DivRL: Disentangled Self-Similarity Rewards for Diverse Subject-Driven Generation

Qian Wang, Zhenyu Li, Abdelrahman Eldesokey, and Peter Wonka

KAUST, Saudi Arabia
{first.last}@kaust.edu.sa

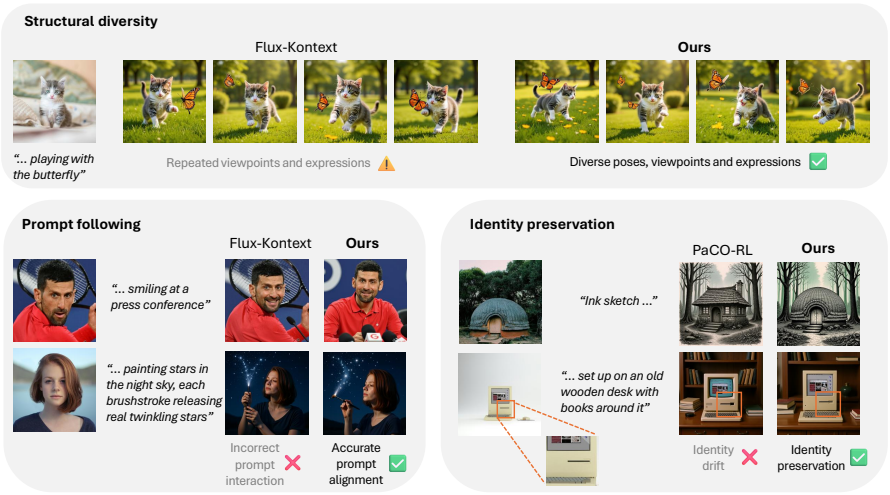

Fig. 1: Our method achieves a better balance between structural diversity, prompt following, and identity consistency compared with strong baselines.

Abstract. Subject-driven image generation faces an "Identity-Diversity Paradox", where strong identity preservation often leads to rigid and low-diversity outputs. We propose a post-training framework called *DivRL* that jointly optimizes identity consistency and structural diversity simultaneously by leveraging disentangled visual features from a robust similarity model. Specifically, we introduce a Negative Self-Similarity Measure (nSSM) to quantify structural diversity, and Visual Semantic Matching (VSM) to evaluate identity consistency. We propose an "Explore-and-Suppress" strategy that treats VSM as a gated constraint: the model freely explores structurally diverse configurations, and only samples that violate the identity threshold are penalized via a quadratic hinge loss. This converts identity preservation from a competing objective into a feasibility constraint, allowing nSSM and VSM to improve jointly. Experiments demonstrate that our method effectively pushes the model

to generate both consistent and diverse images and improves structural diversity while maintaining comparable identity consistency through a gated optimization formulation.

Keywords: Subject-driven Image Generation · Diffusion Models · Reinforcement Learning · GRPO

1 Introduction

Subject-driven image generation aims to synthesize novel images of a specific subject while preserving its visual identity under diverse contexts, poses, and styles. Recent diffusion-based models have demonstrated remarkable fidelity in reproducing reference subjects, enabling applications such as personalized image generation and identity-preserving editing [13, 33, 34, 38]. However, these models often struggle to balance identity consistency with structural diversity, while adhering to the textual prompt as illustrated in 1. Models that strongly enforce identity preservation frequently collapse to near-duplicate images of the reference, producing outputs with similar poses, viewpoints, or expressions. Conversely, methods that encourage diversity often drift away from the subject identity. This tension between preserving identity and enabling structural variation constitutes what we refer to as the "Identity–Diversity Paradox".

A key reason behind this paradox lies in how identity is represented in a reference image. Identity is defined by a subset of visual attributes, but these attributes are entangled with spatial structure in the reference image. When models attempt to preserve identity, they often also preserve the entangled structural information, which reduces diversity. Conversely, when models attempt to generate images with different structures, they risk modifying identity-defining attributes, leading to identity drift.

Recent work [4] shows that diffusion backbones contain visual features that are more robust to spatial transformations by disentangling visual and semantic representations. These features provide a reliable signal for evaluating identity consistency even under pose or viewpoint changes. However, optimizing identity consistency using [4] alone remains insufficient: models can still exploit the reward by reproducing the same spatial configuration as the reference image. In other words, identity-based rewards may still encourage structural mimicry by favoring configurations that closely resemble the reference image. Ideally, we would like the model to preserve identity consistency without sacrificing structural diversity.

To achieve this, we propose a method called *DivRL* for measuring structural diversity through the internal relationships of visual features rather than global feature differences. Specifically, we introduce the *negative Self-Similarity Measure (nSSM)*, which evaluates diversity by comparing the self-similarity matrices of the disentangled visual feature grids extracted from the reference and generated images. Self-similarity matrices capture the intrinsic spatial organization of object parts; therefore, high correlation indicates similar structural

layouts, while lower correlation corresponds to structurally diverse contexts. By maximizing nSSM, the model is encouraged to explore structurally different configurations, which often correspond to variations in pose or viewpoint. While we instantiate nSSM using MTG visual features, the formulation applies to any backbone that produces spatially structured feature grids; the identity gate is likewise replaceable by the corresponding backbone's similarity metric.

Reinforcement Learning (RL) allows the model to explore diverse outputs and select structurally novel yet identity-consistent samples. Directly optimizing consistency and diversity objectives together, however, leads to unstable training and reward hacking. We therefore introduce an "Explore-and-Suppress" optimization strategy built on top of Group Relative Policy Optimization (GRPO). The first stage encourages exploration by maximizing structural diversity through nSSM. Then, the second stage applies an identity-preserving gate using a visual similarity metric (VSM) [4], penalizing samples that deviate excessively from the subject identity. This gated formulation converts the conflict between diversity and identity into a collaborative process: the model first discovers structurally diverse solutions and then undesirable samples are further suppressed during the second-stage optimization.

