Style-Friendly SNR Sampler for Style-Driven Generation

Jooyoung Choi[1,*] Chaehun Shin[1,*] Yeongtak Oh[1] Heeseung Kim[1]
Jungbeom Lee[2,†] Sungroh Yoon[1,3,†]
[1]Data Science and AI Laboratory, ECE, Seoul National University
[2]Korea University
[3]AIIS, ASRI, INMC, ISRC, and Interdisciplinary Program in AI, Seoul National University
{jy_choi,chaehuny,dualism9306,gmltmd789}@snu.ac.kr,jbeomlee@korea.ac.kr,sryoon@snu.ac.kr

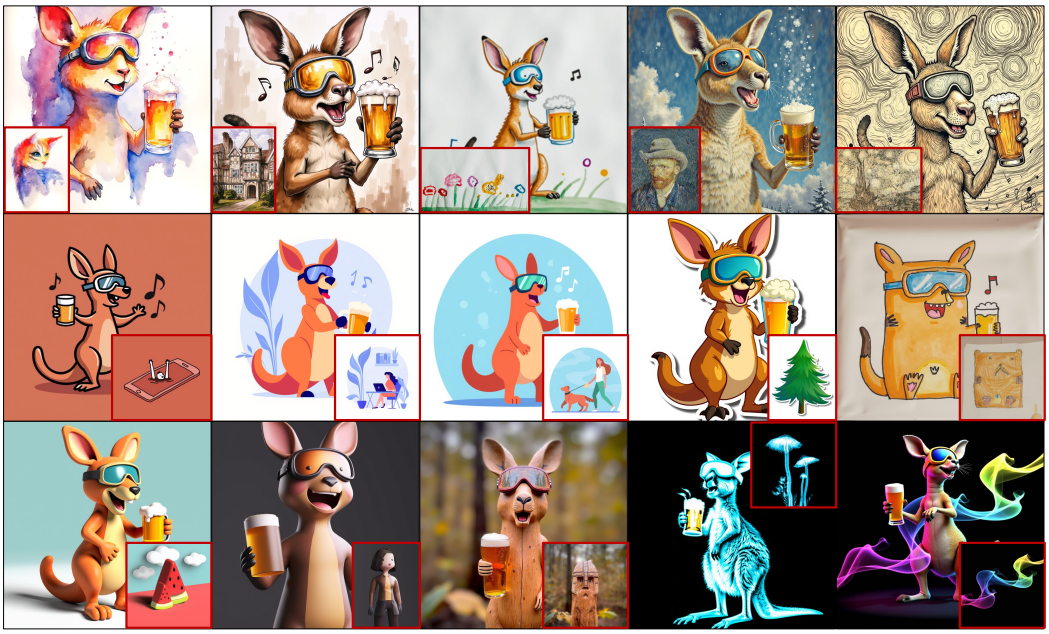


Figure 1. Fine-tuning text-to-image diffusion models on the *style-friendly* noise levels enables learning novel styles from reference images and text prompts. We present 'A kangaroo holding a beer, wearing ski goggles and passionately singing silly songs' in various styles including painting, flat illustration, and 3d rendering styles. References are shown in the red insert box.

Abstract

Recent text-to-image diffusion models generate high-quality images but struggle to learn new styles, which limits the personalized content creation. In response, style-driven generation has become a popular task, wherein users supply reference images capturing the target style, complemented by text prompts that specify stylistic cues. Fine-tuning is a common approach, yet it often blindly utilizes pre-training configurations without modification, especially for noise schedules defined in terms of signal-to-noise ratio (SNR), which determines the amount of image information available at each denoising step. We discover that stylistic features predominantly emerge at low SNR range, leading current fine-tuning methods using regular noise schedules to exhibit suboptimal style alignment. We propose the Style-friendly SNR sampler, which focuses the fine-tuning on low SNR range where stylistic features emerge. We demonstrate improved generation of novel styles that cannot be described solely with a text prompt, enabling high-fidelity personalized content creation.

∗ First authors
† Corresponding authors

1. Introduction

Recently, large-scale text-to-image diffusion models [2, 7, 35] have achieved remarkable progress in visual content creation. In particular, open-weights such as Stable Diffusion series [7, 35] and FLUX [2] have been among the most notable for their photorealistic image quality and language understanding capabilities. Behind this strong performance lies the advancement of the diffusion and flow matching frameworks [27, 28, 47] by noise level scheduling [15, 20], loss weighting [5, 21], and architectural improvements [7, 30, 51]. These advances have largely targeted overall image quality and text-image alignment.

