Style-GRPO: Semantic-Aware Preference Optimization for Image Style Transfer Guided by Reward Modeling

Jianbin Zhao[1*], Chaoran Feng[1*], Miao Yu[2*], Yingtao Li[3], Zhenyu Tang[1], Wangbo Yu[1], Yian Zhao[1], Xiaomin Li[3], Li Yuan[1†], Yonghong Tian[1†]

[1]Peking University, [2]Communication University of China, [3]Dalian University of Technology

Abstract

We present a novel approach to bridging the gap between style transfer guided by reference styles and faithful content preservation in text-to-image generation. First, we introduce ***StyleReward-Dataset****, an expert-annotated benchmark with over 300k adversarial image pairs spanning diverse real-world and virtual styles. Each example is constructed to contrast a faithful result with targeted counterexamples, enabling preference learning over style consistency, content preservation, and perceptual quality. Leveraging StyleReward-Dataset, we propose* ***StyleScore****, an end-to-end multimodal reward model that integrates visual and semantic understanding to provide reliable, human-aligned evaluations of generated images. Additionally, based on StyleReward-Dataset, we develop a two-stage training framework for style transfer, consisting of supervised fine-tuning on StyleReward-Dataset followed by reinforcement learning with Group Relative Preference Optimization that converts relative preferences into stable policy updates. Extensive experiments on both public benchmark and proposed benchmark show that our proposed method achieves substantial improvements in style fidelity and content preservation over strong baselines, with human studies further validating its superior visual realism and alignment with human preference.*

1. Introduction

Recent advances in flow-based diffusion models [10, 29, 58] have led to remarkable progress in Text-to-Image (T2I) generation [5, 7, 14, 25, 32, 44–47, 64, 67, 72], achieving high-quality and diverse image synthesis. Building upon this, diffusion models are increasingly being extended to image editing tasks [30, 36, 57–59], which demand both fine-grained control over style and semantic preservation of the input image. However, when tackling style-aligned image generation, which requires the generated image to adhere to a reference style while maintaining the semantic integrity of the content, existing methods face significant challenges in decoupling style from content images.

Despite this progress, image style transfer remains a challenging task in editing. In this context, the generated image must align with a reference style prompt while faithfully preserving the semantics of the input image. Existing editing models struggle to effectively disentangle global style from content, often resulting in style leakage and semantic drift, especially with complex, compositional instructions [34, 60, 71]. While supervised fine-tuning (SFT) can offer partial solutions, it often overfits to dataset biases and fails to generalize reliably to diverse style representations.

In this work, we incorporate explicit style and content considerations into the editing objective to reduce the gap between intended and realized outputs. We introduce StyleReward-Dataset, an expert-annotated dataset comprising 300k adversarial image pairs that span diverse styles (*cyberpunk*, *ink-wash*, etc.) and include both real-world cultural and virtual game aesthetics. Each example is collected in an adversarial setup with one image consistent with both the content image and the style prompt, and the other counterexamples that violate realism in complementary ways (*style-only*, *content-only*). All annotations are reviewed by vision-language model(VLM) and human experts with reference to authoritative sources to ensure consistency and accuracy.

Building on StyleReward-Dataset, we further propose StyleScore, an end-to-end VLM reward model leveraging diverse expert-level knowledge. StyleScore is designed to evaluate generated images in a manner analogous to a human rater, jointly assessing style consistency, content preservation, and perceptual quality. Our prelimi-

[*] These authors contributed equally to this work.

[†] Corresponding author.

nary results show that STYLESCORE significantly outperforms general-purpose VLMs [2, 42]. This is because such models often conflate general aesthetic appeal with true style fidelity, failing to penalize results that sacrifice content preservation for visual quality. In contrast,STYLESCORE provides a holistic, task-aligned reward by jointly assessing style consistency, content preservation, and perceptual quality, offering a more accurate measure of human preference for this specific task.

To achieve high-fidelity style transfer, we introduce STYLE-GRPO, a novel two-stage training pipeline. The first stage is a crucial domain adaptation step, leveraging supervised fine-tuning (SFT) on the general-purpose FLUX.1[Kontext] [30] model with our STYLEREWARD-DATASET. This SFT stage adapts the model to the nuances of the style transfer task, thereby establishing a stable initial policy that is essential to ground the subsequent reinforcement learning phase and ensure training stability. Building upon this adapted model, the second stage employs Group Relative Preference Optimization (GRPO) [19]. In this refinement phase, STYLESCORE serves as the reward model, providing feedbacks on style consistency, content preservation, and perceptual quality to guide the policy updates.

Our main contributions are summarized as follows:

- We present STYLEREWARD-DATASET, an expert-annotated benchmark with over 300k adversarial image pairs and 150k prompts spanning diverse real-world and virtual styles, enabling preference learning for style consistency, content preservation, and perceptual quality.
- We introduce STYLESCORE, an end-to-end VLM-based reward model that serves as a language-guided verifier for reference-driven style transfer, providing reliable signals for style, content, and quality without reliance on prompt-engineering-heavy pipelines.
- Extensive experiments on public benchmarks demonstrate consistent gains over strong baselines in style consistency and content preservation, with improvements confirmed by both quantitative metrics and user studies.

2. Related Work

2.1. Image Style Transfer

Image style transfer has been widely studied by researchers. It involves the process of applying stylistic representations to content images in order to generate images with specific styles. Early approaches [15, 16, 24, 26, 49] based on CNN and GAN separate and recombine content and style representations to achieve basic style translation. Although effective for simple artistic effects, these methods often struggle to capture complex or high-fidelity style patterns. Recent diffusion-based models [11, 12, 20, 31, 34, 39, 54, 61, 69] and autoregressive model (AR) [60] enable flexible transfer from minimal references such as a single image or text description, significantly improving visual quality. Additionally, Omnistyle [57], StyleBooth [20], DiffStyler [23] and StyleShot [12] introduced high-quality datasets for style transfer, providing diverse reference–content pairs that significantly improve model training and evaluation. However, achieving a balance between global style alignment and content preservation remains challenging, as models trained for local or text-guided edits often suffer from style leakage, inconsistent stylization, and semantic drift. Our method builds upon these advances with the GRPO framework with the style verifier that explicitly optimizes both style consistency and content fidelity.