We evaluate our approach on the DreamBench++ benchmark [25] using the Flux-Kontext [13] backbone and compare against several subject-driven generation methods. Our results show that *DivRL* successfully expands the diversity of generated images while maintaining strong identity consistency, effectively populating the high-utility region where both objectives coexist.
Our contributions are summarized as follows:

- We introduce the negative Self-Similarity Measure (nSSM), a feature-space metric that quantifies structural diversity via self-similarity correlations of disentangled visual features.
- We propose an Explore-and-Suppress optimization strategy that decouples diversity exploration from identity preservation using a gated reward formulation.
- Experiments on DreamBench++ demonstrate that our method improves structural diversity while maintaining competitive identity consistency. We further show that the nSSM formulation and two-stage optimization are backbone-agnostic.

2 Related Work

2.1 Subject-driven image generation

Foundational image diffusion models [5, 27, 29] have achieved superior image generation quality. Building on top of the base generative models, a broad line of work tackling the subject-driven image generation fall into three main categories: optimization-Based [7, 28], adapter-based [3, 8, 14, 31, 41, 42], and native multimodal models [2, 13, 18, 33, 34, 38]. Optimization-based frameworks typically require a small set of images which represent the reference object, and a

dedicated fine-tuning process for every new subject. While enabling faithful personalization, they suffer from lack of high frequency details [7] or catastrophic forgetting [28]. Adapter-based methods utilize a pre-trained visual encoder to extract features from a reference image and inject them into a base model. Specifically, [8] proposed shortcut-routed adapter training to reduce the copy-paste phenomenon. While offering high efficiency and modularity, they often struggle with complex spatial reasoning and structural flexibility. To overcome these bottlenecks, recently native multimodal models have emerged that utilize a unified diffusion transformer architecture, allowing for more fluid in-context interaction between visual and textual tokens.

2.2 Reinforcement learning for visual generation

Inspired by the success of fine-tuning language models with RL [24, 30, 43], emerging works have gained success in aligning image models with the human preferences [1, 6, 10, 22, 23, 32, 44]. Prominent work Flow-GRPO [17] integrates GRPO [30] into flow-based models for text-to-image generation. Flow-GRPO converts the deterministic ODE sampling in the flow models into an equivalent SDE, which brings randomness to support the stochastic sampling requirements of the GRPO framework. TempFlow-GRPO [9] further proposed noise-aware weighting scheme to prioritize different denoising timesteps during sampling stages. To improve the sampling efficiency of Flow-GRPO, DanceGRPO [40] selects only specific critical timesteps to update, MixedGRPO [15] applied the gradient updates only within a sliding window of the denoising timestep.

Multiple features-based rewards [11,16,37,39] or VLM-based rewards [19,20, 35] are designed to measure the image quality from various aspects. To apply RL on the specific subject-driven image generation task, PaCO-RL [26] and Identity-GRPO [21] proposed pairwise consistency preferences reward modeling, measuring relative ranking of which instance is more consistent than the other given a pair of samples. Unlike these pairwise methods that optimize for relative preference, our framework introduces a gated reward mechanism to explicitly decouple identity preservation from structural mimicry, enabling the model to actively explore regions of the solution space that simultaneously satisfy identity consistency and structural diversity.

3 Method

We first introduce the preliminaries about Flow-GRPO. Then, we explain how we design the reward models for identity consistency and structural diversity. Afterwards, we explain our "Explore-and-Suppress" optimization strategy for aggregating two reward models using a two-stage training process.

3.1 Preliminaries

Flow-GRPO optimizes the policy by comparing a group of sampled outputs against their mean reward, which avoids the need for a separate value function

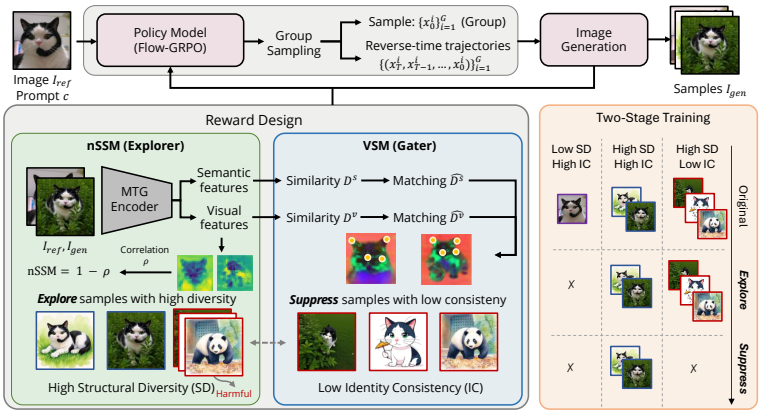

Fig. 2: Pipeline of our method DivRL. We design reward models VSM for *identity consistency* (IC) and nSSM for *structural diversity* (SD). A two-stage "Explore-and-Suppress" strategy is used to generate images that are both identity-consistent and structurally diverse. The first stage encourages free exploration for diverse generation, while the second stage employs a gater to filter out low-consistency samples, maintaining both diversity and consistency.

