MDPI

Article

Flow-Multi: A Flow-Matching Multi-Reward Framework for Text-to-Image Generation

Jaegun Lee, Janghoon Choi *

Major in Data Science Convergence, Graduate School of Data Science, Kyungpook National University, Daegu 41566, Korea; leejken530@knu.ac.kr
* Correspondence: jhchoi09@knu.ac.kr

Abstract

Recent approaches in text-to-image (T2I) generation has actively adopted reinforcement learning (RL) techniques for human preference alignment. However, existing approaches primarily rely on a single reward function, which can lead to overfitting on specific metrics, resulting in issues such as reward hacking and imbalanced optimization among multiple objectives. To address this, we propose **Flow-Multi: A Flow-Matching Multi-Reward Framework for Text-to-Image Generation**. Our method builds upon Flow-matching-based Group Relative Policy Optimization (GRPO) learning. Each sample is evaluated by four reward models—Text-to-Image alignment, Human Preference, Aesthetic Quality, and GenEval—to create a multi-dimensional reward vector. We then utilize the Pareto dominance relationship to remove dominated samples and update the policy using only the non-dominated set. Additionally, we introduce advantage masking during training to suppress the contribution of low-reward samples, ensuring that only high-quality rewards are reflected in policy optimization. Experimental results demonstrate that Flow-Multi achieves balanced improvements across multiple reward criteria compared to the existing Flow-GRPO, validating the effectiveness of multi-reward reinforcement learning for stable alignment in text-to-image generation.

Keywords: flow matching; multi-reward reinforcement learning; text-to-image generation;

1. Introduction

Despite significant advances in text-to-image (T2I) generation [1,2,27], aligning model behavior with human preferences remains a problem with multiple challenges. Beyond mere prompt adherence, users implicitly value several orthogonal qualities—semantic faithfulness, aesthetic appeal, compositional correctness, and robustness on standardized vision-language tests. Optimizing one metric often degrades another, a classic case of "you get what you measure." As a result, single-reward pipelines tend to overfit, exhibiting reward hacking, distributional brittleness, or drift in unmeasured dimensions.

Recent work explores reinforcement learning (RL) for preference alignment in generative models [7,16,18,19,22], borrowing ideas from reinforcement learning from human feedback (RLHF) and policy optimization [12,15,16]. These approaches typically train with a single scalar reward (e.g., preference score, CLIP-based alignment, or a task-specific metric) [10,30,31]. However, collapsing a genuinely multi-objective problem into one dimension produces optimization pathology: (i) metric gaming—improving the chosen proxy while qualitatively regressing elsewhere; (ii) unstable updates—high-variance advantages

Received:
Revised:
Accepted:
Published:

Citation: Lee, J.; Choi, J. Flow-Multi: A Flow-Matching Multi-Reward Framework for Text-to-Image Generation. *Sensors* **2025**, *1*, 0. https://doi.org/

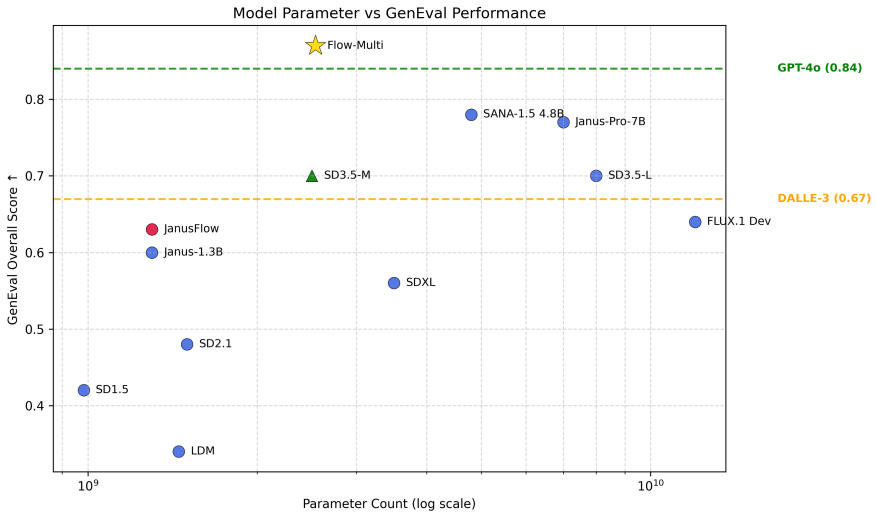


Figure 1. GenEval Performance. The Flow-Multi model demonstrates higher GenEval performance than GPT-4o and the baseline SD3.5-M, outperforming all other compared models.

driven by noisy scorers; and (iii) brittle generalization—improvements that fail to transfer across benchmarks such as DrawBench [27] or GenEval [26].

To address these challenges, we propose **Flow-Multi**, a flow-matching, multi-reward RL for T2I models. Building on Group Relative Policy Optimization (GRPO) in a flow-matching sampler, Flow-Multi evaluates each generated sample with a vector of rewards spanning four complementary dimensions: text–image alignment, human preference, aesthetic quality, and GenEval-style object/attribute correctness. Rather than averaging or hand-tuning weights, we treat the update as a multi-objective selection problem and use Pareto dominance to keep only non-dominated samples within each mini-batch. This preserves trade-offs explicitly: a sample is retained if improving any of its rewards would worsen at least one other reward. In addition, we introduce advantage masking, a simple but effective mechanism that zeroes low-quality advantages so that policy gradients are dominated by consistently good samples instead of noisy tails.

Our method is online, operating in the inner loop of sampling and learning, and it is metric-agnostic: reward heads can be swapped or extended without retooling the optimizer. Practically, the flow-matching backbone keeps training stable and sample-efficient, while mini-batch Pareto selection provides a principled, computation-light approximation to multi-objective policy improvement. The result is a more balanced performance profile across heterogeneous benchmarks, mitigating the typical "win one metric, lose two" failure mode of single-reward optimization.