2.2. Instruction-guided Image Editing

Image editing aims to modify visual content according to instruction guidance while preserving the consistency of unedited regions. Typical editing tasks include object removal, object addition, inpainting, outpainting, and style transfer [27]. Early diffusion-based image edit methods [5, 51, 68] focus primarily on local editing, where the model modifies only specific spatial regions while maintaining the original context. Recent works like Flux.1 [Kontext] [30], Qwen-Image Edit [58], Hunyuan-Image [7] and others [9, 37, 59] show strong controllability for region-aware manipulations but often require detailed textual corrections and struggle to handle complex, compositional edits with sufficient data. This proficiency in local edits, however, does not readily extend to tasks demanding global transformations, with style transfer being a prime example. Unlike local adjustments, style transfer requires holistic consistency across the entire image to maintain both stylistic coherence and semantic fidelity. Consequently, methods optimized for localized edits frequently lead to style inconsistency and content drift, highlighting the unresolved challenge of decoupling style from content.

2.3. Reinforcement Learning in Diffusion Models

Recent advancements in Reinforcement Learning from Human Feedback (RLHF) algorithms have demonstrated remarkable efficacy in aligning models with human preferences in image synthesis [13]. Reinforcement learning algorithms such as PPO [3] and DPO [55], originally developed for large language models, have been successfully adapted to diffusion-based generation, improving task alignment and controllability. Building on this trend, Flow-GRPO [40] and others [21, 25, 33, 56, 63, 73] integrate GRPO-style policy updates into flow-matching models with T2I task, transforming deterministic ordinary differential equation (ODE) sampling into stochastic stochastic differential equation (SDE) formulations to introduce exploration noise for group-based optimization. Beyond policy optimization, the success of RL depends on a high-quality reward signal [17, 28, 32, 35, 38, 62]. Recent meth-

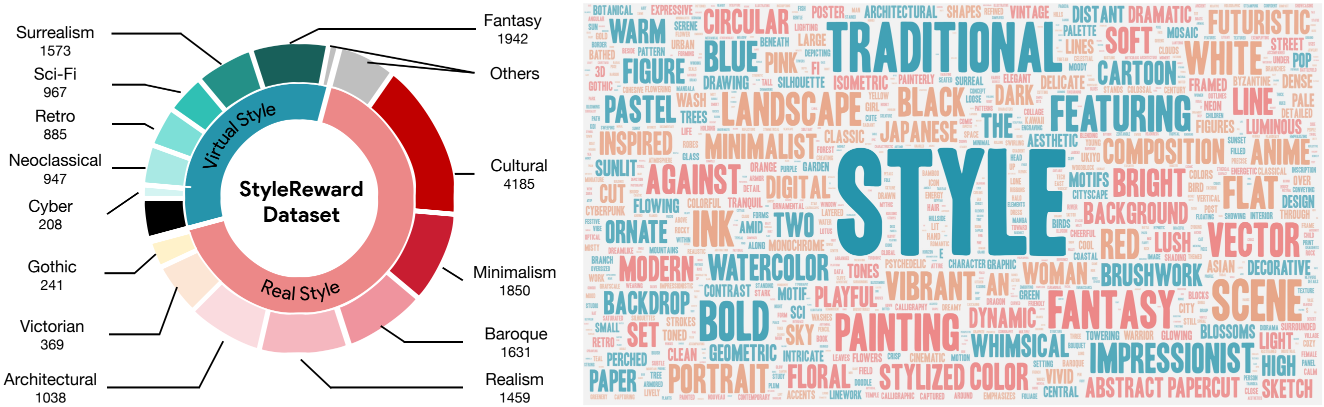

Figure 1. **Data distribution and statistics.** (Left) StyleReward-Dataset is organized into two primary stylized fields: real style and virtual style . Each field is divided into specific categories, with the numbers indicating the volume of implicit prompts collected for each category. (Right) Word cloud of structured prompt in StyleReward-Dataset.

ods [18, 36, 41, 70] employ or fine-tune visionlanguage models as reward models, utilizing pairwise or pointwise comparisons to evaluate generation quality. Building upon these advancements, we extend the visual GRPO algorithm with a semantic-aware reward model for image style transfer, offering the reward signal of style consistency, content preservation, and perceptual quality.

3. Dataset: StyleReward-Dataset

We introduce StyleReward-Dataset, a novel dataset specifically designed to facilitate style-content decoupling in image style transfer models. Unlike conventional datasets, StyleReward-Dataset leverages a large collection of adversarial image pairs to provide fine-grained supervision for human preference.

Dataset Overview. StyleReward-Dataset is constructed to provide large-scale, fine-grained supervision for image style transfer, focusing on both style consistency and content preservation, as illustrated in Figure 1. These tasks are inspired by existing research such as StyleBooth [20] and Style-Tokenizer [34] as well as new concepts developed for this study. Each task is meticulously designed with the following objectives:

- **Comprehensive Style Diversity.** StyleReward-Dataset covers a broad spectrum of visual domains, from real-world aesthetics (e.g., *oil painting, ink-wash, watercolor, photography*) to fictional and digital styles (e.g., *cyberpunk, cel-shading, 3D render*). This diversity allows models to generalize across stylistic domains and to learn domain-invariant representations of style. Further explanations and examples can be found in *Appendix*.
- **Adversarial Pair Construction.** Each instance includes the faithful stylized image and contrastive negative image that intentionally violate style or content alignment (e.g., *correct content but wrong style*, *correct style but wrong content* or *vice versa*). This design supports preference-based supervision and enables fine-grained evaluation of model alignment across multiple dimensions.

Reference Style Curation. To ensure diverse and representative reference styles, we categorize all styles into two major types: real-world styles and virtual-world styles. For data acquisition, we adopt two complementary strategies: real collection and synthetic generation. For the real-world branch, we integrate authentic style images from existing datasets including Style30K [34], Omnistyle [57], and WikiArt [43]. For styles that are difficult to obtain, we employ a T2I generation pipeline [7, 29, 58]. Specifically, we define several high-level style categories and leverage GPT-5 [1] to automatically produce structured templates and style-specific prompts. These prompts are then used to guide T2I models to synthesize high-quality styles.

To ensure data reliability, both real and synthetic style images are filtered using aesthetic score [6], GPT-4o [1] and Gemini 2.5 [8] evaluation, focusing on global image quality, watermark detection, and stylistic coherence. Only images exceeding a predefined quality threshold are retained.

Content Data Curation. For content images and text prompts, we adopt the high-quality text-to-image dataset GenRef-wds [74], which contains over 1M curated text–image pairs. We randomly sample 20k pairs as the base content set for constructing style transfer examples.