and therefore reduce VRAM consumption. Concretely, for each prompt, multiple responses are sampled and their rewards are normalized within the group to form relative advantages. Given a prompt c, the flow model p_θ samples a group of G individual images $\{x_0^i\}_{i=1}^G$ and the corresponding reverse-time trajectories $\{(x_T^i, x_{T-1}^i, ..., x_0^i)\}_{i=1}^G$. The rewards are calculated on the clean images as $R(x_0^i, c)$. The advantage is define as:

$$\widehat{A}_t^i = \frac{R(x_0^i, c) - \text{mean}(\{R(x_0^i, c)\}_{i=1}^G)}{\text{std}(\{R(x_0^i, c)\}_{i=1}^G)}. \tag{1}$$

The policy is then updated to increase the flow objective of samples with above-average rewards while suppressing below-average ones. The objective is to maximize the following objective:

$$J_\theta = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=1}^{T}\left(\min(r_t^i(\theta)\widehat{A}_t^i, \text{clip}(\min(r_t^i(\theta)\widehat{A}_t^i, \epsilon)) - \beta D_{\text{KL}}(\pi_\theta|\pi_{\text{ref}})\right),$$
$$r = \frac{\pi_\theta}{\pi_{\text{old}}}, \tag{2}$$

where π_θ is the current policy and π_{old} is the previous policy, and D_{KL} is the KL regularization term to keep the current policy not to be too far away from the reference policy π_{ref}.

3.2 Reward Design

In subject-driven image generation, given a reference image I_{ref} and a text prompt c, our goal is to generate both consistent (*i.e.* identity consistent) and diverse (*i.e.* structurally diverse) images I_{gen} that follow the given text prompt.

Visual Identity Consistency In the context of RL for post training, we want to find suitable reward signals for identity consistency and structural diversity. Recent work MTG [4] extracts fine visual information by disentangling the visual and semantic features from the backbone of the pre-trained diffusion model. These visual features are robust under varying scales, poses, and contexts, making them a valuable metric to evaluate the performance of the identity consistency between generated images and reference images.

Specifically, we forward I_{ref} and I_{gen} to the MTG network to extract semantic features $\mathcal{F}^s$ and visual features $\mathcal{F}^v$, both with a shape $\mathbb{R}^{48\times48\times c}$, where c is the number of channels. We follow by computing pairwise similarities between individual features to obtain semantic and visual similarity matrices $\mathcal{D}^s$ and $\mathcal{D}^v$, respectively. The maximum similarity score for each point is taken to compute the best per-point match: $\widehat{\mathcal{D}}^s = \max(\mathcal{D}^s)$ and $\widehat{\mathcal{D}}^v = \max(\mathcal{D}^v)$. Semantic correspondences are identified by selecting points whose semantic similarity is above threshold τ_s, *i.e.* $\widehat{\mathcal{D}}^s > \tau_s$. Visual consistency is further assessed over semantically consistent regions:

$$\mathrm{VSM}(\tau_v) = \frac{1}{\mathcal{J}_s} \sum_{j \in \mathcal{J}_s} 1(\mathcal{D}^v > \tau_v) \tag{3}$$

where $\mathcal{J}_s$ denotes the regions that satisfy $\widehat{\mathcal{D}}^s > \tau_s$. A higher VSM generally indicates a higher visual similarity for pixels corresponded to the semantically same object, which imposes a better ID preservation. Figure 2 illustrates this process.

Structural diversity Building on top of the visual features $\mathcal{F}^v_{ref}$ for I_{ref} and $\mathcal{F}^v_{gen}$ for I_{gen}, we further propose negative Self-Similarity Measure (nSSM) for diversity evaluation. We first normalize $\mathcal{F}^v_{ref}$ and $\mathcal{F}^v_{gen}$ across the channel dimension to obtain $\widehat{\mathcal{F}}^v_{ref}$ and $\widehat{\mathcal{F}}^v_{ref}$ of shape $\mathbb{R}^{48\times48}$ each. Afterwards, we calculate the SSM matrices for $\widehat{\mathcal{F}}^v_{ref}$ and $\widehat{\mathcal{F}}^v_{gen}$ to obtain $\mathcal{M}^v_{gen}$ and $\mathcal{M}^v_{ref}$, respectively. The SSM matrices are calculated as:

$$\mathcal{M}^v_{ref} = \widehat{\mathcal{F}}^v_{ref}(\widehat{\mathcal{F}}^v_{ref})^T, \qquad \mathcal{M}^v_{gen} = \widehat{\mathcal{F}}^v_{gen}(\widehat{\mathcal{F}}^v_{gen})^T. \tag{4}$$

The SSM matrices capture the self-organization of the individual feature patches. We also apply the GT object mask to both $\mathcal{M}^v_{ref}$ and $\mathcal{M}^v_{ref}$ to obtain $\widehat{\mathcal{M}}^v_{ref}$ and $\widehat{\mathcal{M}}^v_{gen}$. Afterwards, we compute the Pearson correlation coefficient between

$\widehat{\mathcal{M}}^v_{ref}$ and $\widehat{\mathcal{M}}^v_{gen}$.

$$\rho_{(\widehat{\mathcal{M}}^v_{ref},\widehat{\mathcal{M}}^v_{gen})} = \frac{\mathrm{cov}(\widehat{\mathcal{M}}^v_{ref},\widehat{\mathcal{M}}^v_{gen})}{\sigma_{\widehat{\mathcal{M}}^v_{ref}}\sigma_{\widehat{\mathcal{M}}^v_{gen}}}. \tag{5}$$

Intuitively, if the Pearson correlation coefficient is higher, that means the semantic and structural similarity is also higher. As we want to encourage diversity, we take the negative term of $\rho_{(\widehat{\mathcal{M}}^v_{ref},\widehat{\mathcal{M}}^v_{gen})}$ and define the final nSSM as

$$\mathrm{nSSM} = 1 - \rho_{(\widehat{\mathcal{M}}^v_{ref},\widehat{\mathcal{M}}^v_{gen})}. \tag{6}$$

This nSSM reward is computed on dense latent diffusion feature grid, which incorporates fine-grained geometry and can provide rich structural supervision.