Motivated by the success of text-to-image models, there is a growing need for style-driven generation [24, 37, 43], where the generated samples capture styles desired by individual users or artists. Here, "style" refers to visual properties such as color schemes, layouts, illumination, and brushwork [6, 9, 25, 43]. The most accessible way to request style-driven generation is to write a text prompt describing the target style, but prompts alone rarely suffice when the target style was absent from the model's pre-training data or is hard to describe verbally.

To address this challenge, To address this challenge, recent approaches have explored methods that generate outputs reflecting the style of a user-provided reference image, which fall into two main categories: fine-tuning methods [24, 43] and tuning-free methods [12, 37, 52]. Among these, fine-tuning has remained promising due to its ability to directly adjust model parameters using provided style reference images [38]. Yet existing fine-tuning approaches blindly leverage the noise-sampling schedule that was chosen during pre-training to optimize performance on generic text-to-image benchmarks [10, 18] rather than style-driven generation. This schedule can be expressed by *signal-to-noise ratio* (SNR) [22]: a value that indicates, at each denoising step, how much of the image is still clean signal and how much is noise. We empirically show that high SNR range contributes to sharpening object details and low SNR range contributes to broad color schemes, lighting, and other stylistic cues emerge. Since the pre-training schedule prioritize on high SNR range, using such schedule often fails to reproduce the desired personal style.

We therefore introduce a *Style-friendly SNR sampler*. It prioritize fine-tuning in low SNR range (roughly the first 10% of denoising steps), maximizing exposure to stylistic cues without changing the loss or architecture. With this single modification, diffusion models faithfully learn novel styles while preserving their original text-image alignment.

Through extensive evaluation, we show that our Style-friendly SNR sampler achieves superior performance in capturing novel styles compared to both existing fine-tuning approaches and tuning-free approaches. Comparison against tuning-free baselines confirms that fine-tuning, when guided by our novel SNR-based insight, remains significantly more effective for accurately capturing stylistic details. Furthermore, our analysis provides deeper insight into why existing methods excel at object-centric generation but consistently fail in style-driven contexts. Our method is compatible with diverse backbones such as FLUX [2], SD3.5 [48], and SANA [50], and can be integrated into existing fine-tuning pipelines [24, 40] to provide consistent improvements.

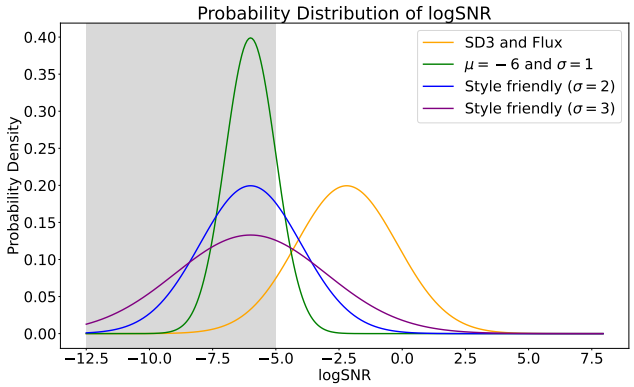


Figure 2. **Probability distribution of Log-SNR.** We bias the fine-tuning towards the shaded SNR range where style features emerge.

2. Training Diffusion Models

2.1. Diffusion Process and SNR Formulation

Diffusion models [14, 27, 28, 42, 47] are based on the forward process that progressively degrades data x_0 into pure noise x_1 as time t progresses from 0 to 1, following the unified formulation below:

$$x_t = \alpha_t x_0 + \sigma_t \epsilon, \tag{1}$$

where α_t and σ_t are predefined noise schedules, and $\epsilon \sim \mathcal{N}(0, I)$ represents standard Gaussian noise.

Recent state-of-the-art flow matching frameworks, such as Stable Diffusion 3 (SD3) [7], FLUX [2], and SANA [50] utilize the noise schedule from rectified flow [27, 28], where $\alpha_t = 1 - t$ and $\sigma_t = t$, with t varying continuously in the range $[0, 1]$. This choice is effective due to straight diffusion trajectories.

Instead of parameterizing the diffusion process using the timestep t, VDM [22] characterize the noise level using the log signal-to-noise ratio (log-SNR), which offers a more intuitive measure of the noise at each step:

$$\lambda_t = \log\left(\frac{\alpha_t^2}{\sigma_t^2}\right). \tag{2}$$

In flow matching frameworks, the log-SNR is simplified as $\lambda_t = 2\log\left(\frac{1-t}{t}\right)$ using timestep t.