Stylized Pair Construction. Each content sample is expanded into two complementary types of stylized pairs. To ensure the quality and correctness of these constructed pairs, we employ a hierarchical filtering pipeline. First, for scalable initial screening, we leverage a suite of advanced open-source VLMs (from the Qwen2.5-VL [2], 7B to 72B) to programmatically identify and remove clear failures. Subsequently, pairs that pass this filter undergo a more refined assessment by proprietary models, GPT-5 [1] and Gemini-2.5 [8], for more nuanced judgment. Finally, all ambiguous cases flagged by models, along with randomly

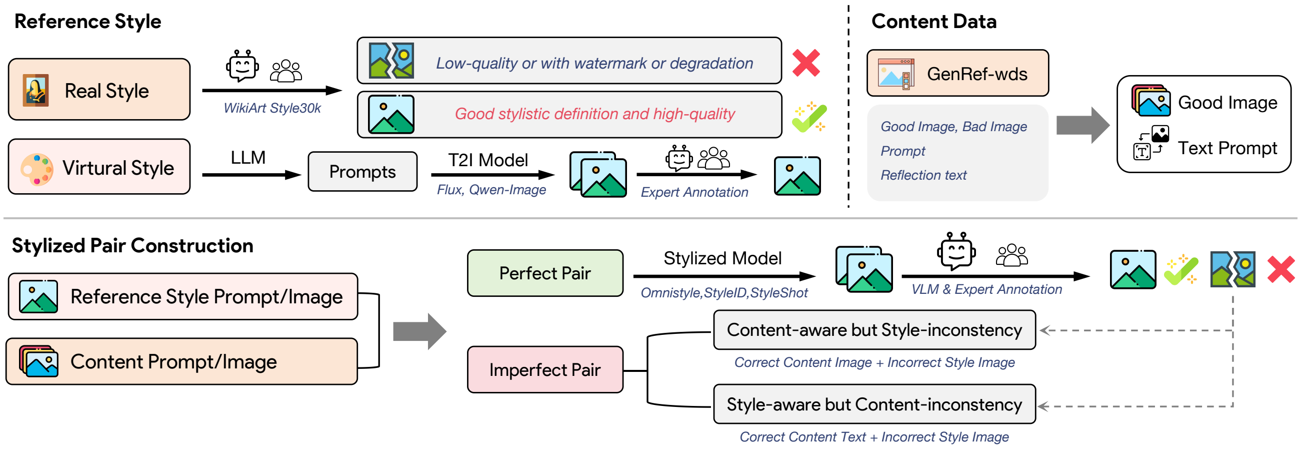
Reference Style
Real Style
WikiArt Style30k
Low-quality or with watermark or degradation
Good stylistic definition and high-quality
Virtural Style
LLM
Prompts
T2I Model
Flux, Qwen-Image
Expert Annotation
Content Data
GenRef-wds
Good Image, Bad Image
Prompt
Reflection text
Good Image
Text Prompt
Stylized Pair Construction
Reference Style Prompt/Image
Content Prompt/Image
Perfect Pair
Stylized Model
Omnistyle,StyleID,StyleShot
VLM & Expert Annotation
Imperfect Pair
Content-aware but Style-inconstency
Correct Content Image + Incorrect Style Image
Style-aware but Content-inconstency
Correct Content Text + Incorrect Style Image

Figure 2. **Data curation pipeline**. Our process begins by sourcing rigorously filtered Reference Styles and Content Data. These are then used to generate and construct stylized pairs with a contrastive objective: each validated *Perfect Pair* is contrasted with *Imperfect Pairs* designed to isolate specific failure modes, namely style inconsistency and content inconsistency.

sampled subsets, are subjected to manual verification by human experts to ensure high quality and resolve edge cases.

- **Perfect Pairs.** High-quality stylized images are generated using state-of-the-art models (e.g., Omnistyle [57], StyleID [31]) and validated by expert annotators for consistent stylistic and semantic fidelity.
- **Imperfect Pairs:** These pairs, which align with common failure modes, are categorized into two types:
  1. **Content-aware but Style-Inconsistent:** Samples retaining correct content but failing on style, exhibiting discrepancies in color tone or texture.
  2. **Style-aware but Content-Inconsistent:** Samples with correct style but deviating content (*semantic drift*). These are generated by altering the content prompt with the style images.

This contrastive construction scheme allows the dataset to explicitly encode style–content disentanglement, supporting both supervised and preference-based training objectives. It also reflects the realistic challenges faced by modern diffusion models, where achieving global style coherence often comes at the cost of semantic accuracy. The complete curation pipeline is illustrated in Figure 2, with additional details provided in the *Appendix*.

4. Method: StyleScore

Developing a reliable reward signal for style transfer uniquely presents a significant challenge, requiring simultaneously assessing *style fidelity* and *content preservation*. Existing general VLM reward models [2, 42] fail in this specific task, struggling to capture the fine-grained style-content decoupling trade-off unique to style transfer, often relying on complex, hallucination-prone prompt engineering. StyleScore is explicitly trained on the STYLEREWARD-DATASET's adversarial pairs to function as a semantic-level verifier, providing a single, unified reward signal that precisely quantifies this trade-off. We detail the model's architecture, based on a frozen VLM backbone with multilayer perceptron reward head, in § Section 4.1, and optimization on preference data in § Section 4.2.

4.1. Reward Model Design

Reward Formation. Reward models are a key component for aligning model outputs with human preferences. Typically, a reward model starts with a pretrained LLM/VLM ϕ and where the LLM head h_i is replaced with a linear reward head l_i, enabling the model to output a scalar reward value. These models are trained using human-provided pairwise comparisons and using a binary cross-entropy loss based on the Bradley-Terry [4] model,

$$P(y_w \succ y_l|\mathbf{x}) = \sigma(r_\phi(y_w|\mathbf{x}) - r_\phi(y_l|\mathbf{x})) \qquad (1)$$

where σ is the sigmoid function. Given a query $\mathbf{x} = (c, x_c)$ which c is the instruction and x_c is the content image, a preferred response y_w and a less preferred response y_l, the reward model is optimized to assign higher rewards to preferred responses:

$$\mathcal{L}_{\text{Reward}}(\theta) = \mathbb{E}_{\mathbf{x},\, y_w,\, y_l}\left[-\log \sigma(r(y_w \mid \mathbf{x}) - r(y_l \mid \mathbf{x}))\right] \qquad (2)$$

Architecture. VLMs are widely used in downstream tasks such as image classification and tagging, owing to their powerful representational capabilities [42]. Thus, we employ Qwen2.5-VL-7B [2] as the backbone. To evaluate a given response, the model jointly processes the query $\mathbf{x} = (c, x_c)$ and the response image y. These multi-modal inputs are tokenized and fed into the backbone to extract a final hidden state representation $\mathbf{h}_{\text{final}} \in \mathbb{R}^d$.