Stabilizing Convergence Recall that the spatial resolution of both $\mathcal{F}^v_{ref}$ and $\mathcal{F}^v_{gen}$ is 48×48. While this fine-grained resolution can capture rich structure-related details and make it a suitable resolution to evaluate the structural similarity, the model is implicitly encouraged to preserve those dense local correlation patterns during the training stage, leading to high-frequency textural artifacts. To solve this issue, we additionally apply a 2×2 average pooling on the normalized visual features to obtain downsampled features $\widehat{\mathcal{F}}^v_{ref}$ and $\widehat{\mathcal{F}}^v_{ref}$, respectively, which are then of size 24×24. The $\widehat{\mathcal{F}}^v_{ref}$ is then calculated as:

$$\begin{aligned}\widehat{\mathcal{F}}^v_{ref} &= \mathrm{avgpool}_{2\times2}(\mathrm{normalize}(\mathcal{F}^v_{ref})),\\ \widehat{\mathcal{F}}^v_{gen} &= \mathrm{avgpool}_{2\times2}(\mathrm{normalize}(\mathcal{F}^v_{gen})).\end{aligned} \tag{7}$$

3.3 Optimization Strategy

We propose a two-stage optimization strategy to facilitate the RL. The only difference between these two stages lies in the definition of the reward for training. In the first stage, we merely use the nSSM as the reward model:

$$R_1 = nSSM, \tag{8}$$

which effectively encourages the model to explore desirable latent region generating both consistent and diverse samples. However, we empirically observe the emergence of inconsistent samples along the training due to the reward hacking (*i.e.*, generating random samples without any consistency can also increase the reward of nSSM).

In the second stage, we introduce the VSM as a "gate" to suppress those noisy samples that provide inconsistent gradient directions. The reward is formulated as

$$R_2 = \begin{cases} \mathrm{nSSM} & \mathrm{VSM} \geq s, \\ \mathrm{nSSM} - \lambda(s - \mathrm{VSM})^2 & \mathrm{VSM} < s, \end{cases} \tag{9}$$

where s is a threshold for identity consistency, and λ is a weighting term for the penalty of the non-similarity degree. Intuitively, the gated reward transforms identity preservation into a feasibility constraint rather than a competing objective. The model is free to explore diverse structural configurations as long as they remain within the identity-consistent region. The introduction of VSM effectively alleviates the reward hacking issue and further facilitates the optimization of nSSM, reduces the tendency of the model to rely on rigid solutions and explores the high-quality latent space where identity and structural novelty coexist. This avoids the optimization instability observed when diversity and identity are directly combined through linear weighting.

4 Experiments

4.1 Implementation details

Settings We adopt the vanilla Flow-GRPO framework to optimize the model and use Flux-Kontext [13] as the backbone for subject-driven image generation. Training is conducted on a 10k subset of the SynCD dataset [12], where each identity is paired with 10 text prompts. For the reward model, we set the semantic and visual thresholds to $\tau_s = 0.7$ and $\tau_v = 0.7$, respectively. The hinge parameters are set to $\lambda = 5$ and $s = 0.5$. Training is performed in two stages: 3200 optimization steps for the exploration stage, followed by 3200 steps for the suppression stage. For Flow-GRPO, the group size is set to 21 ,and the number of denoising steps during sampling is 6. The KL regularization coefficient is $\beta = 0.1$. During inference, we use 28 denoising steps. Training is conducted on 8 NVIDIA A100 (80GB) GPUs, with each training stage taking 24 hours.

Evaluation We evaluate the models from four perspectives: identity consistency, structural diversity, prompt following, and aesthetic quality. For identity consistency, we report the out-of-domain metrics CLIP image cosine similarity and DINO cosine similarity, as well as the in-domain metric VSM. We additionally report the *consistency ratio*, defined as the percentage of generated samples satisfying VSM > 0.6 for in-domain evaluation, and structural DINO sDINO > 0.75 for out-of-domain evaluation. For structural diversity, we report the in-domain metric MTG-nSSM and out-of-domain metrics including DINO-nSSM, scale-invariant IoU, and LPIPS. DINO-nSSM is computed using the same formulation as MTG-nSSM but with DINO features instead of MTG visual features. We further report *diversity-over-consistency*, which measures diversity among identity-consistent samples. The in-domain version computes the average MTG-nSSM over samples with VSM > 0.6, while the out-of-domain version computes the average DINO-nSSM over samples with sDINO > 0.75. We will provide the detailed definition of the metrics in the Supplementary Materials. All evaluations are conducted on DreamBench++ [25], which contains 150 identities spanning animals, humans, objects, and style transfer scenarios, each paired with 9 text prompts.

Table 1: Comparison of different subject-driven generation models across various metrics.