2.2. Diffusion Training and SNR Sampling

Recent flow matching frameworks predict the velocity field $v_\theta(x_t, t)$ by minimizing the following training objective:

$$\mathcal{L}_{\text{FM}}(x_0) = \mathbb{E}_{t\sim p(t)}\left[\left\|\epsilon - x_0 - v_\theta(x_t, t)\right\|^2\right], \quad (3)$$

where $p(t)$ is the timestep sampling distribution, $\epsilon - x_0$ is the target velocity derived from Eq. 1. The representative text-to-image models utilizing flow matching, such as SD3 and FLUX, introduce an SNR sampler for training. This samples the *logit* of t, defined as $\log\left(\frac{t}{1-t}\right)$ from $\mathcal{N}(\mu, \sigma^2)$, where the parameters μ and σ are chosen as 0 and 1 to optimize CLIP [33] and FID [13] scores on COCO-2014 validation set [26].

In addition, they propose shifting timestep t to t_{new} by k for high resolution training:

$$t_{\text{new}} = \frac{kt}{1 + (k-1)t}, \quad (4)$$

which is equivalent to shifting λ_t by $-2\log k$ as follows:

$$\lambda_{t_{\text{new}}} = 2\log\left(\frac{1 - t_{\text{new}}}{t_{\text{new}}}\right) = \lambda_t - 2\log k, \quad (5)$$

where k is defined as 3. Following the above formulation, resulting log-SNR sampling distribution $p(\lambda_t)$ in training time is represented as $\mathcal{N}(-2\log 3, 2^2)$, as visualized by the yellow curve in Fig. 2. This curve demonstrates that pre-training SD3 and FLUX focus on particular SNR values.

3. Method

3.1. Observations

Diffusion Models Struggle to Capture Styles. To understand the fine-tuning capability of recent diffusion model [2], we first revisit the commonly-studied object-driven setting, i.e., DreamBooth [38]. For object-driven generation setting, the standard pre-training configuration, such as SNR schedule, is promising: the generated images exhibit high-fidelity of images of the target object such as a backpack (Fig. 3a). However, the standard pre-training configuration is inadequate for style references. In Fig. 3b, when fine-tuned with a "glowing style" reference, the model only applies glowing effects to specific object details (such as the fur of a sloth), neglecting global stylistic elements like the dark background and blue lighting of the original reference. Similarly, with a Van Gogh oil painting style reference, the model manages to replicate the blue color tone but fails to accurately reproduce distinctive brushstroke characteristics. These indicate that capturing an object's identity and capturing an artistic style *require different treatment during fine-tuning*.

(a) FLUX Object (b) FLUX Style (c) Our Style

Figure 3. **Fine-tuning capability.** (a) While FLUX succeeds in learning objects, (b) it struggles to capture styles, demonstrating that learning novel objects and styles requires distinct strategies. (c) We enable FLUX to learn styles. References are shown in the red insert box.

