persist under domain-specific supervision, ground-truth evaluation, and long reasoning trajectories beyond the response lengths used in our chat experiments.

Table 18: Accuracy (%) of Qwen2.5-1.5B after long-form mathematical-reasoning distillation on 30000 OpenR1-Math-220k problems.

Method	Minerva Math	Gaokao-MathQA
Before Distillation	11.40	43.53
SeqKD	13.97	43.82
GAD	15.07	45.59
GRGC	**17.65**	**49.12**

G.9 Evaluation on Math500

We additionally evaluate mathematical reasoning on Math500 [101]. This setting is intentionally challenging for our training distribution: the student is distilled from GPT-5-Chat on LMSYS-Chat, where only 3.4% of the training prompts are math-related. Thus, Math500 measures whether black-box distillation preserves and transfers sparse reasoning ability from a mostly general conversational dataset, rather than whether the method is specialized for math training.

We compute Math500 accuracy with a rule-based answer-matching protocol rather than an LLM judge. Each model is prompted with the standard instruction wrapper plus an explicit request to put

Table 19: Math500 evaluation for Qwen2.5-1.5B-Instruct distilled from GPT-5-Chat on LMSYS-Chat. Only 3.4% of LMSYS-Chat training prompts are math-related, making this an out-of-domain reasoning stress test rather than a math-specialized training setting.

Method	Before Distill	SeqKD	GAD	Ours
Accuracy	46.4%	39.8%	44.0%	49.6%

Table 20: Multi-seed Qwen2.5-72B automatic evaluation score. Seed=0 is the default evaluation seed used in the main results. The 95% Confidence Interval (CI) represents the range within which the true mean score is expected to lie with 95% confidence.

Method	Seed=0	Seed=1	Seed=2	Seed=3	Seed=4	Mean	Std	95% CI
SeqKD	47.06	47.10	47.05	47.04	47.07	47.06	0.02	[47.04, 47.08]
GAD	48.44	48.44	48.43	48.38	48.43	48.42	0.03	[48.38, 48.46]
Ours	50.19	50.20	50.15	50.22	50.18	50.19	0.03	[50.15, 50.23]

the final answer in `\boxed{}`, and responses are generated greedily. For each output, we extract the final answer by first taking the last `\boxed{}` expression, then falling back to answers after `####`, explicit phrases such as "final answer is", and finally the last mathematical expression or number. The extracted answer is compared against the dataset's `final_answer` field using exact string match after LaTeX normalization, numerical comparison with tolerance 10^{-6} for numbers and fractions, and SymPy-based symbolic equivalence when available. Extraction failures are counted as incorrect, and accuracy is the percentage of correct examples among the 500 Math500 problems.

Table 19 shows that naive sequence-level imitation can hurt mathematical reasoning: SeqKD drops from the undistilled model's 46.4% accuracy to 39.8%. GAD partially recovers this degradation, reaching 44.0%, but still remains below the original student. In contrast, GRGC achieves 49.6%, outperforming the undistilled model and both distillation baselines. This result is consistent with our main claim: by improving grouped advantage construction from adversarial rewards, GRGC provides a better-conditioned on-policy signal that can exploit the limited math-related supervision in LMSYS-Chat without degrading the student's reasoning ability.

G.10 Multi-seed Evaluation Robustness

The automatic evaluation results in the main paper use the default evaluation seed, *i.e.*, Seed=0. To verify that the observed gains do not come from judge-sampling randomness, we additionally repeat the Qwen2.5-72B automatic evaluation with five evaluation seeds while keeping the model generations fixed. Tables 20 and 21 report both the individual seed results and the mean/std across seeds. The results are robust across evaluation seeds: the standard deviation is only 0.03 for our normalized score and 0.54 for win rate, while the improvement of GRGC over GAD is much larger than these fluctuations. This confirms that the performance gains of GRGC are robust to evaluation-seed variation rather than being an artifact of a particular judge sample.

Table 21: Multi-seed Qwen2.5-72B automatic evaluation win rate. Seed=0 is the default evaluation seed used in the main results. The 95% Confidence Interval (CI) represents the range within which the true mean score is expected to lie with 95% confidence.

Method	Seed=0	Seed=1	Seed=2	Seed=3	Seed=4	Mean	Std	95% CI
SeqKD	18.4	18.6	18.4	18.8	18.2	18.46	0.24	[18.16, 18.75]
GAD	25.1	25.5	24.6	24.8	24.8	24.97	0.32	[24.58, 25.36]
Ours	45.3	45.3	44.9	45.3	46.3	45.43	0.54	[44.75, 46.10]

H Complexity Analysis

Let B be the total number of student responses in a minibatch, N the prompt-group size, and $G = B/N$ the number of prompt groups. Let L denote the average generated response length, and let $\mathcal{F}_\pi^{\mathrm{roll}}(L)$, $\mathcal{F}_\pi^{\mathrm{upd}}(L)$, and $\mathcal{F}_D(L)$ denote the model-scale costs of actor rollout, actor update, and critic forward-backward on a length-L response. This notation separates expensive neural-network computation from cheap grouped scalar operations such as BT loss, GRPO normalization, OT calibration, and group power shaping.

