

DiverseGRPO: Mitigating Mode Collapse in Image Generation via Diversity-Aware GRPO

Henglin Liu^{1,2*}, Huijuan Huang^{2†§}, Jing Wang^{2,3*}, Chang Liu¹, Xiu Li^{1†}, Xiangyang Ji^{1†}

¹Tsinghua University, ²Kling Team, Kuaishou Technology, ³Shenzhen Campus of Sun Yat-sen University

Project Page [§]*Project Leader.* [†]*Corresponding Authors.* ^{*}*Work Conducted During Internship.*

liu-hl24@mails.tsinghua.edu.cn

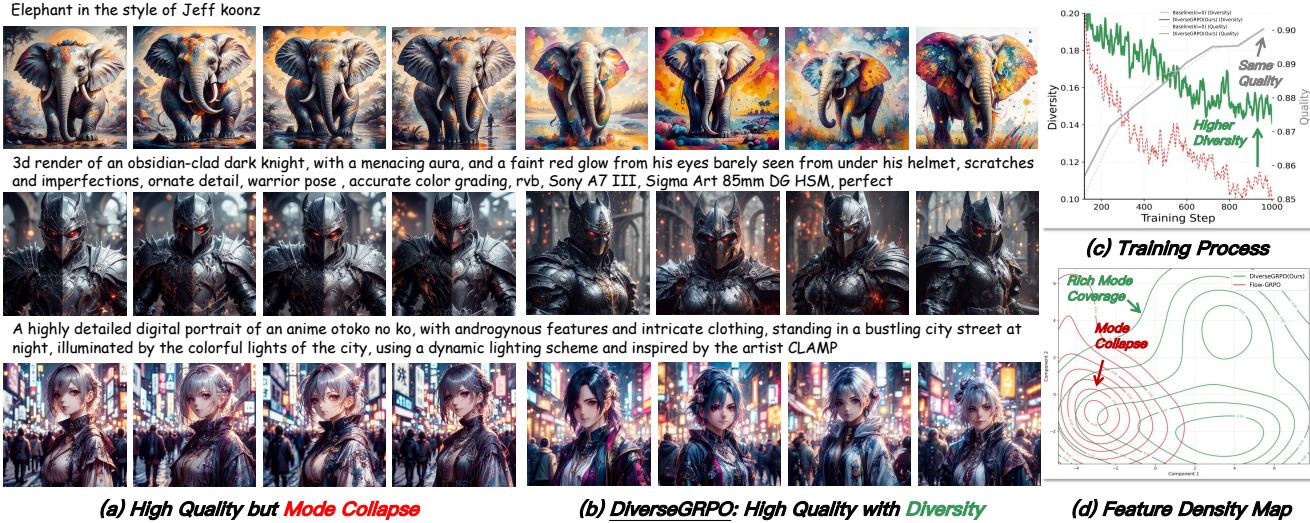


Figure 1. (a) Image generation models trained with GRPO suffer from mode collapse (similar faces, camera angles, etc.), which limits their applicability in creative scenarios. (b) The proposed DiverseGRPO method achieves higher diversity while maintaining comparable quality. (c) DiverseGRPO successfully maintains a healthier level of diversity across the entire duration of training, while the baseline method suffers from a premature collapse. (d) In the Inception feature space, DiverseGRPO generates images that cover a significantly broader range of semantic features, effectively mitigating mode collapse.

Abstract

Reinforcement learning (RL), particularly GRPO, improves image generation quality significantly by comparing the relative performance of images generated within the same group. However, in the later stages of training, the model tends to produce homogenized outputs, lacking creativity and visual diversity, restricting the application scenarios of the model. This issue can be analyzed from both reward modeling and generation dynamics perspectives. First, traditional GRPO relies on single-sample quality as the reward signal, driving the model to converge toward a few high-reward generation modes while neglecting distribution-level diversity. Second, conventional GRPO regularization neglects the dominant role of early-stage denoising in preserving diversity, causing a misaligned regularization budget that limits the achievable quality–diversity trade-off. Motivated by these insights, we

revisit the diversity degradation problem from both reward modeling and generation dynamics. At the reward level, we propose a distributional creativity bonus based on semantic grouping. Specifically, we construct a distribution-level representation via spectral clustering over samples generated from the same caption, and adaptively allocate exploratory rewards according to group sizes to encourage the discovery of novel visual modes. At the generation level, we introduce a structure-aware regularization, which enforces stronger early-stage constraints to preserve diversity without compromising reward optimization efficiency. Experiments demonstrate that our method achieves an 13%~18% improvement in semantic diversity under matched quality scores, establishing a new Pareto frontier between image quality and diversity for GRPO-based image generation.

1. Introduction

The diversity of generated images is a key criterion for evaluating the performance of generative models. A significant loss of diversity represents a major challenge for the practical application, particularly in creative fields such as digital art, advertising, and game design, where novelty and variety are fundamental to their success.

However, reinforcement learning from human feedback (RLHF) [2, 5] techniques for image generation [3], such as Flow-GRPO [21] and DanceGRPO [30], have achieved remarkable progress in aligning text-to-image generation models with human aesthetic preferences, recent studies [6, 11, 31] in large language model(LLM) have revealed a critical limitation of GRPO-based approaches: the degradation of generation diversity. As illustrated in Fig. 1, this phenomenon manifests as homogenized results (nearly identical character appearances and highly similar perspectives), indicating a collapse of semantic diversity. That is because the intrinsic objective of reward maximization tends to overfit the model to a narrow subset of high-reward modes, effectively encouraging the model to reproduce ‘safe’ or ‘high-score’ patterns while suppressing creative or unconventional outputs.