To produce a scalar reward, we replace the model's original language modeling head with a custom reward head. Inspired by BaseReward [70], we design the reward head as

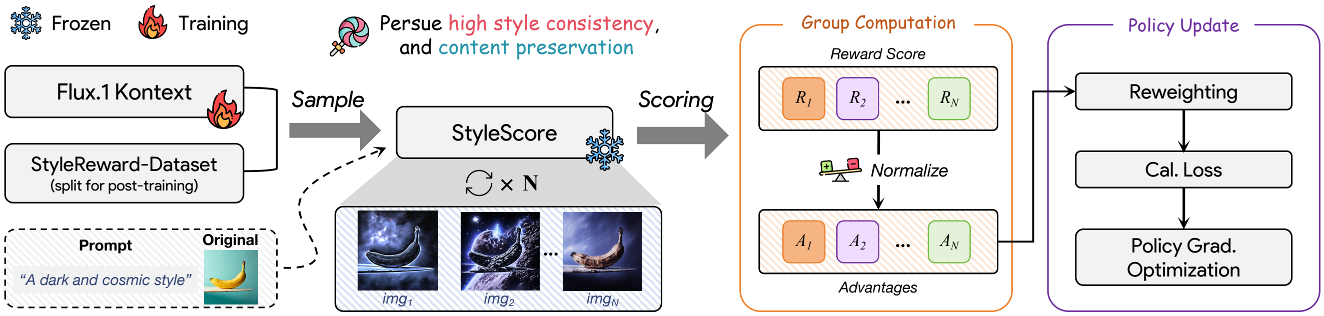

Figure 3. **Overview of the post-training pipeline.** In each iteration, we employ the fine-tuned model based [30] to sample multiple stylistic candidates. These are then adjudicated by our frozen StyleScore reward model, and the resulting scores are normalized into advantages to drive the policy update. This online reinforcement learning cycle allows the model to directly learn the nuanced trade-offs between style consistency and content preservation, moving beyond the supervision of the SFT stage.

a two-layer multilayer perceptron (MLP). Specifically, the first linear layer projects the hidden state $\mathbf{l}_{\text{final}}$ into an intermediate space, followed by a SiLU activation:

$$\begin{aligned} l_{\text{act}} &= \text{SiLU}(\mathbf{W}_1 \mathbf{h}_{\text{final}} + \mathbf{b}_1) \\ r_\phi(y \mid \mathbf{x}) &= \mathbf{W}_2 l_{\text{act}} + \mathbf{b}_2 \end{aligned} \tag{3}$$

where $\mathbf{W}_1 \in \mathbb{R}^{k \times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{1 \times k}$ are the weight matrices, k is the intermediate dimension, and $\mathbf{b}_1$, $\mathbf{b}_2$ are the bias terms. During fine-tuning, only the parameters of this MLP reward head and other lightweight components are updated, while the backbone parameters are kept frozen.

4.2. Reward Model Training

For developing **StyleScore**, we employed a fine-tuning approach on the Qwen2.5-VL-7B with LoRA [22] using STYLEREWARD-DATASET. Each training instance is structured as a tuple (c, x_c, y_w, y_l), where c is the edit instruction, and x_c is the content image, respectively. Correspondingly, y_w and y_l denote the perfect and degraded images.

Reward Calculation and Objective. The training process involves two independent forward passes through the reward model ϕ for each instance. The chosen pair $(\mathbf{x}, y_w)$ is processed by the model to produce a sequence of token-level reward scores, $R_w \in \mathbb{R}^{L_w}$. The rejected pair $(\mathbf{x}, y_l)$ is processed to produce $R_l \in \mathbb{R}^{L_l}$. As defined by our architecture (§ 4.1), these scores are the output of the MLP reward head. We then extract the final scalar reward for each sequence by taking the score associated with the last token:

$$r_i = r_\phi(y_i \mid \mathbf{x}) = R_i[-1], \quad \text{for } i \in \{w, l\} \tag{4}$$

The model's trainable parameters θ are then optimized by minimizing the Bradley-Terry preference loss, as defined in Equation 2. This objective maximizes the margin between the chosen and rejected rewards:

$$\mathcal{L}_{\text{Reward}}(\theta) = \mathbb{E}_{(\mathbf{x},\, y_w,\, y_l) \sim \mathcal{D}} \left[-\log \sigma(r_w - r_i) \right]. \tag{5}$$

Reward Calculation and Objective. Our training objective is to optimize the reward model ϕ such that it assigns higher scores to preferred responses over less-preferred ones, consistent with the collected human preference data. This is achieved by minimizing the Bradley-Terry preference loss, as previously introduced in Equation 2.

5. Two-Stage Fine-Tuning: Style-GRPO

5.1. Supervised Fine-tuning

Recent post-training algorithms, such as PPO [3, 50] or DPO [55], have substantially improved model alignment and fine-tuning efficiency. However, these methods inherently assume that the optimization objectives remain within the pre-trained model's distribution. While this assumption holds for tasks like aesthetic ranking, it becomes restrictive for style transfer, where the target distribution spans diverse and unseen artistic domains. Our preliminary experiments confirm that existing edit models have a limited understanding of style semantics, stemming from a pre-training paradigm reliant on limited descriptive text–image pairs rather than explicit style-conditioned supervision. Consequently, the model struggles to generalize to novel or composite styles, often resulting in style leakage and inconsistent stylization. This shortcoming presents a significant obstacle for subsequent RL-based post-training.

To address this fundamental distribution gap and establish a viable starting policy for RL, our approach begins with supervised fine-tuning on the STYLEREWARD-DATASET. This initial stage is designed to adapt the pre-trained model to the specific domain of style transfer, enhancing its understanding of diverse artistic styles. We adopt FLUX.1[Kontext] [30] as our base model, and the SFT objective is formulated as:

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{t,\, z \sim p_t(z \,|\, c)} \Big[\big\| v_\theta(z, t, c) \; - \; u_t(z \,|\, c) \big\|_2^2 \Big], \tag{6}$$

Here, z denotes latent variable sampled from the interpolated distribution $p_t(z \,|\, c)$, $t \in [0, 1]$ is diffusion time step, and $u_t(z \mid c)$ represents the target velocity field that guides the sample toward data distribution conditioned on c.