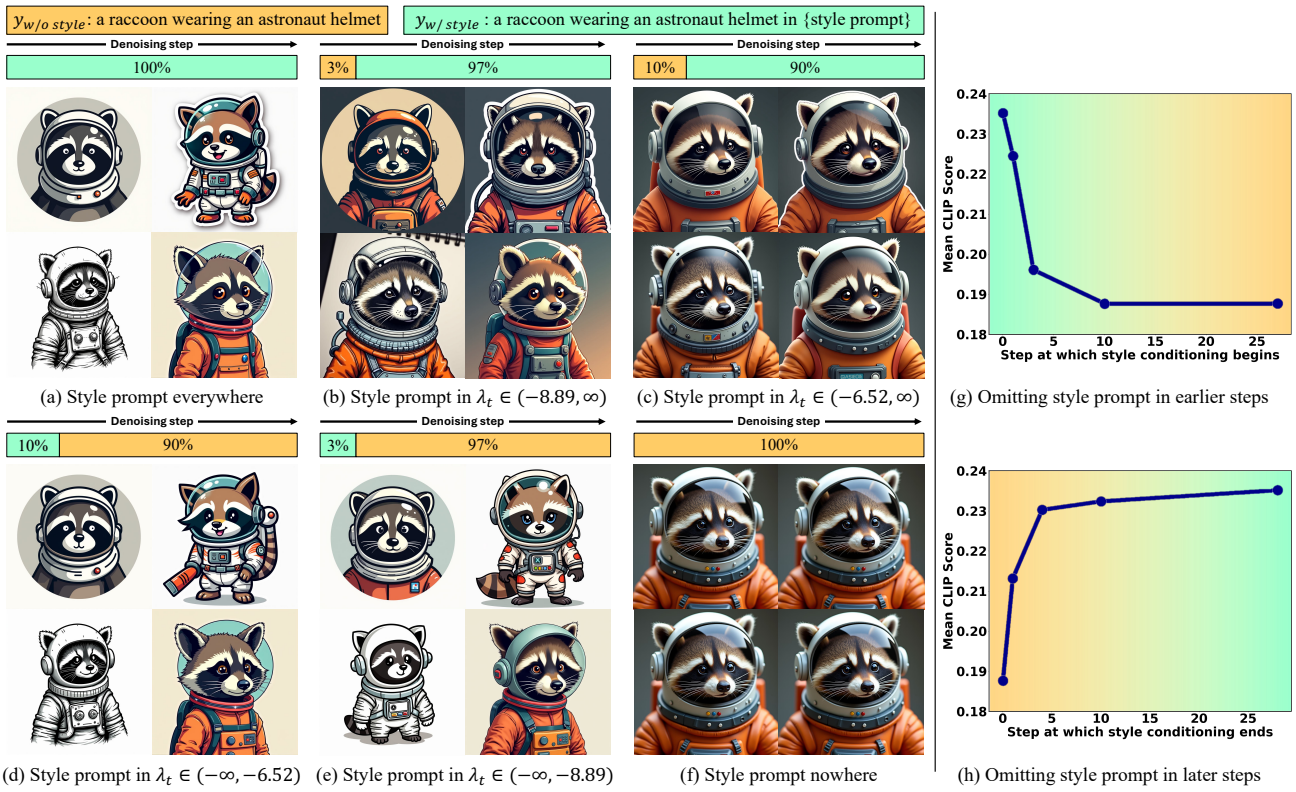
Styles Emerge at Low SNR Range. To better understand why diffusion models struggle to learn new styles, we analyze at which diffusion timestep stylistic features emerge during generation using a pre-trained FLUX [2]. Specifically, we switch from a prompt without style descriptions ($y_{\text{w/o style}}$) to one including style descriptions ($y_{\text{w/ style}}$) at different points in the denoising process. Fig. 4 presents the qualitative examples and quantitative curves showing the averaged CLIP similarity [33] between each generated image and a text describing the target style (e.g., "watercolor painting style").

As shown in Fig. 4c, omitting the style prompt in the initial **10%** of generation steps significantly reduces style alignment. Fig. 4g shows that as we increase the portion of early generation steps where the style prompt is omitted, the CLIP score drops sharply, indicating that style information lost in these early steps cannot be recovered later in the generation process.

Conversely, omitting the style prompt at later denoising steps minimally affects style alignment (Fig. 4d, e). In Fig. 4, as we increase the portion of early steps where the style prompt is included, the CLIP similarity rises steeply and quickly saturates.These results demonstrate that styles are predominantly determined at early denoising steps, corresponding to low log-SNR λ_t values (shaded region in Fig. 2).

3.2. Style-Friendly SNR Sampler

Our earlier observations show that styles primarily emerge during early denoising steps, characterized by low SNRs. However, existing fine-tuning methods use the SNR schdules from pre-training, optimized mainly for object-centric



(a) Style prompt everywhere (b) Style prompt in $\lambda_t \in (-8.89, \infty)$ (c) Style prompt in $\lambda_t \in (-6.52, \infty)$ (g) Omitting style prompt in earlier steps

(d) Style prompt in $\lambda_t \in (-\infty, -6.52)$ (e) Style prompt in $\lambda_t \in (-\infty, -8.89)$ (f) Style prompt nowhere (h) Omitting style prompt in later steps

Figure 4. **Prompt switching during generation.** λ_t indicates log-SNR. The bar graphs above each image represent the denoising steps, illustrating when each prompt is applied and at what point the prompt switch occurs. The style prompts are 'minimalist flat round logo', 'sticker', 'cartoon', and 'detailed pen and ink drawing', clockwise from top left. Styles emerge in the initial 10% of denoising steps; therefore, (c) and (f) fail to capture target styles. In contrast, omitting style prompts in later steps (d, e) still preserves styles well, similar to the fully styled baseline (a). (g) and (h) quantify these observations, showing the average CLIP similarity across 5 prompts and 5 styles when omitting (g) or including (h) the style prompt in earlier steps.

benchmarks [10, 18], as indicated by the yellow curve in Fig. 2. Consequently, standard fine-tuning procedures place insufficient emphasis on SNR range crucial for capturing styles, failing to achieve alignment with reference styles.

Building upon this motivation, we propose to rectify the fine-tuning of diffusion models by biasing towards lower log-SNR λ_t values where stylistic features emerge. Specifically, we sample log-SNR from a normal distribution:

$$\lambda_t \sim \mathcal{N}(\mu_{\text{low}}, \sigma^2_{\text{large}}), \tag{6}$$

with a lowered mean μ_{low} and large σ_{large} to increase fine-tuning density at extremely small log-SNR values critical for learning stylistic cues (shaded region in Fig. 2).