This observation raises a deeper question: **Is diversity degradation an inevitable byproduct of reward optimization, or is it a symptom of misaligned learning objectives and generation dynamics?**

We begin by examining the problem from the reward modeling perspective. As shown in Fig. 2.a, GRPO relies solely on single-sample reward signals, where the reward model assigns individual scores to each image without considering the relationships between samples. This approach leads the generative model to become ‘short-sighted’ and ‘conservative’, focusing on maximizing immediate rewards at the expense of exploration and innovation. To analyze this effect more formally, consider that the conditional generation distribution can be decomposed into K semantic modes: $\pi_\theta(x | p) = \sum_{k=1}^K w_k \pi_\theta^k(x | p)$, where each component π_θ^k represents a distinct visual mode (e.g., composition, lighting, or style), and w_k is its mixture weight. Let the average reward within each mode be: $\bar{r}_k = \mathbb{E}_{x \sim \pi_\theta^k} [r(x, p)]$, then the expected reward objective can be rewritten as $J(\theta) = \sum_{k=1}^K w_k \bar{r}_k$. During optimization, modes with larger average reward \bar{r}_k gain higher sampling probability, yielding the following **replicator dynamics**:

$$\frac{dw_k}{dt} = w_k (\bar{r}_k - \mathbb{E}_j [\bar{r}_j])$$

This equation describes a self-reinforcing selection process: modes with slightly higher average rewards continue to grow in weight, while others are gradually eliminated. At

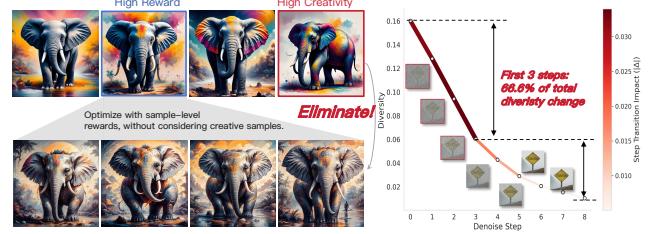


Figure 2. Analysis of the reasons for mode collapse: (Left) Policy model collapse into high-reward modes due to single sample reward modeling. (Right) Conventional regularization neglects the dominant role of early-stage denoising in preserving diversity.

equilibrium, only the dominant high-reward mode survives:

$$w_k = \begin{cases} 1, & \text{if } k = \arg \max_j \bar{r}_j, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, single-sample reward optimization implicitly drives the model toward a unimodal distribution, leading to homogenized outputs and a collapse in visual diversity. Therefore, to prevent mode extinction, it is insufficient to adjust sampling strategies or model architectures; the reward itself must be redesigned to reshape the underlying dynamics. Instead of rewarding individual samples independently, we introduce a distribution-aware reward that depends on the semantic structure of the generated set for each prompt. Specifically, we construct a distribution-level representation via spectral clustering over samples generated from the same text prompt, and adaptively allocate exploratory rewards according to group sizes, which encourages the discovery of novel visual modes.

In addition to the external reward signal, we further investigate how the intrinsic denoising dynamics of diffusion models influence visual diversity. We compute the perceptual similarity between samples generated at different denoising steps, where all samples within the same denoising step share the same latent variable, using the DreamSim [10] model. As shown in Fig. 2.b, The more denoising steps that are shared, the more similar the samples are, which aligns with previous observations [15, 25]. However, it is noteworthy that the decline in diversity is steeper during the early denoising phase (as indicated by the red bar, where the first one-third of the denoising steps accounts for approximately 66% of the overall diversity change). This suggests that the early denoising phase has a disproportionately large impact on the resulting diversity. From the perspective of mitigating mode collapse, the denoising trajectory forms an imbalanced ‘diversity budget’, where early steps are diversity-critical while later steps mainly refine visual quality. Under this perspective, because the diffusion variance is largest in early steps, the KL penalty becomes effectively weakest exactly when the budget should be highest, resulting in a structural mismatch that accelerates mode extinction. To resolve this budget misallocation,

tion, we reformulate diversity preservation as a structure-aware regularization scheduling problem, and replace the uniform KL penalty with a Wasserstein constraint that concentrates the regularization budget in the high-impact early phase, while lifting the constraint in the late phase to preserve reward quality and image fidelity. To validate the effectiveness and generality of our approach, we conduct extensive experiments across multiple diffusion backbones (SD3.5 [9] and Flux [4]) and heterogeneous reward models (PickScore [17] and HPSv3 [22]). In all settings, our method consistently improves semantic diversity while preserving or even enhancing image quality, outperforming existing GRPO-based baselines under matched reward budgets.

Overall, we conducted an in-depth analysis of mode collapse in GRPO for image generation and innovatively designed a new GRPO training paradigm. Specifically, our main contributions are as follows:

- **Diversity-aware reward modeling for mitigating mode collapse.** DiverseGRPO introduces a distribution-level reward formulation that moves beyond conventional single-sample scoring. By applying spectral semantic grouping and assigning exploration bonuses inversely to group size, our approach explicitly incentivizes the discovery and preservation of rare visual modes—directly countering the root cause of mode collapse.
- **Structure-aware regularization tailored to diffusion dynamics.** We uncover a fundamental misalignment between standard regularization and the uneven diversity sensitivity along diffusion trajectories. To correct this, we design a Wasserstein-based structure-aware constraint that applies stronger early-step regularization, where diversity is most fragile, while gradually relaxing the penalty in later stages to enhance the effectiveness of reward optimization.
- **State-of-the-art diversity-quality trade-off in GRPO-based image generation.** Comprehensive experiments demonstrate that DiverseGRPO consistently mitigates mode collapse, achieving up to 18% improvement in semantic diversity under matched quality. Across multiple backbones and reward models, our method establishes a new Pareto frontier for GRPO-driven image generation.