5.2. Post-training with GRPO

While the SFT stage adapts the model to the target style domain, it struggles to achieve the fine-grained style–content disentanglement necessary for high-fidelity stylization. To overcome this limitation, we introduce **Style-GRPO**, the second stage of our pipeline. This online RL framework refines the SFT model by optimizing for a more nuanced objective, using StyleScore as the reward function to quantitatively evaluate both stylization quality and content fidelity.

To implement this framework, we first follow [40] and adopt an ODE-to-SDE conversion strategy. This reformulates the deterministic ODE sampler into a stochastic one, yielding a policy π_θ suitable for exploration. Given this policy, we then apply group relative policy optimization (GRPO) [19] to optimize the policy using rewards from StyleScore. This approach notably improves memory and sample efficiency by forgoing the need for an auxiliary value network. For each editing prompt c, we sample G trajectories $\{x_0^i\}_{i=1}^G$ from the policy and compute the corresponding rewards $R(\hat{x}_0; x_0, c)$. Group-normalized advantages $\hat{A}_t^i$ are calculated as:

$$\hat{A}_t^i = \frac{R(\hat{x}_0^i; x_0^i, c) - \mathrm{mean}(\{R(\hat{x}_0^j; x_0^j, c)\}_{j=1}^G)}{\mathrm{std}(\{R(\hat{x}_0^j; x_0^j, c)\}_{j=1}^G)}, \qquad (7)$$

The policy is then updated using a clipped objective:

$$\begin{aligned}\mathcal{L}_{\text{Style-GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\Big(&\min(r_t^i\hat{A}_t^i,\, \mathrm{clip}(r_t^i)\hat{A}_t^i) \\ &- \beta D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})\Big)\Bigg], \\ \mathrm{clip}(r_t^i) = \mathrm{clip}(r_t^i, 1-\epsilon, 1+\epsilon).& \end{aligned} \qquad (8)$$

where $r_t^i(\theta) = \frac{p_\theta(x_{t-1}^i|x_t^i,c)}{p_{\theta_{\text{old}}}(x_{t-1}^i|x_t^i,c)}$ is denotes as the probability ratio. The KL term ensures that the policy π_θ remains close to the reference policy π_{ref}, preventing overfitting to the reward signal or reward hacking [13]. Furthermore, recognizing that early denoising steps are more critical for diverse global styles, we adopt a timestep-aware reward weighting strategy inspired by [21, 33]. We apply an exponential decay $w(t) = \alpha^{t/T}$ to prioritize the reward signal during these initial and style-critical timesteps, which enhances style-content disentanglement and improves image quality.

6. Experiments

6.1. Experimental Settings

Implementation Details. For our proposed reward model, StyleScore is trained by fine-tuning the Qwen2.5-VL-7B-Instruct [2] backbone with LoRA [22] on our StyleReward-Dataset. This stage uses a learning rate $5e^{-5}$ with a batch size 32 and the LoRA rank is set to 64. To enhance the generative model's style transfer capabilities, we fine-tune the base model Flux.1[Kontext] [30] in two phases. First, the SFT stage is conducted on the 'perfect pairs' subset of our dataset using the LoRA rank 128 with a batch size of 32 and the batch size is set to 32. Second, for the online reinforcement learning, we follow [40] and utilize a GRPO setup where we adopt LoRA-based fine-tuning for GRPO training, with a LoRA rank of 128, a learning rate of $5e^{-4}$, an importance clipping range of 1×10^{-4}, a group size of 16, and a KL-penalty coefficient of 0.01. We employ StyleScore, CLIP Score [48] and Aesthetic Score as the reward signals in the RL stage and the resolution of all input image is 1024×1024 in the whole training pipeline. All models are trained and evaluated on 8 × NVIDIA H200 GPUs. Further details are provided in the *Appendix*.

Evaluation Baselines. We compare our approach to existing instruction-guided style transfer methods, including image editing models such as InstructPix2Pix [5] and Flux Kontext [30], as well as text-guided style transfer methods like StyleBooth [20], Omnistyle [57], and DiffStyler [23].

Evaluation Metrics. The stylization quality of our method is assessed on the ImgEdit [65] and AnyEdit [66] benchmarks. ImgEdit employs a GPT-4o-powered score evaluation of content preservation, style consistency and aesthetics. AnyEdit offers quantitative metrics, using CLIP [48] and DINO [52] as metrics utilized in prior works [51, 53].

6.2. Quantitative Experiments

As shown in Table 1, it demonstrates that our Style-GRPO framework achieves state-of-the-art performance across diverse benchmarks. On the LLM-judged ImgEdit benchmark, our method surpasses all baselines under both GPT-4o and Gemini-2.5-Pro evaluations, indicating superior overall quality and human preference alignment. The results on AnyEdit offer a more granular insight into our model's core strengths. Notably, our method achieves the lowest L1 Distance and highest DINO similarity by a significant margin, showcasing its exceptional ability to preserve content structure and semantic fidelity which is a critical challenge in style transfer. This is achieved while simultaneously attaining the highest CLIP$_{img}$ score, confirming top-tier style application. It is worth noting that our CLIP$_{text}$ score is slightly lower than some baselines. This is an expected behavior and highlights a favorable trade-off: our model prioritizes fine-grained visual cues from the reference image over generic textual descriptions. Unlike baselines that may overfit to broad text priors at the cost of specific style details or content structure, our approach ensures that the generated result faithfully mirrors the unique artistic traits of the reference style. This balanced excellence validates the effectiveness of our two-stage pipeline,

Table 1. **Quantitative comparison with the state-of-the-art methods on the public and proposed benchmarks.** We highlight the best scores with the **bold**, and second-best scores with underlined, respectively.