2. Related Work

2.1. Diffusion and Flow-Matching Foundations

Diffusion models learn to reverse a Gaussian noising process and have become the dominant paradigm for high-fidelity image synthesis, with sampling implemented via discrete DDPM steps or continuous-time probability-flow ODE/SDE solvers [1–3]. Flow matching instead trains a continuous-time normalizing flow by directly matching the velocity field, enabling efficient deterministic sampling with far fewer steps while preserving quality [4,5]. Recent analyses connect diffusion and flow under a unified SDE/ODE view, clarifying when stochastic versus deterministic solvers are preferable and how model parameterizations transfer across the two families [15]. Building on this foundation, modern

T2I systems often adopt flow-matching backbones for superior speed–quality trade-offs in both image and video generation.

2.2. Reinforcement Learning for T2I Alignment

Preference alignment for diffusion/flow generators has followed multiple tracks. Policy-gradient style methods adapt RLHF ideas to T2I: DPO [8] directly fine-tunes diffusion policies from human preferences, while Diffusion-DPO [7]imports DPO as a simpler, value-free alternative to PPO-like RLHF [7]. Training-free alignment aims to steer preference at inference time without additional optimization [11]. More recently, group-relative policy optimization (GRPO [12])—a value-free policy gradient introduced for LLMs—has been instantiated on flow-matching generators: Flow-GRPO [15] integrates online RL into flows, and concurrent works (e.g., DanceGRPO[16], PREF-GRPO [35]) explore injecting stochasticity (converting the ODE to an SDE) or fitting pairwise preferences to stabilize advantages and mitigate reward hacking . While effective, single-reward training often exhibits metric gaming and trade-off regressions on unoptimized dimensions [20] (e.g., aesthetics vs. compositional correctness), motivating multi-reward formulations.

2.3. Multi-Objective Alignment and Reward Design

T2I post-training has been pursued via (1) direct scalar-reward fine-tuning [20], (2) reward-Backpropagation (AlignProp [10]), (3) DPO-style preference fitting [7,8,17], (4) PPO/GRPO policy gradients [9,12], and (5) training-free steering [11] . Many systems use linear scalarization of multiple rewards (e.g., CLIP alignment + aesthetics) with hand-tuned weights, as in Promptist [13] or differentiable reward formulations in DRaFT [19]. Parrot [14] explores multi-objective optimization by mapping language-encoded preference vectors to Pareto-optimal solutions, and model-averaging along the Pareto frontier has also been studied [22]. However, scalarization fixes a single operating point and is brittle when reward scales or reliabilities shift. Complementary to scalarization, process rewards seek finer credit assignment beyond sparse terminal signals: step-level PRMs and verification-guided supervision (ThinkPRM [34]) improve reasoning- or step-aware learning but carry substantial annotation/training cost . In diffusion/flow settings, step-aware preference models (e.g., LPO [23]) target both noisy and clean images; online surrogates (PRIME [32]) approximate process rewards from outcomes alone . Our perspective aligns with the latter: use multiple outcome rewards to reflect orthogonal desiderata (alignment, preference, aesthetics, compositionality) and select dominates samples via Pareto dominance, complemented by advantage masking to suppress low-quality gradients—thereby approximating multi-objective improvement without training explicit PRMs.

2.4. Benchmarks and Evaluation for T2I

Evaluation has evolved beyond perceptual fidelity to probe compositional reasoning, world knowledge, typography, and controllability. Datasets such as DrawBench [27] and GenEval [26] stress attribute binding, counts, spatial relations, and text rendering; newer suites (e.g., TIIF-Bench [28]) vary prompt length, typography, and style to test robustness under prompt perturbations, while UNIGENBENCH [29] expands coverage across fine-grained sub-criteria . Preference-oriented metrics (e.g., PickScore [25]), alignment metrics (e.g., CLIPScore [30]), and learned aesthetics predictors [31] complement task-specific pass/fail scores. Our experiments follow this trend, reporting multi-dimensional metrics (alignment, preference, aesthetics, GenEval correctness), which together reveal trade-offs that single metrics often conceal.

3. Method

3.1. Preliminaries

We adopt the Rectified Flow framework [36] to define a generative process interpolating between a clean data sample $x_0 \sim X_0$ and a noise sample $x_1 \sim \mathcal{N}(0, I)$ via a linear path $x_t = (1-t)x_0 + tx_1$. A Transformer-based velocity field $v_\theta(x_t, t)$ is trained to minimize the flow-matching objective:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],\, x_0\sim X_0,\, x_1\sim X_1}\Big[\, \big\|(x_1 - x_0) - v_\theta(x_t, t)\big\|_2^2 \,\Big]. \tag{1}$$

To leverage reinforcement learning (RL) for fine-tuning, we formulate the denoising process as a finite-horizon Markov Decision Process (MDP) where the state is $s_t = (x_t, t, c)$, the action corresponds to the state update $a_t = \Delta x_t$, and the policy is induced by the flow model. Unlike standard deterministic ODE samplers, we adopt a Stochastic Differential Equation (SDE) formulation to enable exploration and define a valid probability density for the policy. The transition follows the Euler-Maruyama update

$$x_{t+\Delta t} = x_t \;+\; \mu_\theta(x_t, t, c)\,\Delta t \;+\; \sigma_t\sqrt{\Delta t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{2}$$

Here, the drift is parameterized as

$$\mu_\theta(x_t, t, c) = v_\theta(x_t, t, c) \;+\; \frac{\sigma_t^2}{2t}\Big(x_t + (1-t)\, v_\theta(x_t, t, c)\Big), \tag{3}$$

to maintain consistency with the ODE marginals, where $\sigma_t = a\sqrt{t/(1-t)}$ controls the noise schedule. Consequently, the policy $\pi_\theta(\cdot|s_t)$ becomes a Gaussian distribution $\mathcal{N}(x_t + \mu_\theta\Delta t, \sigma_t^2\Delta t I)$, which facilitates gradient-based optimization.