Baseline GAD. For one minibatch, baseline GAD contains four conceptually distinct pieces:

1. **Actor rollout.** Generating B student responses of average length L costs
$$\mathcal{O}\big(B\,\mathcal{F}_\pi^{\mathrm{roll}}(L)\big). \tag{112}$$
2. **Critic scoring and BT training.** The critic evaluates the G teacher responses and the B student responses, then forms Bradley–Terry losses on the resulting sequence scores. Thus the exact critic-side cost is
$$\mathcal{O}\big((B+G)\,\mathcal{F}_D(L)\big) + \mathcal{O}(B), \tag{113}$$
where the extra $\mathcal{O}(B)$ term is the scalar BT loss itself after sequence scores are formed. Since $G = B/N$ and N is fixed in our setting, this simplifies to
$$\mathcal{O}\big(B\,\mathcal{F}_D(L)\big) + \mathcal{O}(B). \tag{114}$$
3. **GRPO normalization.** Groupwise mean/std computation and advantage construction operate only on B scalar sequence rewards:
$$\mathcal{O}(B). \tag{115}$$
4. **Actor update.** PPO/GRPO updates over the generated trajectories cost
$$\mathcal{O}\big(B\,\mathcal{F}_\pi^{\mathrm{upd}}(L)\big). \tag{116}$$

Hence the total time complexity of GAD can be written as
$$\mathcal{C}_{\mathrm{GAD}} = \mathcal{O}\big(B\,\mathcal{F}_\pi^{\mathrm{roll}}(L)\big) + \mathcal{O}\big(B\,\mathcal{F}_D(L)\big) + \mathcal{O}\big(B\,\mathcal{F}_\pi^{\mathrm{upd}}(L)\big) + \mathcal{O}(B), \tag{117}$$
where the final $\mathcal{O}(B)$ term absorbs BT loss formation and GRPO normalization. In other words, the dominant cost of GAD comes from rollout, critic forward-backward, and actor update; BT loss and GRPO normalization themselves are low-order grouped scalar operations.

Critic-side OT calibration. Our OT term is added after the critic has already produced one scalar sequence reward for each student response. For each prompt group, OT sorts the N scalar rewards and matches them to N Gaussian target quantiles. Sorting dominates this step, yielding
$$\mathcal{O}(N \log N) \tag{118}$$
time per prompt group and therefore
$$\mathcal{O}(B \log N) \tag{119}$$
time per minibatch. The additional memory cost is linear in the number of grouped scalar rewards:
$$\mathcal{O}(B). \tag{120}$$
Because N is small and fixed in our experiments (e.g., $N = 8$), this overhead is low-order relative to the critic forward-backward pass itself.

Policy-side group power modulation. The group power transform computes a per-group mean and standard deviation, standardizes the grouped rewards, and applies an elementwise signed power map. These are all scalar operations on B rewards:
$$\mathcal{O}(N) \tag{121}$$
per prompt group and
$$\mathcal{O}(B) \tag{122}$$
per minibatch, with
$$\mathcal{O}(B) \tag{123}$$
additional memory. Thus group power is asymptotically cheaper than OT and negligible relative to rollout or model updates.

Table 22: Asymptotic training complexity comparison. Here B is the number of sampled student responses in a minibatch, N is the prompt-group size, and L is the average response length. Since N is a small fixed group size in our setting, grouped operations with complexity $\mathcal{O}(B \log N)$ behave as low-order overheads and do not change the dominant model-scale training complexity. As a result, the total wall-clock training time of GRGC is expected to remain nearly identical to that of GAD.

Method	Time complexity	Memory complexity	Dominant terms
GAD	$\mathcal{O}\big(B\,\mathcal{F}_\pi^{\mathrm{roll}}(L)\big)$ $+\mathcal{O}\big(B\,\mathcal{F}_D(L)\big)$ $+\mathcal{O}\big(B\,\mathcal{F}_\pi^{\mathrm{upd}}(L)\big)$ $+\mathcal{O}(B)$	$\mathcal{M}_\pi + \mathcal{M}_D$ $+\mathcal{O}(B)$	rollout critic forward/backward actor update
GRGC (ours)	$\mathcal{O}\big(B\,\mathcal{F}_\pi^{\mathrm{roll}}(L)\big)$ $+\mathcal{O}\big(B\,\mathcal{F}_D(L)\big)$ $+\mathcal{O}\big(B\,\mathcal{F}_\pi^{\mathrm{upd}}(L)\big)$ $+\mathcal{O}(B \log N)$	$\mathcal{M}_\pi + \mathcal{M}_D$ $+\mathcal{O}(B)$	rollout + critic + actor update OT sorting/matching group power shaping