2. Background

2.1. RL for Diffusion and Flow Models.

Building on the success of reinforcement learning (RL) in Large Language Models (LLMs), algorithms such as PPO [3, 23] and DPO [24] have been adapted to diffusion models for preference alignment and task-specific optimization. This trend has extended to flow-based models. Flow-GRPO [21] and DanceGRPO [30] integrate GRPO-style policy updates into flow matching, transforming de-

terministic ODE sampling into stochastic SDE formulations to introduce exploration noise. Subsequent works like Mix-GRPO [19] propose a hybrid ODE-SDE sampling strategy to improve training efficiency. Addressing the noise inconsistency issue in SDE sampling, Flow-CPS [26] introduces a noise-consistent scheme for more accurate reward estimation. Further innovations tackle the challenge of reward sparsity in multi-step trajectories. TempFlowGRPO [12] move beyond assigning a single global reward. In a parallel and significant development, BranchGRPO [20] introduces a tree-structured branching mechanism within the diffusion/flow matching inversion process. It allows multiple trajectories to share prefixes and split at intermediate steps, enabling dense, layer-wise reward fusion.

Most existing methods focus on improving the efficiency of policy optimization but overlook mode collapse issue, which severely diminishes visual diversity and limits the practical applicability of the models. In this work, we conduct an in-depth analysis of this problem and propose an effective solution.

2.2. Mode-Collapse in Generation Models

Research on preventing mode collapse in large language models (LLMs) can be broadly divided into two approaches: one targeting sample or token selection, and the other integrating multiple reward signals to guide generation. DivPO [18] selects preference pairs of samples online, using ‘high-quality and rare’ generated outputs as positive examples and ‘common but low-quality’ results as negative examples, improving diversity while maintaining generation quality. Cui et al. [6] addresses entropy collapse by pruning action probabilities and applying KL regularization to high-covariance tokens (action probability and logits changes), helping the policy avoid entropy collapse and enhancing diversity. Another line of work integrates multiple reward signals to improve diversity in generation. CPO [16] modularizes the injection of multiple creative dimensions into the preference optimization objective, adjusting the weight of each dimension to adapt to varying task needs. In the context of image generation, Astolfi et al. [1] use Pareto fronts to compare image generation models. They find that newer models like LDM-Turbo achieve greater consistency and realism but are less diverse, while older models such as LDM offer superior diversity. DiADM [8] decouples diversity and quality by introducing a diversity-aware module with pseudo-unconditional feature inputs. Ding et al. [7] utilize human judgments on similarity and combining latent space projection with contrastive learning to progressively infer diversity metrics. However, it introduces significant complexity due to the multi-stage training process. In this work, we address the diversity degradation problem from the dual perspectives of distributional reward modeling and generation dynamics.

DiverseGRPO Framework

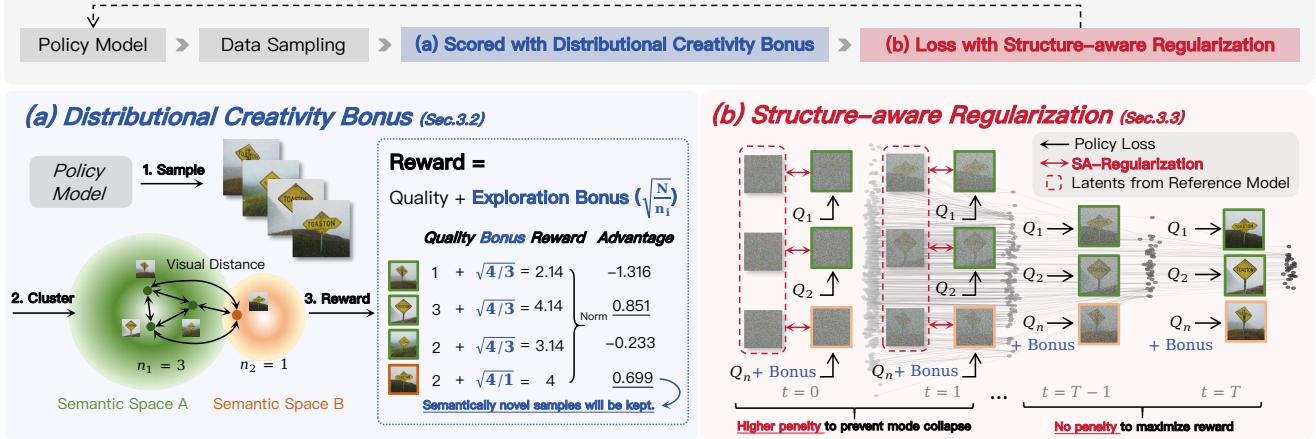


Figure 3. DiverseGRPO employs two primary strategies to mitigate mode collapse: (a) A distributional creativity bonus mechanism based on semantic grouping. It begins by applying spectral clustering to images generated from the same caption, then assigns exploratory rewards according to cluster size to encourage the emergence of novel visual modes. (b) Structure-aware regularization imposes stronger constraints during the initial denoising stages to preserve sample diversity, while gradually relaxing the penalty in later stages to enhance the effectiveness of reward optimization.