To optimize this policy without a value function, we employ Group Relative Policy Optimization (GRPO). For a given prompt q_i, we sample a group of K trajectories $\{\tau_{i,k}\}_{k=1}^K$ using the current policy and compute rewards $R_{i,k}$. We then calculate the group-relative advantage

$$A_{i,k} = \frac{R_{i,k} - \bar{R}_i}{\max\big(\text{Std}[R_{i,\cdot}],\, \epsilon\big)}, \qquad \bar{R}_i = \frac{1}{K}\sum_{k=1}^{K} R_{i,k}, \tag{4}$$

to reduce variance. To construct the objective, we first define the likelihood ratio $\rho_{i,k} = \pi_\theta(o_{i,k} \mid q_i)/\pi_{\theta_{\text{old}}}(o_{i,k} \mid q_i)$ against the old policy. The PPO-style clipped surrogate loss is then formulated as

$$\mathcal{L}_{\text{clip}}(\theta) = \mathbb{E}_{i,k}\Big[\min\Big(\rho_{i,k}\, A_{i,k},\; \text{clip}\,\big(\rho_{i,k},\, 1-\epsilon_{\text{clip}},\, 1+\epsilon_{\text{clip}}\big)\, A_{i,k}\Big)\Big]. \tag{5}$$

In addition to the surrogate loss, we incorporate a KL divergence penalty to a frozen reference policy π_{ref}. While the general form relies on the density ratio $r(o) = \pi_{\text{ref}}/\pi_\theta$, our diffusion-based policy with a fixed covariance schedule allows for a closed-form stepwise calculation:

$$D_{\text{KL}}\big(\pi_\theta(\cdot \mid s_t) \;\|\; \pi_{\text{ref}}(\cdot \mid s_t)\big) = \frac{\|\bar{x}_{t+\Delta t,\theta} - \bar{x}_{t+\Delta t,\text{ref}}\|_2^2}{2\,\sigma_t^2\Delta t}, \tag{6}$$

where $\bar{x}_{t+\Delta t,\theta}$ and $\bar{x}_{t+\Delta t,\text{ref}}$ are the mean updates of the active and reference policies, respectively. This simplifies to a Euclidean distance between drifts, directly regularizing the velocity field. Finally, combining these terms, the policy is updated by maximizing the objective:

$$\max_\theta\; \mathcal{J}_{\text{GRPO}}(\theta) = \mathcal{L}_{\text{clip}}(\theta) \;-\; \mathbb{E}_i\big[\beta\, D_{\text{KL}}\big(\pi_\theta(\cdot \mid q_i) \;\|\; \pi_{\text{ref}}(\cdot \mid q_i)\big)\big]. \tag{7}$$

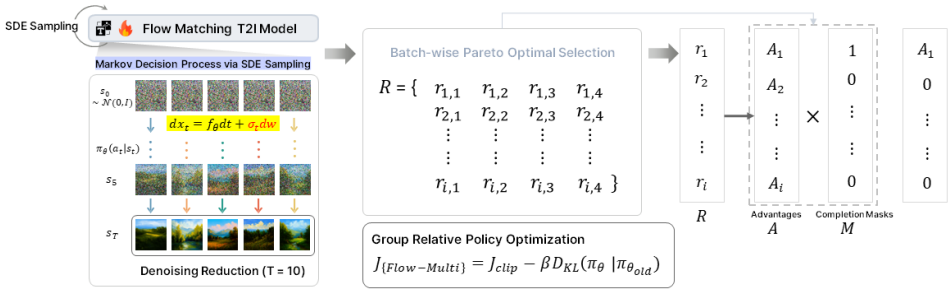


Figure 2. Overview of Flow-Multi, a multi-reward reinforcement learning framework incorporating Pareto optimization and advantage masking.

To implement this optimization efficiently, we utilize Low-Rank Adaptation (LoRA). Instead of updating the full weights W, we freeze the pre-trained parameters W_0 and train low-rank matrices A and B such that the effective weight becomes $W = W_0 + BA$. This approach allows us to adapt the heavy text-to-image backbone to complex reward signals with minimal computational overhead.

3.2. Prompt Dataset and Grouping

For the input prompts, we adopt the GenEval [26] prompt set as our training/evaluation source. Each record in the metadata (`train_metadata.jsonl`) specifies a natural-language prompt and structured constraints such as object classes, counts, colors, and spatial relations (e.g., `tag` $\in$ {`counting`, `colors`, `position`, `color_attr`}; fields `include/exclude` provide the target composition). We take the `prompt` string for generation and retain the structured fields to compute the *compositional* reward.

For each prompt q_i, we generate a group of $K=4$ images $\{x_{i,k}\}_{k=1}^{K}$ using our diffusion policy (Sec. 3.3). In our default setting, each epoch uses $P=48$ prompts (mini-batches), yielding $N = P \times K = 192$ images per epoch. Grouping by prompt is essential because both our advantage estimation and our Pareto selection are done within the set $\{x_{i,k}\}_{k=1}^{K}$ for a fixed q_i.

3.3. Flow-Matching Sampling with GRPO

We use a Stable Diffusion-3.5-medium model and perform sampling with a flow-matching scheduler. Unless otherwise noted, we use $S = 10$ sampling steps for training (exploration), $S_{\text{eval}} = 40$ for evaluation, classifier-free guidance scale $= 4.5$, and 512×512 resolution. For a prompt q_i, the policy $\pi_\theta(\cdot \mid q_i)$ induces a per-step conditional Gaussian transition; we log the per-step log-likelihood ratio to construct the PPO-style clipped objective in GRPO (Sec. 3.7).

3.4. Reward Suite

Each generated image $x_{i,k}$ is evaluated by four metrics, which together form a 4D reward vector

$$R(x_{i,k}) = \left[r_{i,k}^{\text{aesth}},\ r_{i,k}^{\text{pref}},\ r_{i,k}^{\text{align}},\ r_{i,k}^{\text{comp}}\right] \in \mathbb{R}^4.$$

Concretely,

- **Aesthetic Quality** (r^{aesth}): a learned aesthetic predictor score.
- **Human Preference** (r^{pref}): PickScore (CLIP ViT-H-14 with preference head).
- **Text–Image Alignment** (r^{align}): CLIPScore between the prompt and the image.
- **Compositional Compliance** (r^{comp}): GenEval-style object/attribute/position/count correctness, computed using the metadata (`include/exclude`) with a pretrained detector.