The newly sampled log-SNR λ_t is mapped to its corresponding diffusion timestep using $t = (1+\exp(\lambda_t/2))^{-1}$, and this timestep is directly plugged into the standard fine-tuning objective Eq. 3 as shown in the pseudo code of Appendix.

Trainable Parameters. We fine-tune FLUX-dev [2], SD3.5 [7, 48], and SANA [50] by applying LoRA [16] to the attention layers of their underlying transformer architectures. FLUX-dev and SD3.5 employ Multi-Modal Diffusion Transformer (MM-DiT), which contains separate attention modules for text tokens and image tokens; we train LoRA adapters on attention layers of *both* modalities.

4. Experiments

We compare our method for style-driven generation against baselines from several categories, including both fine-tuning and tuning-free methods.

- **Fine-tuning**: SD3 sampler [7], Direct Consistency Optimization (DCO) [24], and StyleDrop [43] are selected as representative fine-tuning methods. The SD3 sampler utilizes a flow matching loss [27, 28] with timestep shifting in Eq. 5, whereas DCO is based on preference learning [34] and employs a regularized loss. For StyleDrop, due to the lack of official open-source implementations, we adopt the unofficial implementation provided by [1].

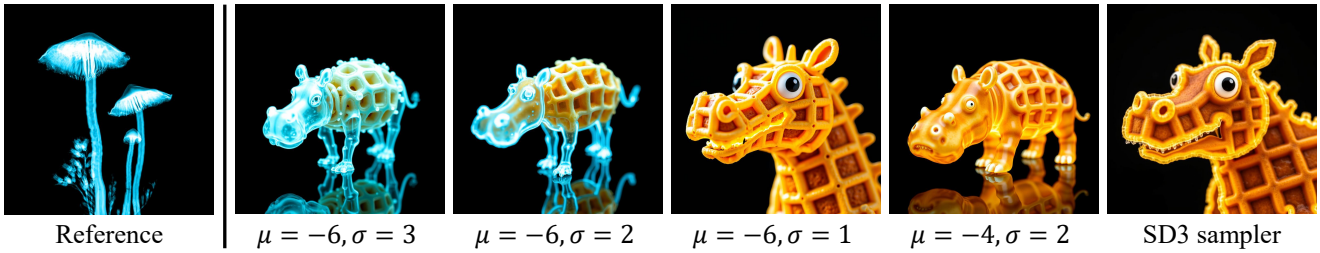
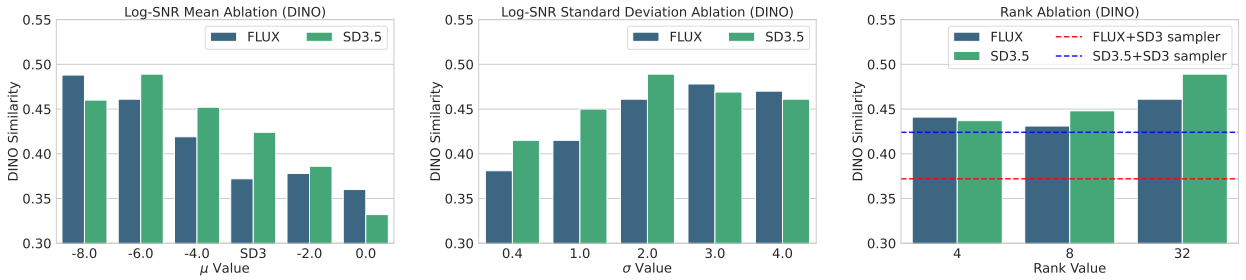


Figure 5. **Effect of varying μ and σ.** Diffusion models start to capture the reference glowing style when μ is lower and σ is larger. The prompt is 'a hybrid creature that is a mix of a waffle and a hippopotamus, in glowing style'. Samples are generated with the same seed.



(a) Varying μ (b) Varying σ (c) Varying LoRA Rank

Figure 6. **SNR sampler analysis.** DINO similarities of varying SNR sampler parameters with FLUX and SD3.5-8B. Dotted lines in (c) indicate results of SD3 sampler [7]. Unless specified, we use $\mu = -6$, $\sigma = 2$, and rank 32. CLIP scores are shown in Appendix.

- **Image variation**: IP-Adapter [52] and RB-Modulation [37] generate image variations using CLIP [33] or CSD [44] image embeddings respectively. IP-Adapter reconstructs images to best reflect the information contained in the CLIP image embedding, whereas RB-Modulation applies gradient guidance at test time to maximize the similarity score evaluated by CSD.
- **Editing**: Style-Aligned [12] with DDIM inversion [45] edits images by manipulating self-attention weights to enforce the style consistency.
- **Detailed prompt**: We name this baseline *GPT-4o prompt*, where we utilize GPT-4o [19] to generate detailed style descriptions from reference images and apply these in text-to-image generation.