3. Method

3.1. Preliminary

The goal of Reinforcement Learning is to learn a policy that maximizes expected cumulative reward. GRPO achieves this by optimizing its policy model to maximize the following objective:

$$\mathcal{J}_{\text{Flow-GRPO}}(\theta) = \mathbb{E}_{c \sim \mathcal{C}, \{\mathbf{x}^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | c)} f(r, \hat{A}, \theta, \epsilon, \beta) \quad (1)$$

$$\begin{aligned} f(r, \hat{A}, \theta, \epsilon, \beta) &= \frac{1}{G} \sum_{i=1}^G \frac{1}{T} \sum_{t=0}^{T-1} \left(\min \left(r_t^i(\theta) \hat{A}_t^i, \right. \right. \\ &\quad \left. \left. \text{clip} \left(r_t^i(\theta), 1 - \epsilon, 1 + \epsilon \right) \hat{A}_t^i \right) \right) \\ &\quad - \beta D_{\text{KL}} (\pi_\theta \| \pi_{\text{ref}}), \\ r_t^i(\theta) &= \frac{p_\theta (\mathbf{x}_{t-1}^i | \mathbf{x}_t^i, \mathbf{c})}{p_{\theta_{\text{old}}} (\mathbf{x}_{t-1}^i | \mathbf{x}_t^i, \mathbf{c})}, T \text{ is the timestep.} \end{aligned} \quad (2)$$

Given a prompt \mathbf{c} , the flow model p_θ samples a group of G individual images $\{\mathbf{x}_0^i\}_{i=1}^G$ and the corresponding reverse-time trajectories $\{(\mathbf{x}_T^i, \mathbf{x}_{T-1}^i, \dots, \mathbf{x}_0^i)\}_{i=1}^G$. Then, the advantage of the i -th image is calculated by normalizing the group-level rewards as follows:

$$\hat{A}_t^i = \frac{R(\mathbf{x}_0^i, \mathbf{c}) - \text{mean} \left(\{R(\mathbf{x}_0^i, \mathbf{c})\}_{i=1}^G \right)}{\text{std} \left(\{R(\mathbf{x}_0^i, \mathbf{c})\}_{i=1}^G \right)} \quad (3)$$

Flow-GRPO transforms the original deterministic ODE into an SDE. A key property of this transformation is that the resulting SDE preserves the original model's marginal probability density function at every point in time. The ODE and

SDE is as follows:

$$d\mathbf{x}_t = \mathbf{v}_t dt \quad (4)$$

$$\begin{aligned} \mathbf{x}_{t+\Delta t} = \mathbf{x}_t + & \left[\mathbf{v}_\theta(\mathbf{x}_t, t) + \frac{\sigma_t^2}{2t} (\mathbf{x}_t + (1-t)\mathbf{v}_\theta(\mathbf{x}_t, t)) \right] \Delta t \\ & + \sigma_t \sqrt{\Delta t} \epsilon \end{aligned} \quad (5)$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ injects stochasticity and $\sigma_t = a \sqrt{\frac{t}{1-t}}$. The KL divergence between π_θ and the reference policy π_{ref} is a closed form:

$$\begin{aligned} D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) &= \frac{\|\bar{\mathbf{x}}_{t+\Delta t, \theta} - \bar{\mathbf{x}}_{t+\Delta t, \text{ref}}\|^2}{2\sigma_t^2 \Delta t} \\ &= \frac{\Delta t}{2} \left(\frac{\sigma_t(1-t)}{2t} + \frac{1}{\sigma_t} \right)^2 \\ &\quad \times \|\mathbf{v}_\theta(\mathbf{x}_t, t) - \mathbf{v}_{\text{ref}}(\mathbf{x}_t, t)\|^2 \end{aligned} \quad (6)$$

3.2. Distributional creativity bonus

After GRPO training, policy models tend to generate a narrow range of outputs that match the surface-level features preferred by the reward model. This limitation stems from reward models' inability to account for distributional diversity. They can only assess the quality of individual images in isolation, failing to recognize the range of valid visual interpretations for a given caption. To address this, we propose a distributional creativity reward that encourages the model to explore a wider range of visual modes. Our method consists of two stages: (1) distance calculation to quantify the visual differences perceived by humans between generated images and (2) spectral clustering to group the images based on these distances, followed by targeted exploration rewards for the underrepresented clusters.

Perception distance calculation. We begin by defining the pairwise visual distance between images. Given a set of images $\{x^1, x^2, \dots, x^n\}$, we compute the perceptual distance between each pair using *DreamSim* [10], a model designed to align with human visual similarity judgments.

The resulting pairwise distance matrix D is an $n \times n$ matrix where each element D_{ij} represents the perceptual distance between images x_i and x_j . The diagonal entries satisfy $D_{ii} = 0$, as the distance between an image and itself is zero. This matrix serves as the basis for subsequent similarity analysis or clustering tasks.

$$D = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix} \quad (7)$$

This matrix D serves as the basis for determining the visual diversity of the generated images.