Stacking $R(x_{i,k})$ over all N images yields a matrix $R \in \mathbb{R}^{N\times 4}$ per epoch.

3.5. Group-wise Pareto Non-Dominated Selection

Within each prompt group i, we consider the K reward vectors $\{R(x_{i,k})\}_{k=1}^{K}$. A sample $x_{i,a}$ **dominates** $x_{i,b}$ (denoted $x_{i,b} \prec x_{i,a}$) if it is **no worse** on all metrics and strictly better on at least one:

$$R(x_{i,a}) \succeq R(x_{i,b}) \quad \Longleftrightarrow \quad \big(R_d(x_{i,a}) \geq R_d(x_{i,b})\ \forall d\big)\ \wedge\ \big(\exists d:\ R_d(x_{i,a}) > R_d(x_{i,b})\big),$$

We mark each non-dominated sample with a binary mask

$$M_{i,k} = \begin{cases} 1, & \text{if } x_{i,k} \text{ is non-dominated within group } i, \\ 0, & \text{otherwise.} \end{cases}$$

Only the survivors $\{k:\ M_{i,k} = 1\}$ enter the subsequent statistics and policy update. This *batchwise* Pareto filtering preserves diverse trade-offs among the four objectives while removing samples strictly inferior to others under the same prompt.

3.6. Scalarization and Survivor-Only Advantage (as implemented)

After group-wise Pareto filtering (Sec. 3.5), we form a single scalar reward per image by a fixed weighted sum of the four metrics.This filtering step is crucial; unlike naive scalarization which might favor samples that exploit a single reward while degrading others (reward hacking), Pareto selection ensures that only solutions offering valid trade-offs across all objectives contribute to the policy update. Let $w = (w^{\text{aesth}}, w^{\text{pref}}, w^{\text{align}}, w^{\text{comp}})$ be the weights from configuration (`config.reward_fn`); by default $w_d = \frac{1}{4}$. For each image $x_{i,k}$ we define

$$r_{i,k}^{\text{comb}} \ = \ \sum_{d=1}^{4} w^{(d)}\, r_{i,k}^{(d)},$$

which matches the `avg` in our implementation.

Let $M_{i,k} \in \{0,1\}$ denote the Pareto non-dominance mask within each prompt group (Sec. 3.5), and let $\mathcal{S} = \{(i,k) : M_{i,k} = 1 \wedge \texttt{valid}(i,k)\}$ be the set of surviving, valid samples across the whole distributed mini-epoch (all processes). If $\mathcal{S}$ is empty, we fall back to a single top-1 valid sample by r^{comb} to avoid degeneracy.

We compute global survivors-only statistics

$$\mu_{\mathcal{S}} \ = \ \frac{1}{|\mathcal{S}|} \sum_{(i,k)\in\mathcal{S}} r_{i,k}^{\text{comb}}, \qquad \sigma_{\mathcal{S}} \ = \ \max\Big(\text{Std}\{r_{i,k}^{\text{comb}} : (i,k) \in \mathcal{S}\},\ \epsilon\Big),$$

and define the advantage

$$A_{i,k} \ = \ \frac{r_{i,k}^{\text{comb}} - \mu_{\mathcal{S}}}{\sigma_{\mathcal{S}}} \ \cdot\ M_{i,k}.$$

Thus dominated or invalid samples ($M_{i,k} = 0$) contribute zero advantage. We do *not* compute per-step advantages. For terminal-only rewards, we use a single scalar $A_{i,k}$ per image and apply it to all stepwise log-prob terms; equivalently, we set $A_{i,k,t} := A_{i,k}$ for notation/shape matching only.

Algorithm 1 Flow-Multi with Pareto Masking (mini-batch)

Require:
π_θ : Online policy model (initialized from pretrained weights)
π_{ref} : Reference model
$\mathcal{Q}$:prompt batch
K : group size (default 4)
$\sigma_t = a\sqrt{t/(1-t)}$: noise schedule
μ_θ (Eq.(3): drift ; β: KL weight ; LoRA rank r; reward weights $w \in \mathbb{R}^4$ (default $w_d{=}\frac{1}{4}$)

Ensure: updated policy π_θ (LoRA-only updates)

Initialize policy π_θ (LoRA on Q/K/V/O); set reference $\pi_{\text{ref}} \leftarrow \text{stopgrad}(\pi_\theta)$
for epoch $= 1, \ldots$ **do**
  Sample a mini-batch of prompts $\{q_i\}_{i=1}^{P}$ from $\mathcal{Q}$
  for each prompt q_i **do**
    Sampling (SDE rollout): generate K images $\{x_{i,k}\}_{k=1}^{K}$ using (2)
    Reward vector for each $x_{i,k}$: $R(x_{i,k}) = (r_{\text{aesth}}, r_{\text{pref}}, r_{\text{align}}, r_{\text{comp}})$
    Group-wise Pareto non-dominated mask $M_{i,k} \in \{0,1\}$ (§3.5)
    Scalarization $r_{i,k}^{\text{comb}} = \sum_{d=1}^{4} w^{(d)} r_{i,k}^{(d)}$ (§3.6)
  end for
  Gather survivors $S = \{(i,k) : M_{i,k} = 1 \wedge \text{valid}(i,k)\}$ across devices; if $S = \emptyset$ then fallback to top-1 by r^{comb}
  Compute survivors-only stats μ_S,σ_S
  Advantage (masked) $A_{i,k} = \frac{r_{i,k}^{\text{comb}} - \mu_S}{\sigma_S} \cdot M_{i,k}$
  Compute likelihood ratios $\rho_{i,k} = \frac{\pi_\theta(o_{i,k}|q_i)}{\pi_{\theta_{\text{old}}}(o_{i,k}|q_i)}$
  PPO-style objective with Pareto mask (Eq. (8), §3.7)
  KL penalty to reference (stepwise or sequence): (Eq. (9), §3.7)
  Update $\theta \leftarrow \theta + \eta \nabla_\theta J_{\text{GRPO}}(\theta)$ ▷ LoRA params only
  Logging: Pareto keep ratio $\left(\sum_{i,k} M_{i,k}/(P \cdot K)\right)$, each reward mean/var, A stats, KL, loss terms, learning curves
  Periodically refresh reference $\pi_{\text{ref}} \leftarrow \text{stopgrad}(\pi_\theta)$
end for

