

StyleDoctor: Towards Specialist Reward Model for Style-centric Generation Tasks

Supplementary Material

1. Implementation Details

For the downstream generation tasks, we merely replace the original reward model with our StyleDoctor, while keeping the set of hyperparameters the same as in the previous works. All the experiments are conducted on 80GB A100 GPUs.

For the curation process of the Style Perception Reward Dataset (SPRData), we employ ChatGPT to generate a list of approximately 500 simple object concepts. The prompt for GPT-4o to generate reward modeling textual annotations is: *You are an expert art critic and visual reasoning assistant that performs fine-grained style analysis. Your task is to compare two stylized images (Image A and Image B) against a given reference style image and determine which image better matches the reference in artistic style. You must reason explicitly about stylistic attributes such as color palette, brushstroke texture, composition, and overall mood before concluding which image aligns more closely with the reference style.* During the style understanding evaluation, we use the same prompt for GPT-4o to verify the output for the reward model.

2. Dataset Sample

We have shown two examples from the SPRData in Fig. 1 and Fig. 2.

3. User Study Details

To further validate the effectiveness of StyleDoctor in real user perception, we conduct a human study on four representative style-related generation tasks, including reference image-guided style generation, instruction-guided style transfer, reference image-guided style transfer, and style customization. We recruit 50 participants from diverse backgrounds, and each participant completes 20 pairwise comparisons for each task. In every comparison, the participant is shown a baseline result and the corresponding result enhanced by StyleDoctor, and is asked to select the one that better matches the target style condition. The judgment criterion is style consistency, which mainly reflects the alignment in overall stylistic coherence, color palette, and texture characteristics. As shown in Table 1, StyleDoctor is consistently preferred across all four tasks, indicating that the improvement is not limited to automatic metrics but is also clearly perceived by human evaluators.

4. Additional Reward Analysis

We provide additional analyses to better understand why StyleDoctor is more suitable for style-centric generation than existing alternatives. First, following the rebuttal, we evaluate reward selection accuracy on paired data with user-validated negative samples constructed by weakening the style injection strength. As reported in Table 2, StyleDoctor substantially outperforms CLIP and CSD, showing that it can more reliably distinguish stylistically insufficient results from better aligned ones. We further compare StyleDoctor with more general-purpose reward models, including ImageReward and PickScore. Since these models are mainly designed for human preference or text-image alignment rather than explicit style consistency, their performance is notably lower on both style understanding and downstream style generation, whereas StyleDoctor achieves clear advantages on all dimensions, as shown in Table 3.

5. Model Scaling Study

We additionally evaluate a larger 7B version of StyleDoctor to examine whether the proposed framework benefits from increased model capacity. The results in Table 4 show that scaling from 3B to 7B consistently improves both multimodal style understanding and generation-related evaluation metrics. In particular, the 7B model achieves stronger performance on style understanding and better style/content scores on downstream generation, suggesting that the proposed style-centric reward modeling paradigm scales effectively with model size. Due to computational constraints, we do not further perform RL fine-tuning with the 7B variant in this work.

6. Reinforcement Learning Details

For the reinforcement learning experiments in Sec. 4.2 of the main paper, we adopt Flow-GRPO as the primary optimization framework and use StyleDoctor as the reward model. In our implementation, the reward is constructed from both the intermediate Chain-of-Thought style analysis and the final judgment produced by StyleDoctor. More specifically, we average the rubric scores from the multi-dimensional reasoning process, including aspects such as global style consistency, color palette, brushstroke texture, and composition, and combine them with the final preference judgment to form the reward signal used during training. This design allows StyleDoctor to provide richer supervision than a single scalar preference score. Besides Flow-

User:

Analyze Image 2's and Image 3's style consistency to the source style image (Image 1).



assistant:

<think>

1. Global-view style consistency: Image 2 (7/10) - Image 2 retains a similar atmosphere and mood, with a soft, muted color palette that resonates with the overall impressionistic feel of Image 1. Image 3 (4/10) - While it shares some elements, the bright colors and more defined forms diverge significantly from the softer, more muted tones of Image 1.

2. Color palette consistency: Image 2 (8/10) - The colors in Image 2 closely resemble those in Image 1, with earthy tones that echo the muted palette of Image 1. Image 3 (5/10) - The colors are more vibrant and varied, which detracts from the cohesive feel of Image 1.

3. Brushstroke texture consistency: Image 2 (9/10) - The brushstrokes in Image 2 are loose and expressive, mirroring the impressionistic style of Image 1. Image 3 (3/10) - The brushwork is more precise and defined, lacking the fluidity and spontaneity found in Image 1.

4. Form and composition consistency: Image 2 (7/10) - The forms in Image 2 are somewhat reminiscent of the abstract nature of Image 1, with a focus on shapes that echo the impressionistic style. Image 3 (4/10) - The composition is more structured and rigid, which contrasts with the more organic and dynamic forms in Image 1.

Total score: Image 2: 7 + 8 + 9 + 7 = 31 Image 3: 4 + 5 + 3 + 4 = 16

</think>

<answer>Image 2 has a more similar style to the source style image</answer>

Figure 1. A sample from the SPRData.