Method	**ImgEdit** [65]		**AnyEdit** (Global style transfer) [66]				↑ **StyleScore**
	↑ `GPT-4o` [1]	↑ `Gemini-2.5-Pro` [8]	↑ CLIP$_{img}$ [48]	↑ CLIP$_{text}$ [48]	↓ L1 Distance	↑ DINO [52]	
InstructP2P [5]	3.55	2.65	0.8260 ± 0.1461	0.1717 ± 0.0369	$\underline{0.1550 \pm 0.0756}$	0.7104 ± 0.0952	3.21
DiffStyler [23]	1.51	1.65	0.4900 ± 0.0788	$\underline{0.1889 \pm 0.0221}$	0.2395 ± 0.0697	0.5875 ± 0.0612	2.03
StyleBooth [20]	4.33	3.88	0.8221 ± 0.1037	$\mathbf{0.1986 \pm 0.0346}$	0.2075 ± 0.0604	0.7230 ± 0.0845	3.46
Omnistyle [57]	3.77	2.38	0.7590 ± 0.0689	0.1797 ± 0.0242	0.1907 ± 0.0497	0.6981 ± 0.0755	2.96
FLux.1 Kontext [30]	$\underline{4.55}$	$\underline{4.29}$	$\underline{0.8215 \pm 0.1319}$	0.1857 ± 0.0351	0.2457 ± 0.1255	$\underline{0.7311 \pm 0.0832}$	$\underline{3.77}$
Ours	**4.74**	**4.46**	$\mathbf{0.8452 \pm 0.0608}$	0.1664 ± 0.0257	$\mathbf{0.0944 \pm 0.0285}$	$\mathbf{0.7583 \pm 0.0912}$	**3.91**

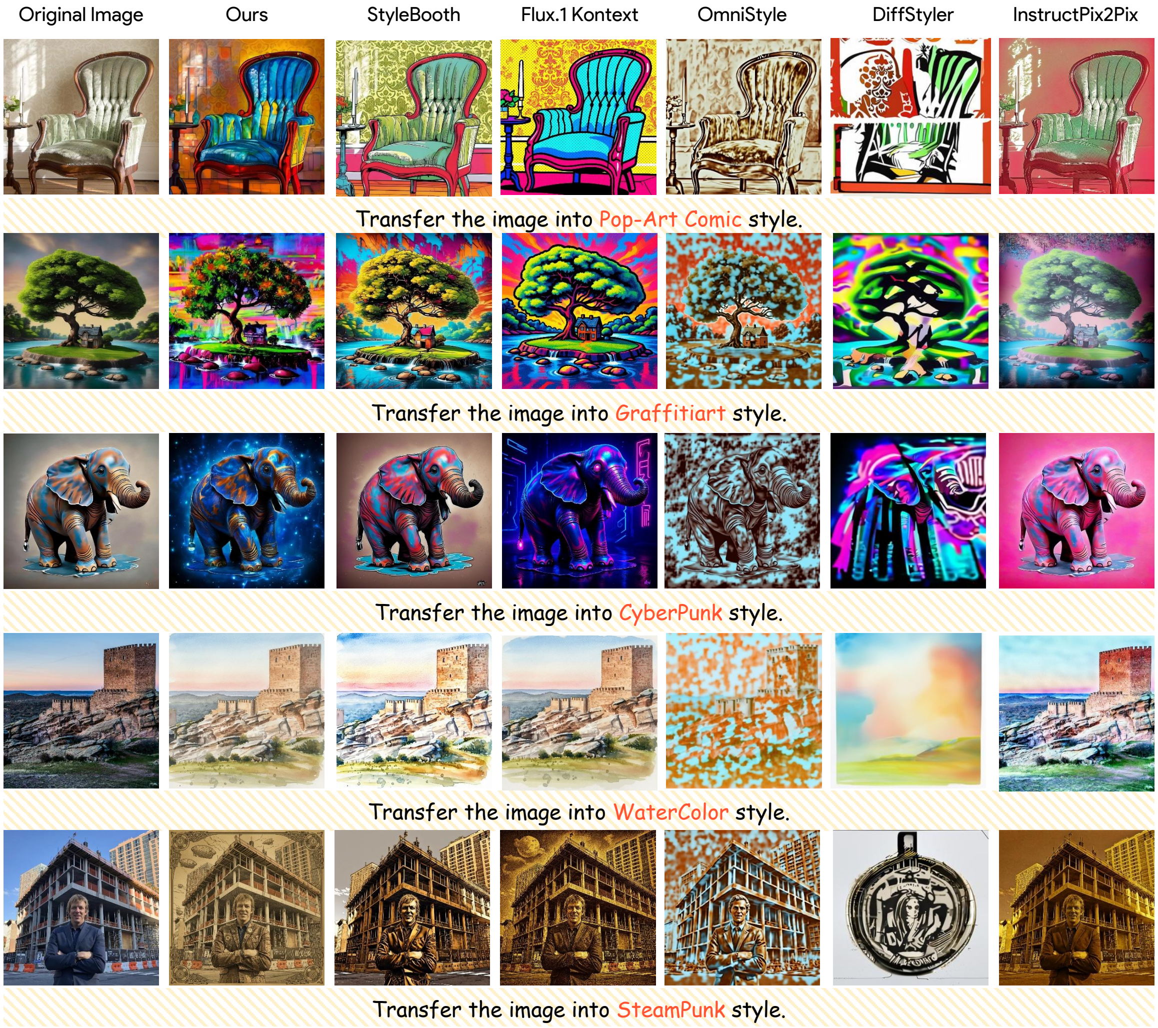
Original Image
Ours
StyleBooth
Flux.1 Kontext
OmniStyle
DiffStyler
InstructPix2Pix
Transfer the image into Pop-Art Comic style.
Transfer the image into Graffitiart style.
Transfer the image into CyberPunk style.
Transfer the image into WaterColor style.
Transfer the image into SteamPunk style.

Figure 4. Qualitative comparison on the prompts with different styles.

Table 2. Reward model preference accuracy.

	Qwen2.5-VL [2]	**ImageReward** [62]	**Ours**
Accuracy ↑	65.2%	48.7%	**98.6%**

Table 3. The user study for style transfer tasks.

Method	**Omnistyle**	**DiffStyler**	**StyleBooth**	**InstructP2P**	**Flux Kontext**	**Ours**
Rank 1	0.2%	0.1%	1.1%	0.8%	10.3%	**87.5%**

where the GRPO stage, guided by StyleScore, masterfully navigates the trade-off between style fidelity and content preservation. Finally, our model predictably earns the highest score on our internal StyleScore benchmark, confirming its strong alignment with our target objectives.