Spectral Clustering. To effectively partition images based on their perceived differences, we use spectral clustering to divide the images into distinct clusters according to their visual similarity. We begin by constructing an affinity matrix A using a Gaussian kernel, which measures the similarity between images. The affinity between two images x_i and x_j is defined as:

$$A_{ij} = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right) \quad (8)$$

where σ is a scaling factor that controls the width of the Gaussian kernel. This affinity matrix nonlinearly maps the pairwise similarities into connection weights, thereby forming a graph that captures the complex intrinsic relationships within the data. Next, we compute the degree matrix D , which is a diagonal matrix where each entry D_{ii} is the sum of the affinities of image x_i with all other images:

$$D_{ii} = \sum_{j=1}^n A_{ij} \quad (9)$$

Using the degree matrix D and the affinity matrix A , we construct the normalized Laplacian matrix L :

$$L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \quad (10)$$

The Laplacian matrix captures the structure of the graph of images, where nodes (images) that are more similar are strongly connected, and nodes that are less similar are weakly connected. To identify distinct clusters, we perform an eigenvalue decomposition of the Laplacian matrix, obtaining the eigenvectors corresponding to the smallest eigenvalues. These eigenvectors encode the most significant components of the data. The resulting eigenvectors are used to form a new matrix V , where each row corresponds

to the eigenvector of an image. The rows of V are then clustered using k-means clustering to partition the images into k distinct clusters. This step allows us to identify groups of visually similar images, where each cluster represents a distinct visual mode.

Reward Allocation. After partitioning the images into clusters, we assign exploration rewards based on the cluster sizes. Smaller clusters, which represent underexplored visual modes, receive higher rewards. Specifically, the exploration reward for an image x_i in cluster C_k is inversely proportional to the number of images in that cluster:

$$E_i = \sqrt{\frac{N}{n_k}} \quad (11)$$

where n_k is the number of images in cluster C_k , N is the total number of samples with the same caption. Taking the square root further moderates the influence, balancing between rewarding diversity and maintaining model stability. By using Eq. 11, we ensure that clusters with fewer images (which are more likely to represent visually distinct modes) receive proportionally higher rewards. This incentivizes the model to generate images that are visually distinct and underrepresented, promoting diversity in the generated outputs.

The final reward for an image x^i is a combination of its quality score Q_i and the exploration reward E_i . The two components are weighted by a factor β , which balances the emphasis on quality and diversity:

$$R_i = Q_i + \beta \cdot E_i \quad (12)$$

where Q_i represents the intrinsic quality score by reward model, and β controls the influence of the diversity bonus. The combined score guides the model to generate high-quality images that also exhibit greater diversity. We then employ the group-relative advantage (Eq. 3).

3.3. Structure-aware Regularization

Rethinking KL penalty. To prevent a decline in diversity, existing methods [12, 20, 21] introduce a KL regularization term (as shown in Eq. 6), which constrains the divergence between the Gaussian distributions generated by the current policy model and the original base model at each denoising step. As shown in Fig. 2, we observe that the diversity of generated samples correlates strongly with the early denoising stages of the diffusion process. Intuitively, these early steps define the global structure and semantic modes of the output distribution, thus constituting a limited diversity budget. Preserving this budget requires strong constraints at early stages, where exploration determines the range of possible generations, while later stages should allow freer adaptation for high-reward refinement. However, in Eq. 6, σ^2 decreases rapidly as denoising progresses, leading to adaptive imbalance: In early stages, large variance σ^2

Table 1. Comparative evaluation of different backbone models and reward models. Higher values (\uparrow) indicate better performance for DreamSim, BeyondFID(abbreviated as BFID), ImageReward(abbreviated as ImR), PickScore, and UnifiedReward(abbreviated as UniReward), while lower values (\downarrow) are better for FID. The Improvement is calculated as $\frac{\text{Ours} - \text{Flow-GRPO}}{\text{Flow-GRPO}} \times 100\%$. For the FID metric, where lower values are better, the improvement percentage is calculated as $\frac{\text{Flow-GRPO} - \text{Ours}}{\text{Flow-GRPO}} \times 100\%$.

Algorithm	Diversity				Quality			
	DreamSim (\uparrow)	FID (\downarrow)	BFID (\uparrow)	SSIM (\uparrow)	CLIP (\uparrow)	ImR (\uparrow)	PickScore (\uparrow)	UniReward (\uparrow)
SD3.5-M / Pickscore								
Flow-GRPO	0.1278	56.206	0.0667	0.1701	0.3278	1.3650	0.8809	3.5817
Ours	0.1517	43.115	0.1895	0.2137	0.3339	1.4009	0.8837	3.5852
Improvement	+18.8%	+23.3%	+184.2%	+25.6%	+1.9%	+2.7%	+0.3%	+0.1%
Flux.1-dev / Pickscore								
Flow-GRPO	0.1382	68.746	0.0766	0.1578	0.3198	1.3497	0.8750	3.5882
Ours	0.1575	62.505	0.1059	0.1818	0.3266	1.3959	0.8779	3.6039
Improvement	+13.9%	+9.1%	+38.2%	+15.2%	+2.2%	+3.4%	+0.3%	+0.4%
SD3.5-M / HPSv3								
Flow-GRPO	0.1625	34.040	0.0971	0.1967	0.3309	1.3037	0.8445	3.5825
Ours	0.1851	29.820	0.1646	0.2103	0.3343	1.3239	0.8462	3.5894
Improvement	+13.9%	+12.4%	+69.4%	+6.9%	+1.0%	+1.5%	+0.2%	+0.2%

downscales the KL penalty, weakening regularization when diversity should be preserved. In later stages, small variance amplifies the penalty, excessively constraining high-frequency refinement and discouraging reward-driven improvements. This mismatch violates the desired diversity budget allocation, as it leads to insufficient regularization when global structures form and excessive restriction when only local details are refined.