Model	Prompt foll.	Identity consistency			Diversity			Aesthetic
	CLIP-T↑	CLIP-I↑	DINO↑	VSM-0.7↑	DINO-nSSM↑	Scale-inv IOU↓	MTG-nSSM↑	HPS↑
Flux-IP-Adapter	0.274	0.805	0.568	0.481	0.531	0.593	0.765	0.275
OmniGen	0.293	0.735	0.474	0.464	0.546	0.583	0.786	0.301
OmniGen2	0.301	0.802	0.559	0.460	0.496	0.638	0.727	0.311
UNO	0.209	0.710	0.387	0.454	0.590	0.637	0.816	0.215
Flux-Kontext	<u>0.283</u>	0.781	0.594	0.605	0.433	0.694	0.659	<u>0.294</u>
PaCo-RL	**0.287**	0.766	0.570	0.576	0.445	<u>0.683</u>	0.668	**0.302**
DivRL (VSM only)	0.278	**0.806**	**0.670**	**0.688**	0.383	0.757	0.608	0.291
DivRL (nSSM only)	**0.287**	0.756	0.520	0.562	**0.489**	**0.634**	**0.704**	0.293
DivRL (VSM + nSSM)	0.279	<u>0.786</u>	<u>0.600</u>	<u>0.614</u>	<u>0.453</u>	0.694	<u>0.689</u>	0.287

Baselines We compare our method with Flux-IP-Adapter[1], OmniGen [38], OmniGen2 [34], UNO [36], PaCo-RL [26], and the original Flux-Kontext [13]. Flux-IP-Adapter injects image conditioning into the Flux backbone through an IP-Adapter module. OmniGen and OmniGen2 are unified multimodal models capable of handling various image-conditioned generation tasks. UNO is a multi-image conditioned multimodal model designed for subject-driven generation. PaCo-RL introduces a pairwise consistency reward model and fine-tunes Flux-Kontext using Flow-GRPO. All methods are evaluated at a resolution of 1024×1024.

4.2 Quantitative results

We present a quantitative evaluation of different methods in terms of prompt following, identity consistency, structural diversity ,and aesthetic in Table 1. We categorize the methods into two different categories: Flux-Kontext-based models and others. In addition to our full model, we report two variants trained with a single reward signal: VSM-only and nSSM-only, which illustrate the optimization boundaries of identity consistency and structural diversity. Optimizing with VSM alone significantly improves identity consistency compared to the baselines, but leads to a noticeable drop in diversity metrics. In contrast, optimizing with nSSM alone substantially increases structural diversity but reduces identity consistency. By combining the two rewards through our proposed optimization strategy, our method achieves a better balance between identity consistency and structural diversity. In particular, it improves diversity over the Flux-Kontext baseline while maintaining comparable identity consistency.

Interestingly, even without explicitly optimizing for text alignment, our method achieves strong prompt-following performance, with the nSSM-only variant obtaining the highest text-alignment score. This phenomenon can be explained by the geometric rigidity induced by identity preservation during pretraining. Encouraging structural diversity through nSSM relaxes this rigidity, allowing the model to better adapt the subject to the contextual requirements of the prompt.

[1] https://huggingface.co/XLabs-AI/flux-ip-adapter

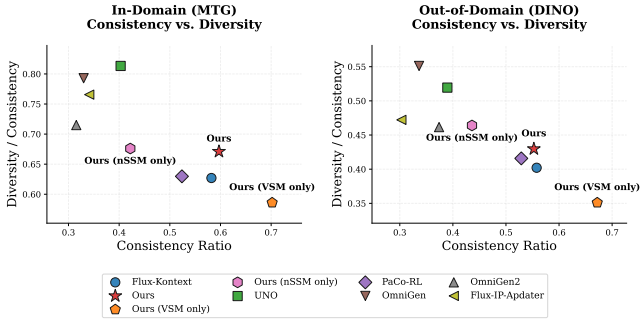

Fig. 3: Quantitative evaluation of the trade-off between identity consistency and structural diversity. Results are reported using both in-domain metrics (MTG) and out-of-domain metrics (DINO).

We further analyze the trade-off between consistency and diversity in Figure 3 using the consistency ratio and diversity-over-consistency metrics. Both in-domain and out-of-domain evaluations show that non-Flux-Kontext-based models achieve high diversity for samples that pass the consistency threshold, but their overall consistency ratio is significantly lower than that of Flux-Kontext-based methods. This suggests that these models struggle to reliably preserve subject identity, which is a fundamental requirement in subject-driven generation. Notably, OmniGen2 demonstrates markedly improved global feature similarity over OmniGen, yet its VSM score and consistency ratio remain low, indicating that it still falls short of reliable fine-grained identity preservation and cannot close the gap with Flux-Kontext-based methods on this dimension.

Among Flux-Kontext-based methods, the base Flux-Kontext model already achieves strong identity preservation, reflected by its high consistency ratio. PaCo-RL produces slightly more diverse outputs but exhibits a lower consistency ratio compared to the base model. The two variants of our method demonstrate the expected extremes: the nSSM-only model achieves higher diversity but lower consistency, while the VSM-only model achieves the highest consistency ratio but the lowest diversity. Our full method combines the advantages of both variants, maintaining a consistency level comparable to the baselines while achieving improved diversity.

4.3 Qualitative results

We present qualitative comparisons between our method and the baselines in Figure 4. The examples include both rigid objects (*e.g.*, bicycle and van) and articulated objects (*e.g.*, stork and human). Our method preserves the fine visual details of the reference subject while following the text prompt and exhibiting structural diversity, such as changes in pose and viewpoint.