Overall comparison with GAD. Combining the two additions gives

$$\mathcal{C}_{\mathrm{GRGC}} = \mathcal{C}_{\mathrm{GAD}} + \mathcal{O}(B \log N) + \mathcal{O}(B) = \mathcal{C}_{\mathrm{GAD}} + \mathcal{O}(B \log N), \tag{124}$$

so our method preserves the same model-scale dominant terms as GAD and adds only grouped scalar processing on top of the critic-emitted sequence rewards. The key practical implication is that GRGC does *not* introduce an extra rollout loop, an extra critic network, or another model-scale optimization stage; its additional cost comes only from cheap per-group operations after the main neural computation has already been done. In particular, BT loss formation, GRPO normalization, OT matching, and group power shaping are all low-order operations on scalar group rewards after the dominant neural computation is finished. Since N is small and fixed in our setting, these terms do not materially affect end-to-end training time, so the overall wall-clock training cost of GRGC is expected to be nearly the same as that of GAD. Table 22 summarizes this asymptotic comparison.

I Broader Impact

Our work studies how to make black-box distillation more effective when the teacher is accessible only through output queries. On the positive side, better black-box distillation can reduce the cost of obtaining capable small language models, which may improve accessibility for research labs, educational use, and resource-constrained deployment settings. It may also enable more systematic study of proprietary-model behavior using open students that are easier to inspect, evaluate, and stress-test.

At the same time, improvements in distillation efficiency can also accelerate the transfer of undesirable behaviors from proprietary teachers into cheaper and more deployable students. More capable distilled models may be misused for low-cost spam generation, manipulative persuasion, or unsafe domain advice. In addition, more effective distillation can make capability transfer easier to reproduce, which is a dual-use concern. Our method does not require additional access to private model internals or private data, but it can increase the effectiveness of capability transfer from black-box teachers. We therefore view stronger safety evaluation, misuse screening, and downstream deployment constraints as important complements to progress in black-box distillation.

J Limitations

Our method is built on top of GAD and inherits both its strengths and its limitations. In particular, although GRGC improves the quality of on-policy black-box distillation, it still relies on an adversarial critic and on-policy rollouts, and is therefore typically slower than simple supervised baselines such as SeqKD. This is a real tradeoff: our goal is not to make on-policy distillation cheaper than off-policy imitation in absolute terms, but to make the extra optimization cost buy a better-conditioned and more effective training signal.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the core claim that GRGC improves on-policy black-box distillation by conditioning prompt-group reward geometry, and this matches the method, experiments, and limitations discussed in the paper; see the abstract, Section 1, and Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes an explicit Limitations discussion in the appendix, covering the additional training cost relative to SeqKD; see Appendix "Limitations".

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The appendix provides explicit assumptions, formal propositions, and corresponding derivations and proof sketches for the theoretical claims used in the paper; see Appendix "Mechanism Analysis for OT Calibration and Group Power".

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper specifies datasets, model families, teacher models, evaluation benchmarks, key hyperparameters, training schedule, and hardware used for the reported experiments; see Section 4.1 and the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper and appendix provide the datasets, model families, evaluation prompts, pseudocode, hyperparameters, and reproduction-relevant implementation details needed to reproduce the main experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 and the appendix describe datasets, evaluation splits, teacher and student models, baselines, training schedule, group size, learning rates, OT and power hyperparameters, sensitivity analyses for the two method-specific hyperparameters, and hardware.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper reports multi-seed automatic evaluation results with mean and standard deviation across five evaluation seeds in the Appendix, showing that the gains remain much larger than judge-sampling variability. It also supports the empirical conclusions through consistent trends across multiple teacher-student pairs, multiple test sets, sensitivity analyses, training-stability analyses, and complementary human evaluation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies the accelerator type and count, the training schedule in epochs and optimization steps, the key sequence lengths, and the overall training recipe used for the reported experiments; see Section 4.1 and Appendix "Complexity Analysis".

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our knowledge, the research conforms to the NeurIPS Code of Ethics; the paper also discusses broader impacts and limitations, including dual-use concerns.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both positive impacts (cheaper, more accessible small models) and negative impacts (more effective transfer of harmful behaviors), and also mentions the need for safeguards; see the main-paper Broader Impact paragraph and the appendix Broader Impact section.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This submission does not release a new high-risk model, model checkpoint, or scraped dataset asset; the paper instead discusses broader-impact risks and deployment considerations.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper credits the datasets, models, and baseline methods used in the experiments and identifies the main external assets explicitly in the experimental sections and references.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The current submission does not release a new dataset, benchmark, model checkpoint, or public code package as an asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The appendix describes the human-evaluation setup, the annotator pool, the per-model sample allocation, and the evaluation criteria used by the annotators. The study does not rely on an external crowdsourcing platform.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The human evaluation concerns expert assessment of model outputs only, does not involve collection of personal or sensitive participant data, and is described in the appendix together with the evaluation procedure and scope of human involvement.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core contribution is a reward-geometry conditioning method for black-box distillation rather than a method that uses an LLM as a novel algorithmic component; teacher and judge models are standard experimental ingredients and are documented in Section 4.1.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.