Structure-aware Wasserstein Distance To address this imbalance, we propose a structure-aware regularization that replaces the KL term with a stage-dependent metric. Specifically, for the first K denoising steps, we apply a Wasserstein Distance constraint between the current and reference policies:

$$\mathcal{L}_{WD} = \frac{\|\bar{x}_{t+\Delta t, \theta} - \bar{x}_{t+\Delta t, \text{ref}}\|^2}{2} \quad (13)$$

which removes the variance denominator in the KL formulation. This modification yields a stronger constraint on early-stage updates, forcing the model to maintain semantic coverage and structural diversity across distinct modes. For later steps ($t > K$), we remove the regularization entirely, allowing the policy to freely optimize toward higher reward fidelity. Formally, the overall regularization is defined as:

$$\mathcal{L}_{\text{reg}}(t) = \begin{cases} \frac{\|\bar{x}_{t+\Delta t, \theta} - \bar{x}_{t+\Delta t, \text{ref}}\|^2}{2}, & t \leq K, \\ 0, & t > K. \end{cases} \quad (14)$$

4. Experiments

4.1. Experimental Setting

Implementation Details: We assess the ability of our approach to maintain generation diversity against baseline methods under two model architectures—SD3.5-M [9] and Flux.1-dev [4], and two preference reward functions (Pickscore [17] and HPSv3 [22]). For the Flux.1-dev architecture, we use 6 steps during training and 28 steps during evaluation, with a classifier-free guidance scale of 3.5. For the SD3.5-M architecture, we use 10 for training and 40 for evaluation, and a guidance scale of 4.5. We apply LoRA fine-tuning to all models, with LoRA rank $r = 32$, scaling factor $\alpha = 64$, learning rate 3×10^{-4} , and clip range 1×10^{-4} . We use gradient accumulation over $g = 12$ steps and a per-GPU batch size of 2. The exploration reward coefficient β is set to 0.7, and regularization coefficient K is set to 4.

Evaluation Metrics: We evaluate the generated images from two perspectives: diversity and quality. To quantify sample diversity, we use DreamSim [10], which measures perceptual similarity between image pairs; lower similarity indicates higher diversity. To further assess the degree of mode collapse, we compute SSIM [28], FID [14], and BeyondFID [8]. These metrics capture distributional deviations between images generated by the base model and those produced after training. Smaller deviations correspond to weaker collapse and better preservation of the orig-

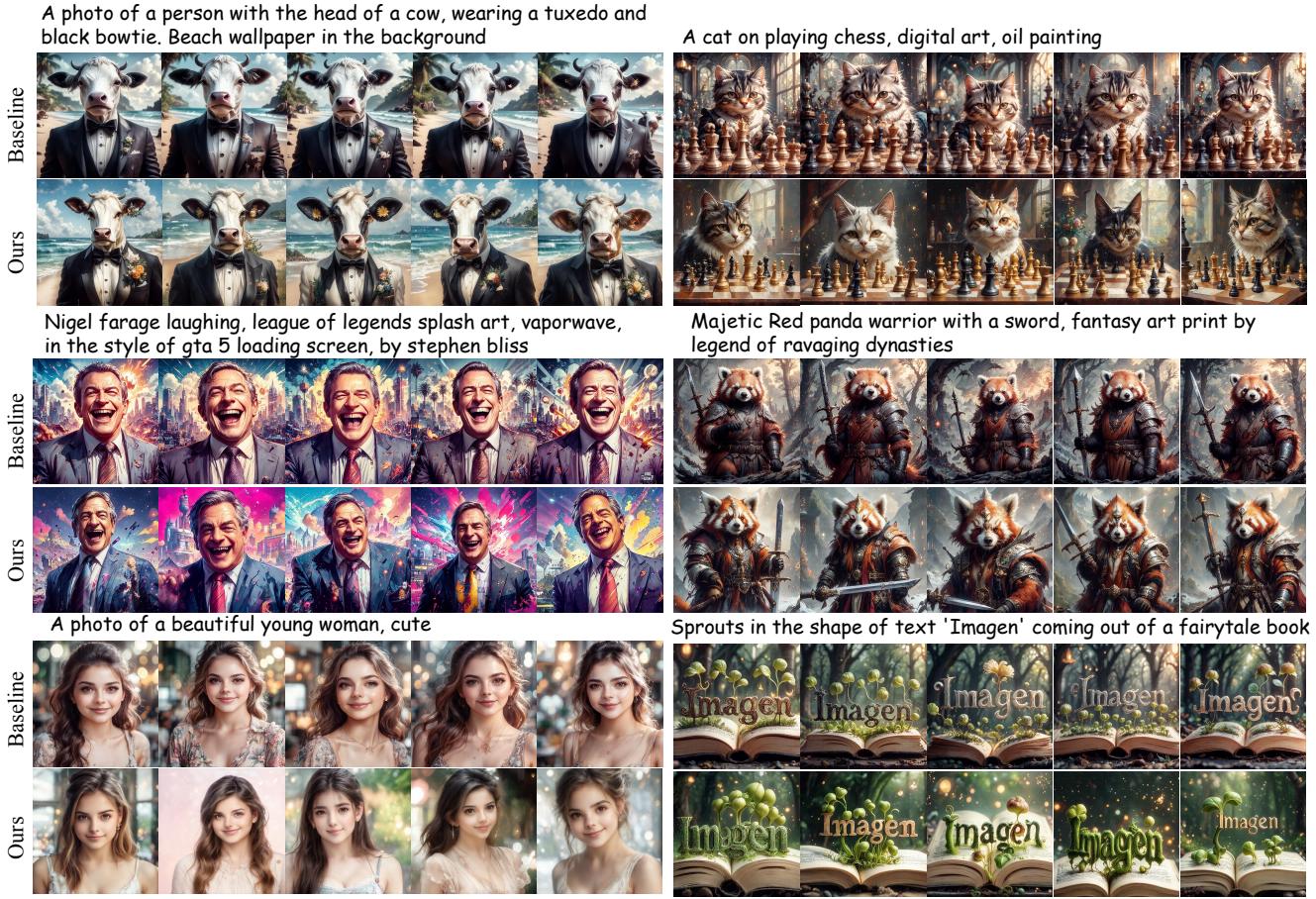


Figure 4. Qualitative experiments on diversity, the baseline method exhibits mode collapse in the generation of main subjects (such as facial features, poses, font colors, etc.), whereas our method achieves greater diversity and creativity while maintaining image quality and consistency with the captions.