3.7. GRPO with Pareto Masking

Let $\rho_{i,k} = \frac{\pi_\theta(o_{i,k}|q_i)}{\pi_{\theta_{\text{old}}}(o_{i,k}|q_i)}$ be the likelihood ratio for the sampled trajectory/action. The PPO-style clipped surrogate with Pareto masking is

$$\mathcal{L}_{\text{clip}}^{\text{mask}}(\theta) = \mathbb{E}_i\left[\frac{1}{\sum_k M_{i,k}} \sum_{k=1}^{K} \min\Big(\rho_{i,k}\, A_{i,k}\, M_{i,k},\ \text{clip}(\rho_{i,k}, 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}})\, A_{i,k}\, M_{i,k}\Big)\right]. \tag{8}$$

We also include a trust-region penalty toward a frozen reference policy π_{ref}:

$$\max_\theta \ \ \mathcal{J}_{\text{GRPO}}(\theta) = \mathcal{L}_{\text{clip}}^{\text{mask}}(\theta) \ - \ \mathbb{E}_i\big[\beta\, D_{\text{KL}}\big(\pi_\theta(\cdot \mid q_i) \,\|\, \pi_{\text{ref}}(\cdot \mid q_i)\big)\big], \tag{9}$$

where D_{KL} is computed from log-likelihoods (or, under equal-covariance Gaussian steps, via a per-step closed form). In practice, we ascend on $\mathcal{J}_{\text{GRPO}}$ with ϵ_{clip} and β tuned to balance stability and exploration.

Implementation notes.

At each epoch we (i) sample $N = 192$ images (48 prompts $\times$ 4 images), (ii) compute the four rewards and build $R \in \mathbb{R}^{N\times 4}$, (iii) apply *groupwise* Pareto to obtain $\{M_{i,k}\}$, (iv) compute survivor-only normalized advantages, and (v) update the policy with the masked GRPO objective (8)–(9). The *Pareto keep ratio* ($\sum_{i,k} M_{i,k}/N$) is tracked to monitor how selective the filtering is during training.

4. Experiments

Table 1. GenEval image generation benchmark. Best and second-best scores are highlighted in blue and green, respectively. All results are obtained from our own experiments under a unified evaluation environment and identical hyperparameter settings, *except* for DALLE-2, DALLE-3, and GPT-4o, which are reported from official sources. For **Flow-GRPO**, we report the reproduced results under our experimental setup using the same hyperparameters as the original paper, which differ from the originally reported score (0.92 Overall).

Model	Overall	Single Obj.	Two Obj.	Counting	Colors	Position	Attr. Binding
Diffusion Models							
LDM	0.34	0.90	0.26	0.19	0.66	0.01	0.04
SD1.5	0.42	0.95	0.37	0.35	0.75	0.03	0.06
SD2.1	0.48	0.98	0.45	0.40	0.82	0.08	0.14
SD-XL	0.56	0.98	0.75	0.43	0.88	0.13	0.21
DALLE-2	0.52	0.94	0.66	0.49	0.77	0.10	0.19
DALLE-3	0.67	0.96	0.87	0.47	0.83	0.43	0.45
Autoregressive Models							
Janus-1.3B	0.60	0.96	0.62	0.27	0.82	0.48	0.46
JanusFlow	0.63	0.97	0.55	0.43	0.85	0.55	0.42
Janus-Pro-7B	0.77	0.97	0.87	0.55	0.88	0.73	0.62
GPT-4o	0.84	0.99	0.92	0.85	0.92	0.75	0.61
Flow Matching Models							
FLUX.1 Dev	0.64	0.98	0.79	0.75	0.77	0.20	0.39
SD3.5-L	0.70	0.99	0.89	0.69	0.82	0.27	0.55
SANA-1.5 4.8B	0.78	0.99	0.91	0.80	0.87	0.60	0.54
SD3.5-M	0.70	0.99	0.89	0.69	0.82	0.27	0.55
Flow-GRPO†	0.72	1.00	0.90	0.73	0.82	0.30	0.59
Flow-Multi(Ours)	0.87	0.99	0.99	0.88	0.81	0.80	0.80

Table 2. Task performance and image-quality results on compositional generation, text rendering, and human-preference benchmarks, evaluated using CLIPScore, ImageReward, and UnifiedReward.

Model	Task Metric		Image Quality		Preference Score		
	GenEval	CLIP	Aesthetic	DeQA	ImgRwd	PickScore	UniRwd
Diffusion Models							
LDM	0.34	0.19	5.06	3.35	-1.95	19.17	1.36
SD1.5	0.42	0.22	5.21	3.58	-1.48	19.78	1.58
SD2.1	0.48	0.24	5.45	3.66	-0.73	20.55	1.98
SD-XL	0.56	0.26	5.88	3.92	0.15	21.78	2.61
Autoregressive Models							
Janus-1.3B	0.60	0.21	5.28	2.87	-1.08	19.86	1.50
JanusPro-7B	0.77	0.27	5.99	3.51	0.77	22.00	2.67
Flow matching Models							
FLUX.1 Dev	0.64	0.27	6.26	4.35	1.03	22.87	3.34
SD3.5-L	0.70	0.28	5.96	4.18	0.95	22.77	3.24
SANA-1.5 4.8B	0.78	0.27	6.15	4.07	1.06	22.73	3.21
SD3.5-M	0.70	0.28	5.80	4.14	0.77	22.32	3.02
Flow-GRPO	0.72	0.28	5.63	3.93	0.73	22.16	3.03
Flow-Multi	0.87	0.28	5.86	4.15	0.76	22.23	2.95
Δ	+0.15	0.00	+0.23	+0.22	+0.03	+0.07	-0.08

Δ denotes the performance difference between **Flow-Multi** and **Flow-GRPO**.