To validate StyleScore, we evaluate its preference accuracy on 500 test pairs split from the propose dataset. As presented in Table 2, StyleScore achieves 98.6% accuracy, demonstrating a significant advantage over general-purpose VLMs like Qwen2.5-VL [2]. This superiority stems from its specialized training on our constructed dataset, enabling it to discern subtle style and content deviations that generalist models overlook. This confirms STYLESCORE's reliability as a reward signal for our Style-GRPO framework.

6.3. Qualitative Experiments

The qualitative results in Figure 4 compellingly illustrate how our Style-GRPO framework resolves the central trade-off between style fidelity and content preservation, a challenge where competing methods consistently falter. Baselines typically fail at one of two extremes: DiffStyler and OmniStyle, sacrifice structural integrity for style application, leading to catastrophic content corruption (e.g., *the 'Watercolor' castle*). Others like InstructP2P, error in the opposite direction, preserving content but exhibiting only weak or insufficient stylization. Even stronger methods like StyleBooth and Flux.1 Kontext struggle to find a balance, often succumbing to semantic drift where the subject's identity is subtly altered to fit the new aesthetic (e.g., *the 'Pop-Art' chair*). In contrast, our method consistently excels. By leveraging STYLESCORE, Style-GRPO learns to decouple these competing objectives, simultaneously rendering the target style with fine-grained precision while rigorously maintaining the semantic and structural integrity of source images. The visual results strongly corroborate our quantitative findings, validating our approach as a principled solution to the style-content disentanglement problem. However, to formally validate these qualitative observations and mitigate potential author bias, we conduct a comprehensive user study, detailed in the following section.

6.4. User Study

We conducted a large-scale blind user study to assess human preference. 36 participants ranked the outputs of our method against baselines on 50 diverse prompts, evaluating for style consistency, content preservation, and overall quality. As shown in Table 3, our Style-GRPO was the clear winner, selected as Rank 1 in 87.5% of cases. This result provides strong evidence that our method better aligns with human perceptual standards for high-fidelity style transfer.

6.5. Ablation Study

Effect of SFT and Post-Training Stages. To dissect the contributions of each component in our pipeline, we conducted an ablation study with results in Table 4. We observe that applying either the SFT stage or the GRPO post-training stage individually yields significant improvements over the original Flux.1 [Kontext] baseline, confirming that both are highly effective optimization strategies.

Table 4. Ablation study on different training variants and stages.

Method	ImgEdit [65]		StyleScore
	GPT-4o	Gemini	
Flux.1 [Kontext] [30]	4.55	4.29	3.77
+SFT	4.67	4.34	3.82
+SFT + Post-Training	**4.74**	**4.46**	**3.91**
+Post-Training (only)	4.68	4.30	3.85

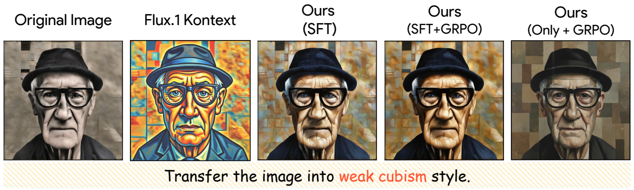

Figure 5. The performance on different training stages.

More revealingly, applying GRPO directly to the baseline can achieve comparable or even superior performance to the SFT-only stage, highlighting the power of direct preference optimization guided by our STYLESCORE. However, our full pipeline, which sequentially combines SFT with post-training, consistently achieves the best overall performance shown as in Figure 5. This key result demonstrates that while GRPO is a potent tool, its effectiveness is maximized when initialized from a domain-adapted policy. The SFT stage provides a more stable and knowledgeable foundation, enabling the subsequent GRPO phase to explore more effectively and converge to a superior final model. Therefore, these results validate our two-stage design as the optimal configuration, where the strategic combination of domain adaptation and preference refinement unlocks state-of-the-art performance.

7. Conclusion

This work introduces a novel framework that significantly advances reference-guided style transfer by addressing the core challenge of style-content disentanglement.Our solution is built upon three synergistic contributions:the large-scale, adversarially-constructed STYLEREWARD-DATASET, which provides the necessary fine-grained supervision; STYLESCORE, a reward model trained on this dataset to accurately reflect human judgments on this specific task; and the Style-GRPO pipeline, which leverages STYLESCORE to effectively refine a powerful generative model. Extensive experiments validate our approach, demonstrating improvements in both style fidelity and content preservation over strong baselines. While the current work focuses on text-guided style transfer, a key direction for future research is to extend our framework to accept reference images as style prompts. This would not only enhance the model's versatility but also expand possibilities for detailed artistic control and creative use cases.

8. Acknowledgment

This work was supported in part by Natural Science Foundation of China (No.62332002, No.62202014), and Shenzhen KQTD (No.20240729102051063)

References

[1] Gpt-4o. https://openai.com/index/hello-gpt-4o/. 3, 7

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 4, 6, 7, 8

[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, 5

[4] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 4

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 1, 2, 6, 7

[6] Laura Jannes Burger. Laion: Image data, ai, and dispossession. Master's thesis, 2023. 3

[7] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025. 1, 2, 3

[8] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3, 7

[9] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 2

[10] 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. 1

[11] Chaoran Feng, Wangbo Yu, Xinhua Cheng, Zhenyu Tang, Junwu Zhang, Li Yuan, and Yonghong Tian. Ae-nerf: Augmenting event-based neural radiance fields for non-ideal conditions and larger scene. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2924–2932, 2025. 2

[12] Junyao Gao, Yanan Sun, Yanchen Liu, Yinhao Tang, Yanhong Zeng, Ding Qi, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[13] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023. 2, 6

[14] Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025. 1

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2

[16] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3985–3993, 2017. 2

[17] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023. 2

[18] Yuan Gong, Xionghui Wang, Jie Wu, Shiyin Wang, Yitong Wang, and Xinglong Wu. Onereward: Unified mask-guided image generation via multi-task human preference learning. *arXiv preprint arXiv:2508.21066*, 2025. 3

[19] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 6

[20] Zhen Han, Chaojie Mao, Zeyinzi Jiang, Yulin Pan, and Jingfeng Zhang. Stylebooth: Image style editing with multimodal instruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1947–1957, 2025. 2, 3, 6, 7

[21] 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. 2, 6

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5, 6

[23] Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Weiming Dong, and Changsheng Xu. Diffstyler: Controllable dual diffusion for text-driven image stylization. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 2, 6, 7