inal generative distribution following [10, 20]. For image quality, we report CLIPScore [13] to evaluate text–image consistency, together with three human-preference-aligned reward models—PickScore [17], ImageReward [29], and UnifiedReward-Qwen [27]. These metrics collectively measure visual fidelity, semantic alignment, and human-perceived attractiveness.

4.2. Main Results

Quantitative results: We compare our method’s diversity to the baseline Flow-GRPO [21] (The KL term was omitted from the baseline because it significantly slows quality improvement, making a direct comparison of the Pareto fronts infeasible). As shown in Table 1, our approach consistently enhances all diversity metrics, with improvements of up to +171.4% in BeyondFID and +18.8% in DreamSim. This confirms that our reward promotes exploration of novel visual modes, preventing convergence to a few high-reward patterns. The results demonstrate that jointly optimizing for distribution-level exploration and structured generation leads to a more balanced reward optimization process.

Qualitative results: As shown in Fig. 4, we provide a visual comparison between Flow-GRPO and our method. Under the condition of comparable image quality, our approach demonstrates significantly stronger diversity while remaining faithful to the semantic constraints. The baseline results, on the other hand, exhibit clear signs of mode collapse, generating highly similar and repetitive outputs. For example, given prompts such as ‘A photo of a beautiful young woman, cute’, the baseline model produces images constrained to nearly identical facial expressions and viewpoints. In contrast, our method yields diverse stylistic interpretations while still satisfying semantic constraints. These visual comparisons conclusively show that our method effectively avoids mode collapse, leading to superior diversity across composition, stylistic details, and dynamic expression.

4.3. Ablation Study

Contribution of SA-Reg and creativity reward: We assess the individual contributions of the Structure-Aware Regularization (SA-Reg) and Creativity Reward modules.

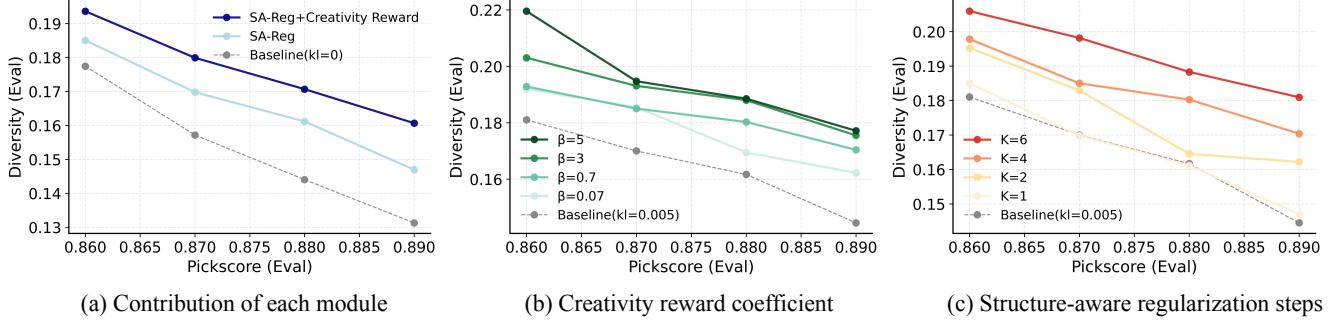


Figure 5. Ablation study on the Pareto front of quality and diversity for different modules and parameters.

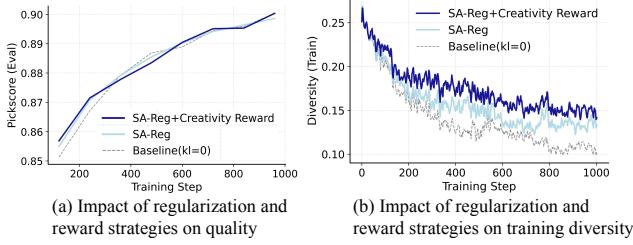


Figure 6. During the training process, DiverseGRPO achieves quality scores comparable to baseline methods, but exhibits a significantly slower decline in diversity.

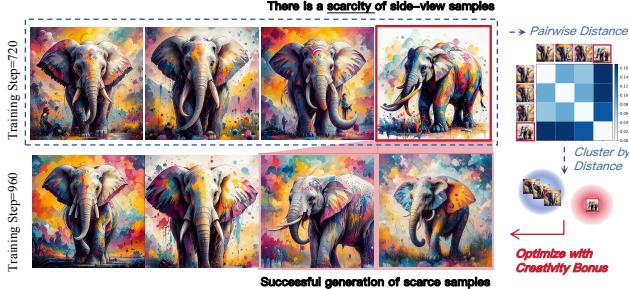


Figure 7. Sample visualization during the training process. Due to the bonus for innovative samples in DiverseGRPO, model can still generate diverse samples in the later stage of training, while it is difficult to mine innovative samples in the baseline (Fig. 2).