Table 1 presents the GenEval image generation results. Our method, **Flow-Multi**, substantially improves upon the base SD3.5-M model, raising the overall score from 0.70 to 0.87 (approximately +26%). Under a unified evaluation environment, Flow-Multi also outperforms the reproduced Flow-GRPO baseline (0.87 vs. 0.72 Overall), while achieving consistently strong performance across all sub-tasks. In particular, Flow-Multi attains

near-perfect accuracy on multi-object composition (Two Obj.: 0.99), significant gains in counting (0.88), spatial reasoning (Position: 0.80), and attribute binding (0.80), demonstrating balanced improvements across diverse objectives. These results highlight the effectiveness of multi-reward reinforcement learning in enhancing text-to-image alignment beyond single-objective optimization.

As shown in Table 2, low-Multi achieves a more balanced trade-off between text–image alignment and perceptual image quality compared to Flow-GRPO.While Flow-GRPO improves alignment performance over the SD3.5-M baseline, this gain is accompanied by a noticeable degradation in visual quality, with its Aesthetic score decreasing from 5.80 (SD3.5-M) to 5.63. This result suggests that optimizing a single alignment-oriented reward can adversely affect perceptual fidelity.

In contrast, Flow-Multi not only substantially improves compositional understanding, achieving a higher GenEval score (0.87 vs. 0.72), but also preserves—and in some cases enhances—image quality. Specifically, Flow-Multi recovers the Aesthetic score to 5.86 and improves DeQA to 4.15, outperforming Flow-GRPO by **+0.23** and **+0.22**, respectively, as highlighted in the Δ row. These results demonstrate that the proposed multi-reward optimization effectively mitigates the common trade-off between alignment and perceptual quality, enabling robust improvements without sacrificing visual realism.

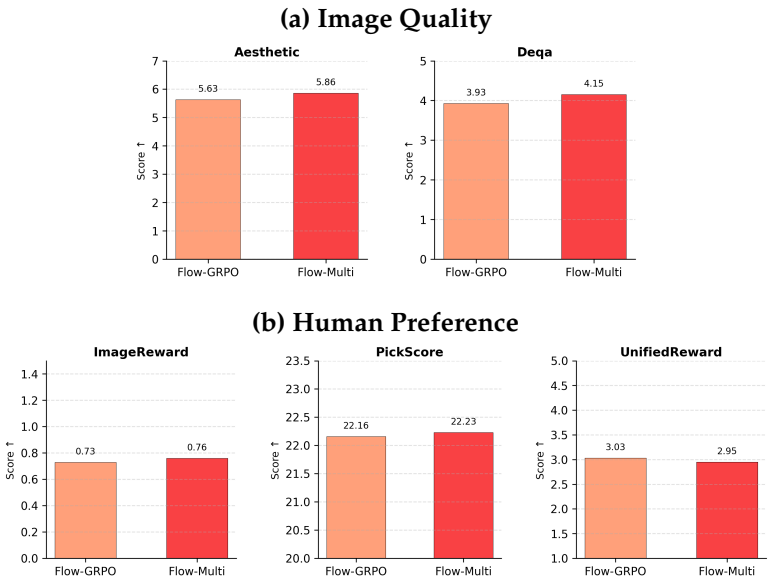


Figure 3. (a) Image quality metrics (Aesthetic, DeQA) remain stable across models. **(b)** Human preference metrics (ImageReward, PickScore, UnifiedReward) improve after Flow fine-tuning.

Figure 3 compares the perceptual quality and human preference metrics of Flow-GRPO and Flow-Multi. As observed in (a), Flow-GRPO tends to sacrifice image quality (lower Aesthetic and DeQA scores) in exchange for strict text alignment. In contrast, Flow-Multi effectively mitigates this trade-off, achieving higher Aesthetic and DeQA scores. This improvement extends to (b) Human Preference metrics; Flow-Multi outperforms Flow-GRPO in both ImageReward and PickScore. Although the UnifiedReward is comparable, the consistent gains in aesthetic and preference scores confirm that Flow-Multi generates more visually appealing images while maintaining alignment.

Fig. 4. Qualitative comparison on compositional prompts. Flow-Multi accurately aligns multiple objects and reliably handles counting prompts (“two snowboards"), generating the correct number of distinct, non-overlapping instances. It also preserves object identity in challenging cases such as zebras and giraffes, and achieves strong color–attribute consistency , showing improved compositional fidelity over prior SD-based models