Flux-Kontext and PaCo-RL produce visually plausible results but often exhibit two failure modes. First, they occasionally fail to preserve fine-grained visual details. Second, although they maintain identity consistency, they some-

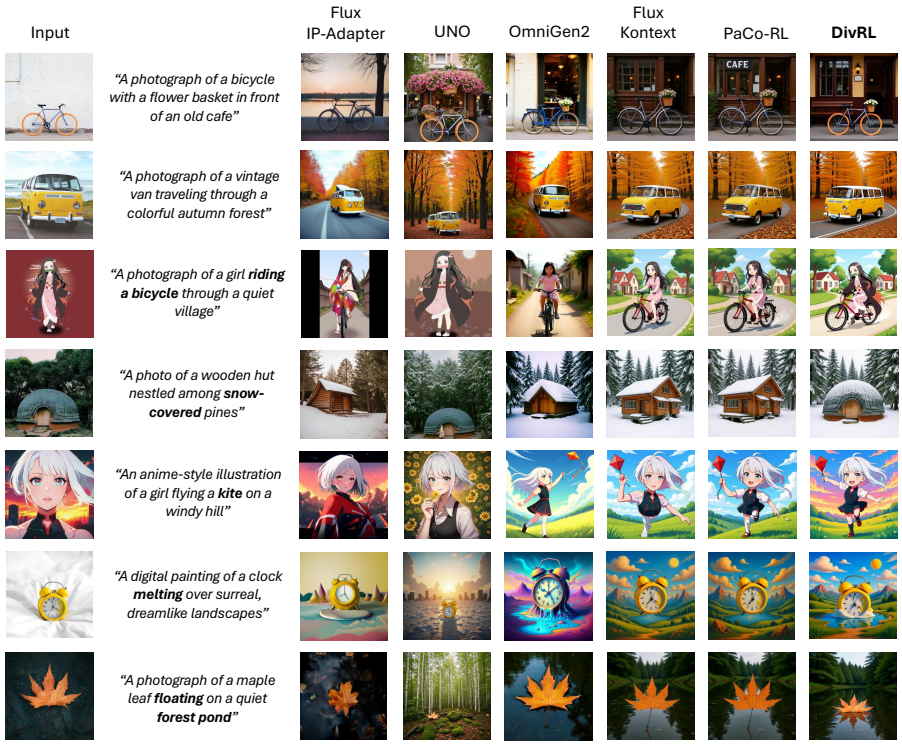

Fig. 4: Visual comparison with baselines. Our method produces structurally diverse images while preserving identity consistency and maintaining good prompt alignment.

times fail to follow the prompt interactions. This behavior stems from the strong emphasis on identity preservation, which encourages the model to reproduce a subject configuration similar to the reference rather than adapting it to the contextual instructions.

Other non–Flux-Kontext-based models tend to produce more structurally diverse outputs, which is consistent with the quantitative results. However, they exhibit different limitations. Flux-IP-Adapter often lacks fine visual details and overall fidelity. UNO generates visually consistent subjects but struggles with prompt alignment. OmniGen produces diverse and visually appealing images but frequently fails to preserve identity. These observations highlight the difficulty of simultaneously achieving strong identity preservation, structural diversity, prompt following, and high visual quality. Building upon the strong identity preservation capability of Flux-Kontext, our method relaxes the structural rigidity of the base model, enabling more diverse subject configurations while maintaining identity consistency and improving prompt alignment.

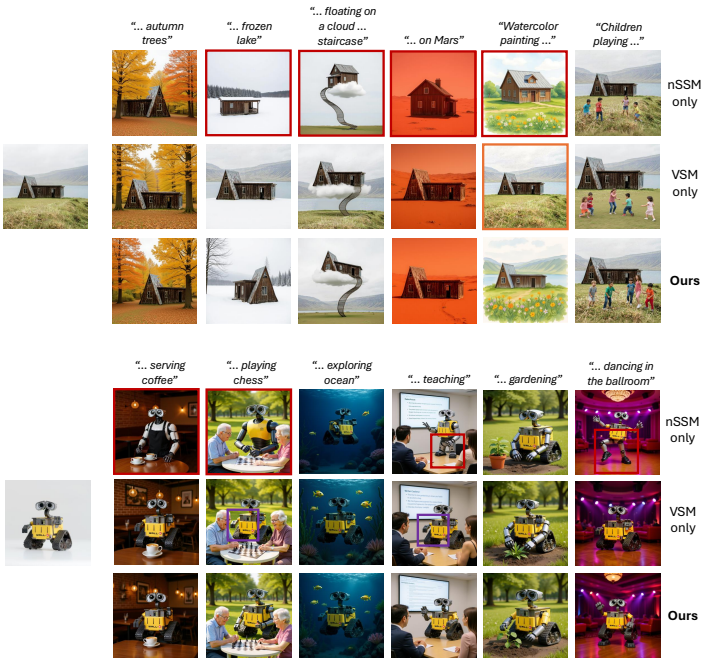

Fig. 5: Qualitative comparison between optimizing using nSSM only, VSM only, and ours. We highlight cases of identity inconsistency, prompt misalignment, and structural rigidity. Our method DivRL achieves the best balance between prompt following, identity preservation, and structural diversity.

4.4 Ablation Studies

Single-Reward Optimization We further compare our method with the two reward variants in Figure 5. Optimizing with nSSM alone produces higher structural diversity by generating subjects with different poses and viewpoints, but it may alter the subject's core attributes and lead to identity drift. In contrast, optimizing with VSM alone achieves strong identity preservation but may result in overly constrained outputs with poses and viewpoints similar to the reference image. In some cases, the model even preserves the original visual style of the reference, ignoring stylistic instructions from the prompt. By combining nSSM and VSM, our method achieves a better balance between identity consistency, structural diversity, and prompt following.