User:

Analyze Image 2's and Image 3's style consistency to the source style image (Image 1).



assistant:

<think>

1. Global-view style consistency: Image 2 (8/10) - Image 2 maintains a similar abstract quality and emotional resonance as Image 1, reflecting a similar approach to form and expression. Image 3 (5/10) - While it shares some stylistic elements, the overall composition and execution diverge significantly from the source style, making it less consistent.

2. Color palette consistency: Image 2 (7/10) - The colors in Image 2, although not identical, share some similarities with the muted tones found in Image 1, particularly in the use of greens and blues. Image 3 (4/10) - The color palette is more vibrant and varied, which contrasts with the subdued tones of Image 1, leading to a lower score.

3. Brushstroke texture consistency: Image 2 (8/10) - The brushstrokes in Image 2 exhibit a similar texture and fluidity, aligning well with the expressive brushwork of Image 1. Image 3 (3/10) - The brushstrokes in Image 3 appear smoother and less dynamic, which detracts from its stylistic alignment with Image 1.

4. Emotional resonance: Image 2 (8/10) - The emotional impact of Image 2 resonates with the spirited and emotive qualities of Image 1, creating a sense of connection. Image 3 (4/10) - The emotional tone is more subdued and lacks the energy present in Image 1, reducing its emotional resonance.

Total score:

Image 2 8+7+8+8=31

Image 3 5+4+3+4=16

</think>

<answer>Image 2 has a more similar style to the source style image</answer>

Figure 2. A sample from the SPRData.

Task	Model	User Preference (Baseline vs. +StyDoc.)
Ref-img Sty. Gen.	OmniGen2	0% vs. 100%
Instr. Sty. Transfer	OmniStyle	37% vs. 63%
Ref-img Sty. Transfer	OmniStyle	29% vs. 71%
Customization	B-LoRA	27% vs. 73%

Table 1. User study results on four representative style-related generation tasks.

Model	Acc. (%)
CLIP	36.60
CSD	55.38
StyleDoctor	75.28

Table 2. Reward selection accuracy on paired data with user-validated negative samples.

Model	Understanding	Generation
ImageReward	58.35 / 47.20	0.37 / 0.59
PickScore	53.08 / 46.45	0.52 / 0.43
StyleDoctor	75.46 / 71.14	0.78 / 0.69

Table 3. Comparison with existing reward models on style understanding and generation.

Model Size	Understanding	Generation
3B	71.14	0.78 / 0.69
7B	75.12	0.84 / 0.75

Table 4. Scaling performance of StyleDoctor with different model sizes.

GRPO, we also study a DPO-style setting, where StyleDoctor is only used to judge generated image pairs and provide preference supervision, without introducing additional architectural changes to the diffusion model. In all experiments, we follow the original hyperparameter settings of the corresponding baseline methods and only replace the reward component with StyleDoctor.

We fine-tune OmniGen2 using the Flow-GRPO framework for instruction-guided image editing on SPRData for 1 epoch. Specifically, for each input instruction, we sample 16 candidate images and compute rewards over the group for policy updates. To improve training efficiency, we adopt the Flow-GRPO-Fast setting. Instead of performing full denoising, we train on only 1–2 denoising steps per trajectory. We adopt the AdamW optimizer with a learning rate of 3×10^{-6} , $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay of 1×10^{-4} , and $\varepsilon = 1 \times 10^{-8}$. Gradient clipping is applied with a maximum norm of 1.0. For stochastic trajectory construction, we convert deterministic ODE sampling into SDE sampling using CPS-based noise injection. We set the noise level to

0.8, the SDE window size to 3, and apply the SDE window within the first half of the trajectory. We exclude the final timestep during training by setting the timestep fraction to 0.99 to avoid numerical instability in low-noise regions. The same setup is used for fine-tuning OmniStyle with Flow-GRPO, as this is the general setting from the official Flow-GRPO repository.

7. DPO with SPRData

To further validate the usefulness of SPRData beyond the main experiments, we conduct an additional DPO experiment on text-controlled style generation using SDXL. Owing to computational constraints, this experiment is performed on a randomly sampled subset of 100K examples from the full SPRData. We also manually verify the correctness of the text annotations on small subsets to ensure annotation quality. As shown in Table 5, incorporating Diffusion-DPO with SPRData leads to clear improvements over the original SDXL baseline under all three evaluation protocols, including CSD, GPT-4o, and human preference, which further supports the effectiveness of the proposed dataset for style-centric alignment.

Model	CSD	GPT-4o	Human
SDXL	0.64	67.05	13%
SDXL + Diffusion-DPO	0.75	72.92	87%

Table 5. DPO with SPRData on text-controlled style generation.

8. Reward Fairness Analysis

Because the reference image-guided style generation results in the main paper are evaluated with CSD, we additionally provide a separate comparison to avoid potential unfairness when CSD itself is used as a reward model. Specifically, we compare the original OmniGen2, OmniGen2 optimized with CSD, and OmniGen2 optimized with StyleDoctor under GPT-4o evaluation and human preference. As shown in Table 6, StyleDoctor achieves the best performance under both criteria, while CSD also improves over the plain baseline but remains clearly inferior to StyleDoctor. This result further indicates that the gains of StyleDoctor do not simply come from metric coupling, but from stronger style

understanding and more effective reward guidance.

Model	GPT-4o	User Preference
OmniGen2	-	0%
OmniGen2 + CSD	62.45	38%
OmniGen2 + StyleDoctor	74.80	62%

Table 6. Comparison of different reward models on reference image-guided style generation.