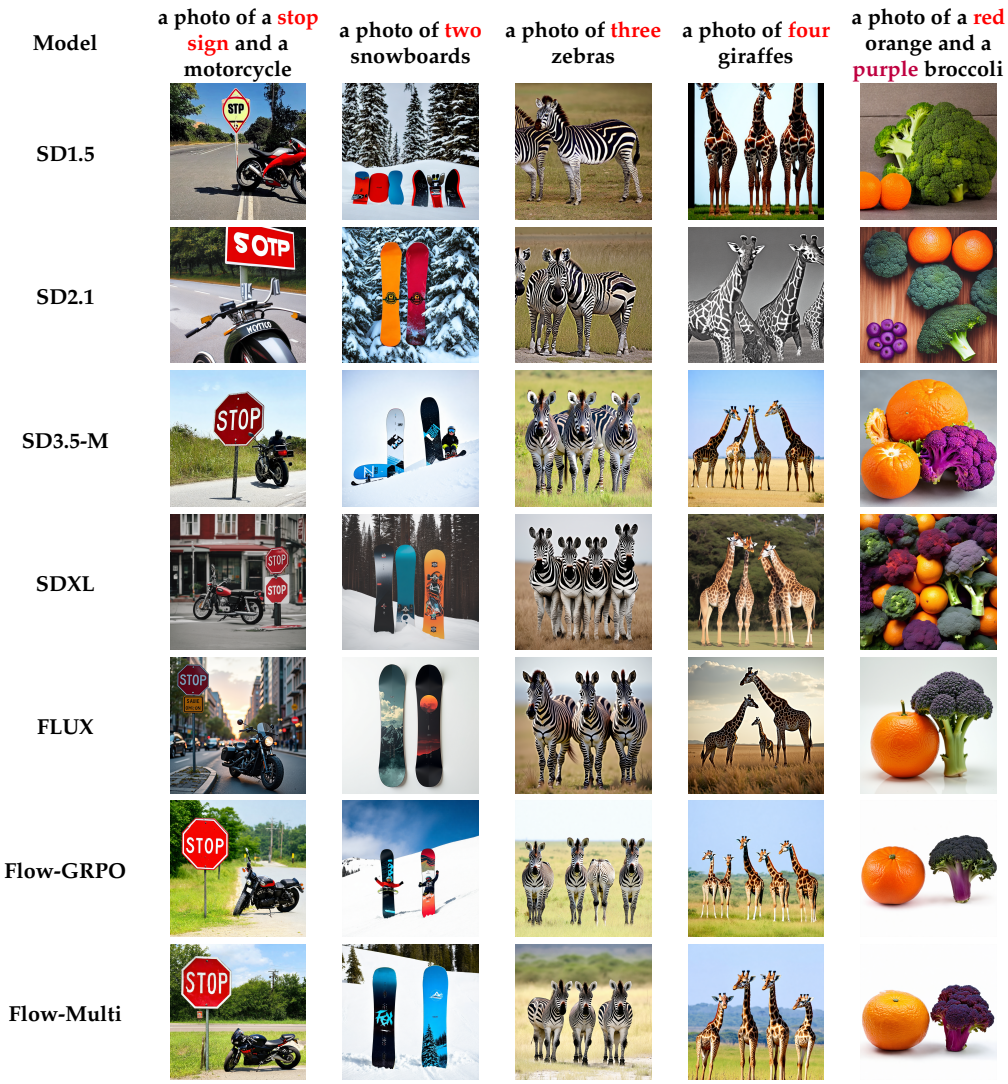
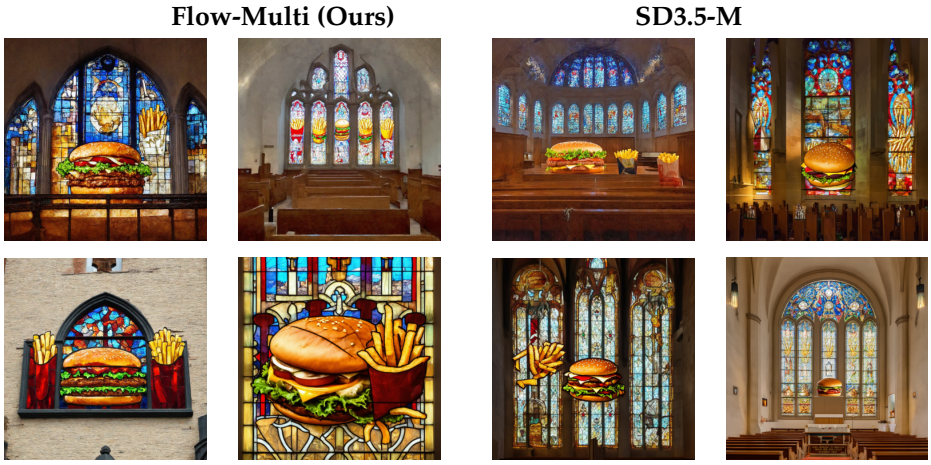


Figure 4. Compared to other models, Flow-Multi outperforms SD-based baselines in color accuracy, counting, and aesthetic quality.



Prompt: "A church with stained glass windows depicting a hamburger and french fries"

Figure 5. Qualitative comparison between Flow-Multi and SD3.5-M on DrawBench. Each prompt block shows four sampled generations from each model. Flow-Multi demonstrates improved text–image alignment and compositional consistency, while preserving visual fidelity across diverse samples.

5. Conclusion

This paper identified fundamental limitations of conventional reinforcement learning (RL) for text-to-image (T2I) generation, where collapsing multi-dimensional human preferences into a single scalar reward leads to brittle generalization and misalignment. Such reward aggregation obscures inherent trade-offs among quality dimensions, resulting in unstable optimization and biased alignment.

To address this issue, we proposed Flow-Multi, a multi-reward RL framework based on flow-matching GRPO. Flow-Multi combines vector-valued rewards, group-wise Pareto non-dominated sample selection, and advantage masking, enabling stable policy updates and balanced optimization across multiple objectives without heuristic tuning.

Experiments on the GenEval benchmark demonstrate clear performance gains: Flow-Multi improves SD3.5-M from 0.70 to 0.87, outperforming the single-reward Flow-GRPO baseline (0.72). In addition to higher aggregate scores, Flow-Multi achieves consistent improvements in spatial reasoning and attribute binding, with human evaluations confirming superior perceptual alignment.

Overall, this work establishes a principled framework for multi-objective alignment in generative modeling. By reframing T2I alignment as a multi-objective optimization problem, Flow-Multi provides a scalable alternative to single-reward RL and can offers a strong baseline for aligning future large-scale multimodal generative models.