Following StyleDrop [43], we select 18 reference styles as our fine-tuning targets. For each style, we use 23 evaluation prompts, generating 2 images per prompt, which results in a total of 828 images per experiment. We fine-tune FLUX-dev [2], SD3.5-8B [7, 48], and SANA [50] with LoRA [16] at rank 32 (except where noted for rank ablation). Following [12, 43] we measure style alignment via human evaluation, DINO ViT-S/16 [3], and CLIP ViT-B/32 [33] image similarity (CLIP-I), and alignment to the text prompts via CLIP text-image similarity (CLIP-T). Implementation details are in Appendix.

4.1. Analysis of Style-Friendly SNR Sampler

In Fig. 5 and Fig. 6, we analyze the impact of varying the parameters of our Style-friendly SNR sampler—specifically, the mean (μ) and the standard deviation (σ) of the log-SNR sampling distribution, as well as the LoRA rank. SD3 sampler denotes the pre-training setting used by FLUX and SD3.5.

Effect of Varying μ. We experiment with μ values ranging from 0 to -8 for both FLUX-dev and SD3.5-8B. As shown in Fig. 5, increasing μ (towards zero) progressively impairs the model's ability to capture reference styles. This trend is quantitatively confirmed in Fig. 6a, where DINO similarity scores decrease with increasing μ, clearly indicating poorer style alignment. Conversely, in Fig. 5, when μ is set to -6, the models begin to capture and reflect the reference styles effectively.

Effect of Varying σ. We also investigate the effect of varying the standard deviation σ of the log-SNR sampling distribution in Fig. 5 and Fig. 6b. When $\sigma < 2$, the fine-tuning does not sufficiently cover the style-emerging region as shown by green curve of Fig. 2, leading to lower style alignment. In Fig. 5, $\sigma \geq 2$ reflects the 'glowing' style successfully. Also, in Fig. 6b, style similarity remains low for

Pre-training configuration uses $\mu = -2\log 3$.