Optimization Strategy Our goal is to balance identity consistency and diversity during optimization. Since Flux-Kontext already provides strong identity preservation, we focus on improving diversity without degrading consistency. To analyze this trade-off, we compare our method with linear combinations of VSM and nSSM, as shown in Figure 6. As the training progression curves show, our method produces a modest but sufficient increase in VSM while achieving a substantial gain in nSSM over the Flux-Kontext baseline. Linear weighting, by contrast, usually fails to simultaneously improve both objectives. This can

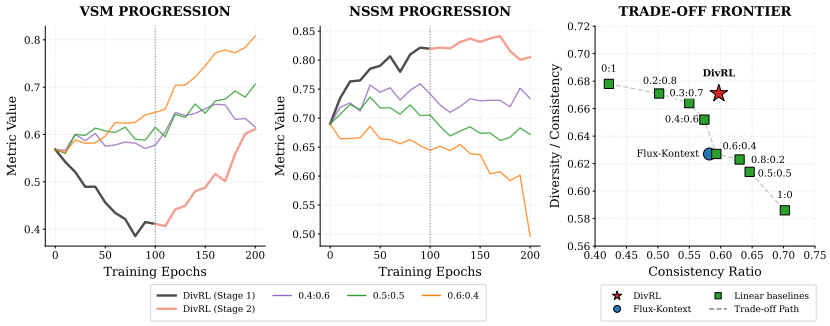

Fig. 6: Comparison between our optimization strategy and linear reward weighting. Left is the training progression curve, while right is the trade-off frontier between the *consistency ratio* and *diversity-over-consistency*. For the linear weighting annotation, it is denoted as the linear ratio between VSM and nSSM.

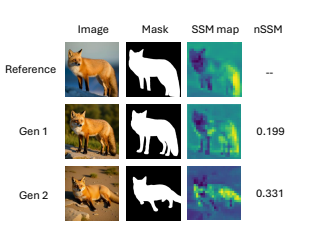

Fig. 7: Visualization of the SSM maps and the corresponding nSSM scores.

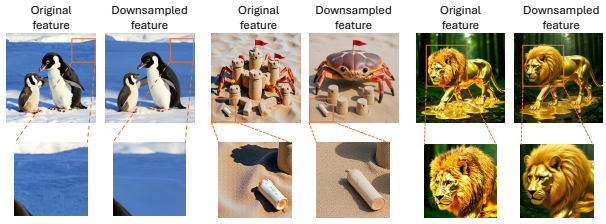

Fig. 8: Comparison between computing nSSM using the original visual feature resolution and using downsampled features. Optimization with the original resolution introduces high-frequency artifacts, whereas downsampled features suppress these artifacts and produce visually pleasing results.

also be verified in the trade-off frontier graph that none of the linear weighting combinations can improve the *consistency ratio* as well as the *diversity-over-consistency* at the same time. This failure occurs because the gradients from the two reward signals are in direct competition under a linear combination. The gated formulation in our method decouples these two objectives: the model first explores structurally diverse solutions, and the identity gate selectively suppresses only those that drift beyond the consistency threshold, allowing VSM and nSSM to improve jointly.

nSSM visualization We visualize the Self-Similarity Measure (SSM) maps in Fig 7. The similarity scores within the masked region are aggregated to produce the SSM maps. When a generated image closely matches the reference structure (Gen 1), the SSM maps exhibit high similarity, resulting in a lower nSSM score. In contrast, a sample with a different pose (Gen 2) produces less similar SSM maps and therefore yields a higher nSSM score. During training, masks are not generated for synthesized images; instead, nSSM is computed using features from all spatial locations of the generated images.

Spatial Resolution of the Self-Similarity Measure The MTG visual features have a spatial resolution of 48×48. While this resolution captures fine structural details, we found that directly computing nSSM at this scale leads to high-frequency artifacts during RL optimization. We hypothesize that optimizing structural similarity at such fine granularity encourages the model to reproduce dense local correlation patterns, allowing it to increase nSSM through high-frequency textures rather than meaningful structural changes, which is a form of reward hacking. To mitigate this issue, we apply 2×2 average pooling to downsample the feature maps to 24×24 before computing nSSM. As shown in Figure 8, the downsampled features effectively suppress high-frequency artifacts and produce visually cleaner results.

5 Discussion and Limitations

Our results highlight an important observation in subject-driven generation: identity preservation rewards alone implicitly encourage structural mimicry. Our framework addresses this issue by explicitly separating structural diversity from identity preservation. An interesting side effect of encouraging structural diversity is the improvement in prompt interaction. As shown in our experiments, relaxing structural rigidity allows the model to better adapt the subject to contextual cues in the prompt.

Limitations. Despite these advantages, several limitations were observed. Firstly, the proposed diversity metric focuses on structural variations in pose and viewpoint, and may not fully capture higher-level semantic diversity or compositional changes. Secondly, our approach relies on RL for post-training, which introduces additional computational overhead compared to purely feed-forward methods. Finally, our base model Flux-Kontext accepts a single reference image by default and may be less effective when the reference contains ambiguous identity cues or complex occlusions. Addressing these limitations through richer diversity metrics, more efficient optimization strategies, and multi-reference conditioning is an interesting direction for future work.

6 Conclusions

In this paper, we investigate the identity–diversity trade-off in subject-driven image generation. We identify structural mimicry as a common failure mode in existing methods, where models tend to reproduce the spatial configuration of the reference image instead of generating diverse structures. To address this challenge, we propose a RL-based framework that decouples structural diversity from identity preservation. We introduce a negative Self-Similarity Measure (nSSM) to quantify structural diversity and employ Visual Semantic Matching (VSM) as an identity consistency gate. Combined with an Explore-and-Suppress optimization strategy based on Flow-GRPO, this formulation encourages the model to explore diverse structural configurations while filtering identity-inconsistent samples. Experiments demonstrate that the proposed method improves structural diversity while maintaining strong identity consistency, achieving a better balance between diversity, prompt following, and identity preservation compared with existing approaches.

Acknowledgement

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST) – Center of Excellence for Generative AI, under award number 5940 and a gift from Google.