References

1. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *Adv. Neural Inf. Process. Syst.* **2020** ,
2. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May **2021**.
3. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May **2021**. doi:10.48550/arXiv.2011.13456.
4. Lipman, Y.; Chen, R.T.Q.; Ben-Hamu, H.; Nickel, M.; Le, M. Flow Matching for Generative Modeling. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May **2023**.
5. Albergo, M.S.; Vanden-Eijnden, E. Building Normalizing Flows with Stochastic Interpolants. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May **2023**.
6. Liu, X.; Gong, C.; Liu, Q. Flow Straight and Fast: Learning to Generate with Rectified Flow. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 28 November–9 December **2022**.
7. Wallace, B.; Chen, S.; Roberts, D.A.; Brooks, T.; Efros, A.A.; Kanazawa, A.; Owens, A. Diffusion Model Alignment Using Direct Preference Optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June **2024**
8. Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C.D.; Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Adv. Neural Inf. Process. Syst.* **2023**
9. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**
10. Prabhudesai, M.; Goyal, A.; Pathak, D.; Fragkiadaki, K. Aligning Text-to-Image Diffusion Models with Reward Backpropagation. *arXiv* **2023**
11. Xie, Z.; Gong, B.; et al. DyMO: Training-free Preference Alignment for Diffusion Models at Inference Time. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), **2025**.
12. DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv* **2025**
13. Hao, Y.; Chi, Z.; Dong, L.; Wei, F. Optimizing Prompts for Text-to-Image Generation. *Adv. Neural Inf. Process. Syst.* **2023**, *36*
14. Lee, S.H.; Li, Y.; Ke, J.; Yoo, I.; Zhang, H.; Yu, J.; Wang, Q.; Deng, F.; Entis, G.; He, J.; et al. Parrot: Pareto-optimal Multi-Reward Reinforcement Learning Framework for Text-to-Image Generation. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October **2024**.
15. Liu, J.; Liu, G.; Liang, J.; Li, Y.; Liu, J.; Wang, X.; Wan, P.; et al. Flow-GRPO: Training Flow Matching Models via Online RL. *arXiv* **2025**,
16. Xue, Y.; Xu, K.; et al. DanceGRPO: Group Relative Policy Optimization for Flow Matching Models. *arXiv* **2025**
17. Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Adv. Neural Inf. Process. Syst.* **2023**, *36*.
18. Fan, Y.; Watkins, O.; Du, Y.; Liu, H.; Ryu, M.; Boutilier, C.; Abbeel, P.; Ghavamzadeh, M.; Lee, K.; Lee, K. DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*.
19. Clark, K.; Vicol, P.; Swersky, K.; Fleet, D.J. Directly Fine-Tuning Diffusion Models on Differentiable Rewards. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May **2024**.
20. Black, K.; Janner, M.; Du, Y.; Kostrikov, I.; Levine, S. Training Diffusion Models with Reinforcement Learning. In Proceedings of the ICML Workshop on Structured Probabilistic Inference and Generative Modeling, Honolulu, HI, USA, 23–29 July **2023**.

21. Hao, Y.; Chi, Z.; Dong, L.; Wei, F. Optimizing Prompts for Text-to-Image Generation. *Adv. Neural Inf. Process. Syst.* **2023**, *36*.
22. Ramé, A.; Couairon, G.; Shukor, M.; Dancette, C.; Gaya, J.B.; Soulier, L.; Cord, M. Rewarded Soups: Towards Pareto-Optimal Alignment by Interpolating Weights Fine-Tuned on Diverse Rewards. *Adv. Neural Inf. Process. Syst.* **2023**, *36*.
23. Zhang, T.; Da, C.; Ding, K.; Yang, H.; Jin, K.; Li, Y.; Gao, T.; Zhang, D.; Xiang, S.; Pan, C. Diffusion Model as a Noise-Aware Latent Reward Model for Step-Level Preference Optimization. *Adv. Neural Inf. Process. Syst.* **2025**, *38*.
24. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the International Conference on Machine Learning (ICML), Virtual Event, 18–24 July **2021**;
25. Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Shahbuland; Penna, J.; Levy, O. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. *Adv. Neural Inf. Process. Syst. Datasets Benchmarks* **2023**, *2*.
26. Ghosh, D.; Hajishirzi, H.; Schmidt, L. GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment. *Adv. Neural Inf. Process. Syst. Datasets Benchmarks* **2023**, *2*.
27. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S.K.S.; Ayan, B.K.; Mahdavi, S.S.; Lopes, R.G.; et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *Adv. Neural Inf. Process. Syst.* **2022**, *35*
28. Wei, X.; Zhang, J.; Wang, Z.; Wei, H.; Guo, Z.; Zhang, L. TIIF-Bench: How Does Your T2I Model Follow Your Instructions? *arXiv* **2025**
29. Wang, Y.; Zang, Y.; Li, H.; Jin, C.; Wang, J. Unified Reward Model for Multimodal Understanding and Generation. *arXiv* **2025**,
30. Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; Choi, Y. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic, 7–11 November **2021**;
31. Liang, Z.; Yuan, Y.; Gu, S.; Chen, B.; Hang, T.; Cheng, M.; Li, J.; Zheng, L. Aesthetic Post-Training Diffusion Models from Generic Preferences with Step-by-step Preference Optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), **2025**;
32. Cui, Y.; Li, J.; Huang, T.; Ma, Y.; Fan, C.; et al. PRIME: Process Reward via Implicit Modeling from Outcome Labels. *arXiv* **2025**,
33. Wang, P.; Li, L.; Shao, Z.; Xu, R.; Dai, D.; Li, Y.; Chen, D.; Wu, Y.; Sui, Z. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *CoRR* **2023**
34. Khalifa, M.; Agarwal, R.; Logeswaran, L.; Kim, J.; Peng, H.; Lee, M.; Lee, H.; Wang, L. Process Reward Models That Think. *arXiv* **2025**,
35. Wang, Y.; Li, Z.; Zang, Y.; Zhou, Y.; Bu, J.; Wang, C.; Lu, Q.; Jin, C.; Wang, J. Pref-GRPO: Pairwise Preference Reward-based GRPO for Stable Text-to-Image Reinforcement Learning. *arXiv* **2025**,
36. Liu, X.; Gong, C.; Liu, Q. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May **2023**.
37. Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In Proceedings of the International Conference on Machine Learning (ICML), Vienna, Austria, 21–27 July **2024**;
38. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 25–29 April **2022**.