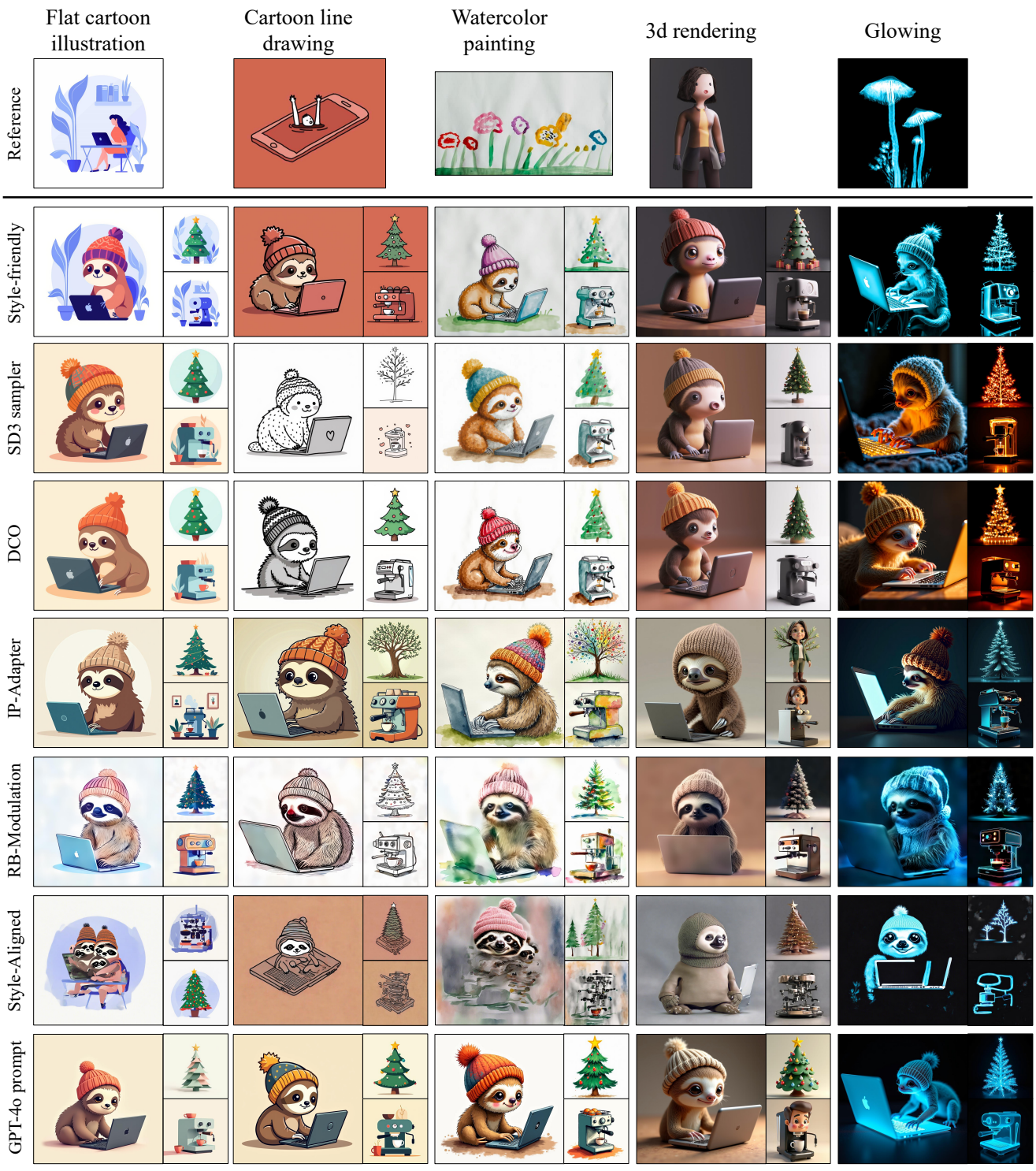


Figure 7. **Qualitative comparison.** We show 'A fluffy baby sloth with a knitted hat trying to figure out a laptop', 'A christmas tree', and 'An espresso machine' in various styles. Reference images and target style prompts are shown in row 1. Our method (row 2) effectively captures the styles indicated by both reference images and their accompanying text prompts. Both fine-tuning methods (row 3 and 4) and tuning-free methods (row 5-8) often miss stylistic nuance.

$\sigma < 2$ but improves when $\sigma \geq 2$.

Effect of Varying Rank. In Fig. 6c, we examine the impact of model capacity by varying the LoRA rank. Notably, with low $\mu = -6$, a rank of 4 achieves higher DINO similarity compared to the SD3 sampler at rank 32 (dotted lines). This demonstrates focusing on lower λ_t has a more pronounced effect on style learning than model capacity alone.

4.2. Qualitative Results

We compare our Style-friendly SNR sampler against recent state-of-the-art methods, including the SD3 sampler [7], Direct Consistency Optimization (DCO) [24], and IP-Adapter [51, 52], which use FLUX-dev as the backbone model; RB-Modulation [37], which uses Stable Cascade [31]; and Style-Aligned [12], which uses SDXL [32].

In Fig. 7, our Style-friendly SNR sampler accurately captures the styles of reference images, reflecting stylistic features including color schemes, layouts, illumination, and brushstrokes. In contrast, fine-tuning FLUX-dev with the standard SD3 sampler often fails to capture key stylistic components, such as layout (column 1), color scheme (columns 2-4), and illumination (column 5).

Fine-tuning FLUX-dev with DCO struggles to learn the reference styles due to strong regularization that prevents significant deviation from the pre-trained model. IP-Adapter with FLUX-dev and RB-Modulation rely on embeddings of CLIP [33] and CSD [44], which may not capture fine stylistic details, leading to inaccurate style reproduction. As seen in column 4 of the IP-Adapter results, a woman in the reference appears, indicating content leakage. Style-Aligned shares self-attention features within the diffusion model, which can cause artifacts such as destroyed structure (columns 1-3) when attention features conflict. While GPT-4o prompt utilizes the detailed style description, it often fails to reflect key stylistic features such as color scheme, highlighting the necessity of image guidance for effective style-driven generation.

4.3. Quantitative Results

We conduct a user study to quantify human preferences using Amazon Mechanical Turk. Following previous work [43], we compare our method to each method with two separate questionnaires: style alignment and text alignment. According to the reference style image and target text prompt, users are asked to select which of the two generated images is more similar to the style in the reference image (style alignment) and represents the text prompt better (text alignment). We obtain 450 answers from 150 participants for each comparison, and the results are presented in Tab. 1. Our method outperforms in *both* aspects ($p < 0.05$ in the Wilcoxon signed-rank test), indicating that it provides superior style alignment without deterioration in content fidelity. More details are provided in Appendix.