References

1. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning (2023)
2. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
3. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6593–6602 (2024)
4. Eldesokey, A., Cvejic, A., Ghanem, B., Wonka, P.: Mind-the-glitch: Visual correspondence for detecting inconsistencies in subject-driven generation. In: Advances in Neural Information Processing Systems (2025)
5. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
6. 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. Advances in Neural Information Processing Systems **36**, 79858–79885 (2023)
7. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=NAQvF08TcyG
8. Goyal, A., Qian, G., Coskun, H., Gupta, A., Tam, H., Ostashev, D., Hu, J., Sagar, D., Tulyakov, S., Aberman, K., Wang, K.C.: Preventing shortcuts in adapter training via providing the shortcuts. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=cZMno8E3yp
9. He, X., Fu, S., Zhao, Y., Li, W., Yang, J., Yin, D., Rao, F., Zhang, B.: TEMPFLOW-GRPO: WHEN TIMING MATTERS FOR GRPO IN FLOW MODELS. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=7mCo3R3Wyn
10. Hu, Z., Zhang, F., Chen, L., Kuang, K., Li, J., Gao, K., Xiao, J., Wang, X., Zhu, W.: Towards better alignment: Training diffusion models with reinforcement learning against sparse rewards. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23604–23614 (2025)
11. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in neural information processing systems **36**, 36652–36663 (2023)
12. Kumari, N., Yin, X., Zhu, J.Y., Misra, I., Azadi, S.: Generating multi-image synthetic data for text-to-image customization. In: IEEE International Conference on Computer Vision (ICCV) (2025)

13. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742
14. Li, D., Li, J., Hoi, S.: BLIP-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=g6We1SwaY9
15. Li, J., Cui, Y., Huang, T., Ma, Y., Fan, C., Yang, M., Zhong, Z.: Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. arXiv preprint arXiv:2507.21802 (2025)
16. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: European Conference on Computer Vision. pp. 366–384. Springer (2024)
17. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., ZHANG, D., Ouyang, W.: Flow-GRPO: Training flow matching models via online RL. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=oCBKGw5HNf
18. Liu, Z., Ren, W., Liu, H., Zhou, Z., Chen, S., Qiu, H., Huang, X., An, Z., Yang, F., Patel, A., et al.: Tuna: Taming unified visual representations for native unified multimodal models. arXiv preprint arXiv:2512.02014 (2025)
19. Long, Y., Yang, Y., Wei, H., Chen, W., Zhang, T., Liu, C., Jiang, K., Chen, J., Tang, K., Wen, B., et al.: Spatialreward: Bridging the perception gap in online rl for image editing via explicit spatial reasoning. arXiv preprint arXiv:2602.07458 (2026)
20. Luo, X., Wang, J., Wu, C., Xiao, S., Jiang, X., Lian, D., Zhang, J., Liu, D., et al.: Editscore: Unlocking online rl for image editing via high-fidelity reward modeling. arXiv preprint arXiv:2509.23909 (2025)
21. Meng, X., Zhang, Z., Zhang, Z., Liao, J., Qin, L., Wang, W.: Identity-grpo: Optimizing multi-human identity-preserving video generation via reinforcement learning. arXiv preprint arXiv:2510.14256 (2025)
22. Miao, Z., Wang, J., Wang, Z., Yang, Z., Wang, L., Qiu, Q., Liu, Z.: Training diffusion models towards diverse image generation with reinforcement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10844–10853 (2024)
23. Na, S., Kim, Y., Lee, H.: Boost your human image generation model via direct preference optimization. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23551–23562 (2025)
24. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
25. Peng, Y., Cui, Y., Tang, H., Qi, Z., Dong, R., Bai, J., Han, C., Ge, Z., Zhang, X., Xia, S.T.: Dreambench++: A human-aligned benchmark for personalized image generation. In: The Thirteenth International Conference on Learning Representations (2025), https://dreambenchplus.github.io/
26. Ping, B., Jia, C., Luo, M., Xia, C., Shen, X., Dang, Z., Qian, H.: Paco-rl: Advancing reinforcement learning for consistent image generation with pairwise reward modeling. arXiv preprint arXiv:2512.04784 (2025)

27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
28. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
29. Sauer, A., Boesel, F., Dockhorn, T., Blattmann, A., Esser, P., Rombach, R.: Fast high-resolution image synthesis with latent adversarial diffusion distillation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
30. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y.K., Wu, Y., Guo, D.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models (2024), https://arxiv.org/abs/2402.03300
31. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8543–8552 (2024)
32. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8228–8238 (June 2024)
33. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324
34. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., Liu, Z., Xia, Z., Li, C., Deng, H., Wang, J., Luo, K., Zhang, B., Lian, D., Wang, X., Wang, Z., Huang, T., Liu, Z.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
35. Wu, K., Jiang, S., Ku, M., Nie, P., Liu, M., Chen, W.: Editreward: A human-aligned reward model for instruction-guided image editing. arXiv preprint arXiv:2509.26346 (2025)
36. Wu, S., Huang, M., Wu, W., Cheng, Y., Ding, F., He, Q.: Less-to-more generalization: Unlocking more controllability by in-context generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18682–18692 (2025)
37. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
38. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13294–13304 (June 2025)
39. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: learning and evaluating human preferences for text-to-image generation. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. pp. 15903–15935 (2023)
40. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)

41. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models (2023)
42. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
43. Zheng, R., Dou, S., Gao, S., Hua, Y., Shen, W., Wang, B., Liu, Y., Jin, S., Liu, Q., Zhou, Y., et al.: Secrets of rlhf in large language models part i: Ppo. arXiv preprint arXiv:2307.04964 (2023)
44. Zhu, H., Xiao, T., Honavar, V.G.: DSPO: Direct score preference optimization for diffusion model alignment. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=xyfb9HHvMe