[24] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2

[25] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025. 1, 2

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[27] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6007–6017, 2023. 2

[28] 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. 2

[29] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1, 3

[30] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 1, 2, 5, 6, 7, 8

[31] Minh-Ha Le and Niklas Carlsson. Styleid: Identity disentanglement for anonymizing faces. *arXiv preprint arXiv:2212.13791*, 2022. 2, 4

[32] Jialuo Li, Wenhao Chai, Xingyu Fu, Haiyang Xu, and Saining Xie. Science-t2i: Addressing scientific illusions in image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2734–2744, 2025. 1, 2

[33] Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025. 2, 6

[34] Wen Li, Muyuan Fang, Cheng Zou, Biao Gong, Ruobing Zheng, Meng Wang, Jingdong Chen, and Ming Yang. Styletokenizer: Defining image style by a single instance for controlling diffusion models. In *European Conference on Computer Vision*, pages 110–126. Springer, 2024. 1, 2, 3

[35] Xiaomin Li, Yixuan Liu, Takashi Isobe, Xu Jia, Qinpeng Cui, Dong Zhou, Dong Li, You He, Huchuan Lu, Zhongdao Wang, et al. Reneg: Learning negative embedding with reward guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23636–23645, 2025. 2

[36] Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, and Li Yuan. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025. 1, 3

[37] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025. 2

[38] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024. 2

[39] Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Yibo Wang, Xintao Wang, Yujiu Yang, and Ying Shan. Stylecrafter: Enhancing stylized text-to-video generation with style adapter. *arXiv preprint arXiv:2312.00330*, 2023. 2

[40] 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, 6

[41] Xin Luo, Jiahao Wang, Chenyuan Wu, Shitao Xiao, Xiyan Jiang, Defu Lian, Jiajun Zhang, Dong Liu, et al. Editscore: Unlocking online rl for image editing via high-fidelity reward modeling. *arXiv preprint arXiv:2509.23909*, 2025. 3

[42] 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. 2, 4

[43] Saif Mohammad and Svetlana Kiritchenko. Wikiart emotions: An annotated dataset of emotions evoked by art. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018. 3

[44] Yuwei Niu, Weiyang Jin, Jiaqi Liao, Chaoran Feng, Peng Jin, Bin Lin, Zongjian Li, Bin Zhu, Weihao Yu, and Li Yuan. Does understanding inform generation in unified multimodal models? from analysis to path forward. *arXiv preprint arXiv:2511.20561*, 2025. 1

[45] Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

[46] Yatian Pang, Peng Jin, Shuo Yang, Bin Lin, Bin Zhu, Chaoran Feng, Zhenyu Tang, Liuhan Chen, Francis EH Tay, Ser-Nam Lim, et al. Next patch prediction for autoregressive visual generation. In *AAAI 2026*, 2024.

[47] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6, 7

[49] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 2

[50] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 5

[51] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman.

Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024. 2, 6

[52] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 6, 7

[53] Xinghai Sun, Changhu Wang, Avneesh Sud, Chao Xu, and Lei Zhang. Magicbrush: Image search by color sketch. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 475–476, 2013. 6

[54] Zhenyu Tang*, Chaoran Feng*, Xinhua Cheng, Wangbo Yu, Junwu Zhang, Yuan Liu, Xiaoxiao Long, Wenping Wang, and Li Yuan. Neuralgs: Bridging neural fields and 3d gaussian splatting for compact 3d representations. In *AAAI 2026*, 2025. 2

[55] 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. 2, 5

[56] Jing Wang, Jiajun Liang, Jie Liu, Henglin Liu, Gongye Liu, Jun Zheng, Wanyuan Pang, Ao Ma, Zhenyu Xie, Xintao Wang, Meng Wang, Pengfei Wan, and Xiaodan Liang. Grpoguard: Mitigating implicit over-optimization in flow matching via regulated clipping, 2025. 2

[57] Ye Wang, Ruiqi Liu, Jiang Lin, Fei Liu, Zili Yi, Yilin Wang, and Rui Ma. Omnistyle: Filtering high quality style transfer data at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7847–7856, 2025. 1, 2, 3, 4, 6, 7

[58] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 1, 2, 3

[59] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 1, 2

[60] Yi Wu, Lingting Zhu, Shengju Qian, Lei Liu, Wandi Qiao, Lequan Yu, and Bin Li. Stylear: Customizing multimodal autoregressive model for style-aligned text-to-image generation. *arXiv preprint arXiv:2505.19874*, 2025. 1, 2

[61] Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski. Stylealign: Analysis and applications of aligned stylegan models. *arXiv preprint arXiv:2110.11323*, 2021. 2

[62] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 2, 7

[63] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. 2

[64] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *International Conference on Machine Learning*, 2024. 1

[65] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025. 6, 7, 8

[66] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26125–26135, 2025. 6, 7

[67] Wangbo Yu*, Chaoran Feng*, Jianing Li, Aofan Zhang, Zhenyu Tang, Mingyi Guo, Wei Zhang, Zhengyu Ma, Li Yuan, and Yonghong Tian. Ea3d: Event-augmented 3d diffusion for generalizable novel view synthesis. In *The Fourteenth International Conference on Learning Representations*, 2026. 1

[68] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2

[69] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models, 2023. 2

[70] Yi-Fan Zhang, Haihua Yang, Huanyu Zhang, Yang Shi, Zezhou Chen, Haochen Tian, Chaoyou Fu, Haotian Wang, Kai Wu, Bo Cui, et al. Basereward: A strong baseline for multimodal reward model. *arXiv preprint arXiv:2509.16127*, 2025. 3, 4

[71] Yian Zhao, Rushi Ye, Ruochong Zheng, Zesen Cheng, Chaoran Feng, Jiashu Yang, Pengchong Qiao, Chang Liu, and Jie Chen. Tune-your-style: Intensity-tunable 3d style transfer with gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19032–19042, 2025. 1

[72] Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025. 1

[73] Yujie Zhou, Pengyang Ling, Jiazi Bu, Yibin Wang, Yuhang Zang, Jiaqi Wang, Li Niu, and Guangtao Zhai. G2rpo: Granular grpo for precise reward in flow models. *arXiv preprint arXiv:2510.01982*, 2025. 2

[74] Le Zhuo, Liangbing Zhao, Sayak Paul, Yue Liao, Renrui Zhang, Yi Xin, Peng Gao, Mohamed Elhoseiny, and Hongsheng Li. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15329–15339, 2025. 3