As shown in Fig. 5(a), the combination of both modules (SA-Reg + Creativity Reward) achieves the optimal balance between quality and diversity. Specifically, the SA-Reg module alone enhances diversity, but the addition of the creativity reward significantly boosts the diversity metric further. This suggests that the structure-aware regularization mitigates diversity degradation by encouraging the generation of diverse image patterns, while the creativity reward drives the model to explore even more diverse modes by incentivizing a broader range of semantic groupings. Fig. 6 shows the image quality and diversity during the training process. Our method exhibits a slower decline in diversity while maintaining image quality comparable to the baseline, demonstrating its effectiveness.

Creativity and SA-Reg coefficients: Figs. 5(b) and 5(c)

analyze the impact of the creativity reward coefficient β and the number of SA-Reg steps K , respectively. Increasing β enhances diversity, especially at $\beta=5$, by promoting exploration, but the diversity gain for $\beta=5$ over $\beta=3$ levels off in the later stages. This could indicate that, after a certain point, the model reaches a balance between exploration and exploitation. Similarly, raising K improves diversity through structured-aware regularization, yet higher steps increase computational cost with limited gains (details in appendix).

4.4. Impact of Exploration Bonus

Figure 7 shows the generated samples during the training process. At training step 720, a rare side-view elephant sample appears. Our method uses spectral clustering to identify these rare samples and assigns higher exploration rewards, allowing the model to continue generating such samples in the later stages of training while also improving the quality of the generated outputs. In contrast, the baseline method (as shown in Fig. 2) faces difficulty in maintaining sample diversity in the later stages of training, even when rare samples are generated. This is due to the fact that the reward model scores individual samples, which fails to recognize the distribution-level creative value, leading to low diversity in the generated samples during later stages of training.

5. Conclusion

We revisit the limitations of GRPO-based reinforcement learning in image generation and show that its conventional reward design and regularization scheme inevitably drive the model toward mode collapse in later training stages. Motivated by these, we introduce DiverseGRPO, which integrates a distributional creativity bonus and structure-aware regularization to recalibrate reward and align denoising-stage constraints with diversity preservation. Extensive experiments show DiverseGRPO substantially mitigates mode collapse, and yields a 13%~18% gain in semantic diversity without compromising visual quality, establishing a new Pareto frontier for image generation.

6. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant U24B6012, 62406167, and Kling Team, Kuaishou Technology.

References

- [1] Pietro Astolfi, Marlene Careil, Melissa Hall, Oscar Mañas, Matthew Muckley, Jakob Verbeek, Adriana Romero Sori-ano, and Michal Drozdzal. Consistency-diversity-realism pareto fronts of conditional image generative models. *arXiv preprint arXiv:2406.10429*, 2024. [3](#)
- [2] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. [2](#)
- [3] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. [2](#), [3](#)
- [4] Black Forest Labs. Flux: Official inference code for flux.1 models. <https://github.com/black-forest-labs/flux>, 2024. Version: commit hash or version number if available. [3](#), [6](#)
- [5] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémie Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023. [2](#)
- [6] Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025. [2](#), [3](#)
- [7] Li Ding, Jenny Zhang, Jeff Clune, Lee Spector, and Joel Lehman. Quality diversity through human feedback. In *Second Agent Learning in Open-Endedness Workshop*, 2023. [3](#)
- [8] Mischa Dombrowski, Weitong Zhang, Sarah Cechnicka, Hadrien Reynaud, and Bernhard Kainz. Image generation diversity issues and how to tame them. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3029–3039, 2025. [3](#), [6](#)
- [9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. [3](#), [6](#)
- [10] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. [2](#), [5](#), [6](#), [7](#)
- [11] Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, et al. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025. [2](#)
- [12] Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. *arXiv preprint arXiv:2508.04324*, 2025. [3](#), [5](#)
- [13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528, 2021. [7](#)
- [14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. [6](#)
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. [2](#)
- [16] Mete Ismayilzada, Antonio Lavergnac Jr, Simone A Lu-chini, Reet Patel, Antoine Bosselut, Lonneke van der Plas, and Roger Beaty. Creative preference optimization. *arXiv preprint arXiv:2505.14442*, 2025. [3](#)
- [17] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023. [3](#), [6](#), [7](#)
- [18] Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*, 2025. [3](#)
- [19] Junzhe Li, Yutao Cui, Tao Huang, Yingping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025. [3](#)
- [20] Yuming Li, Yikai Wang, Yuying Zhu, Zhongyu Zhao, Ming Lu, Qi She, and Shanghang Zhang. Branchgrpo: Stable and efficient grpo with structured branching in diffusion models. *arXiv preprint arXiv:2509.06040*, 2025. [3](#), [5](#), [7](#)
- [21] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. [2](#), [3](#), [5](#), [7](#)
- [22] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15086–15095, 2025. [3](#), [6](#)
- [23] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. [3](#)
- [24] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of*

- the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 3
- [25] Bin Xu Wang and John J. Vastola. Diffusion models generate images like painters: an analytical theory of outline first, details later. *arXiv preprint arXiv:2303.02490*, 2023. 2
 - [26] Feng Wang and Zihao Yu. Coefficients-preserving sampling for reinforcement learning with flow matching. *arXiv preprint arXiv:2509.05952*, 2025. 3
 - [27] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025. 7
 - [28] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6
 - [29] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imageword: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 7
 - [30] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grp on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. 2, 3
 - [31] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. 2