Style Alignment				
Method	Model	*win*	*tie*	*lose*
Style-Aligned [12]	SDXL	61.0 %	7.1%	31.9%
RB-Mod [37]	Cascade	55.6 %	12.6%	31.8%
IP-Adapter [52]	FLUX-dev	59.2 %	8.0%	32.8%
DCO [24]	FLUX-dev	56.0 %	10.2%	33.8%
SD3 sampler [7]	FLUX-dev	56.0 %	9.2%	34.8%
Text Alignment				
Method	Model	*win*	*tie*	*lose*
Style-Aligned [12]	SDXL	60.7%	7.5%	31.8%
RB-Mod [37]	Cascade	54.3%	6.3%	39.4%
IP-Adapter [52]	FLUX-dev	56.0%	4.6%	39.4%
DCO [24]	FLUX-dev	53.2%	10.0%	36.8%
SD3 sampler [7]	FLUX-dev	56.5%	14.0%	29.5%

Table 1. **Human evaluation.** User preference comparing style and text alignments between ours and the baselines.

In Tab. 2, we measure DINO [3] and CLIP image (CLIP-I) similarities to assess style alignment, and CLIP text-image (CLIP-T) similarity for target text alignment. For SD3.5 and FLUX, we use the SD3 Sampler as the baseline; for SANA, we use uniform timestep, following the respective pre-training configurations. Our method consistently improves both DINO and CLIP-I scores across all tested backbones and losses, confirming its effectiveness for capturing styles from reference images. Notably, our approach improves DCO, indicating that the key factor in fine-tuning is not the loss function itself, but rather *which SNR range is emphasized* during training.

While our CLIP-T is slightly lower compared to some methods, we already showed superior text alignment in human evaluation (Tab. 1). This discrepancy arises because textual style descriptions alone can be inherently ambiguous. Consequently, methods that accurately reflect unique styles may deviate from these typical appearance, leading to lower CLIP-T scores despite better alignment to the intended reference styles. Overall, our quantitative results confirm that our method accurately reflects both styles and texts.

4.4. Applications

While Dreambooth [38] paper demonstrates generating multi-panel comics by generating *each panel* with a fine-tuned diffusion model, we define the entire multi-panel comic itself as a unique style. Our method treats multiple panels as a *single image* during fine-tuning, enabling the generation of coherent multi-panel comics from only a single reference (see the first row of Fig. 8). By specifying a

Method	Model	Metrics		
		DINO ↑	CLIP-I ↑	CLIP-T ↑
Style-Aligned [12]	SDXL	0.410	0.675	0.340
RB-Mod [37]	Cascade	0.317	0.647	0.363
IP-Adapter [52]	FLUX	0.361	0.656	0.354
GPT-4o Prompt	FLUX	0.299	0.621	0.338
StyleDrop [43]†	MUSE	0.465	0.665	0.325
Uniform [50]	SANA	0.368	0.645	0.359
+Style-friendly	SANA	0.384	0.655	0.359
SD3 sampler [7]	SD3.5	0.424	0.670	0.350
+Style-friendly	SD3.5	0.489	0.698	0.349
DCO [24]	SD3.5	0.399	0.661	0.355
+Style-friendly	SD3.5	0.478	0.695	0.351
SD3 sampler [7]	FLUX	0.373	0.645	0.350
+Style-friendly	FLUX	0.478	0.691	0.343
DCO [24]	FLUX	0.373	0.643	0.353
+Style-friendly	FLUX	0.488	0.698	0.341

Table 2. **Quantitative comparison.** Style alignment (DINO and CLIP-I) and text alignment (CLIP-T) with 18 styles from [43]. Our Style-friendly exhibits superior style-alignment scores. Rows 1-3 show tuning-free baselines. †: Unofficial implementation [1].

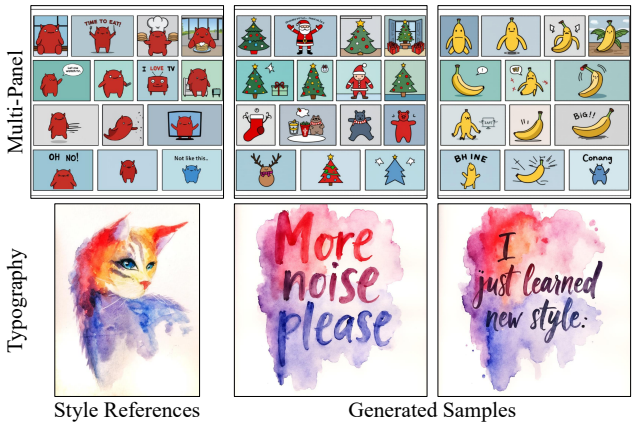


Figure 8. **Multi-panel and typography.** First row demonstrates generating multiple coherent panels as a *single image*. Second row shows customized typography with a unique style.

new subject in the target prompt, the model consistently places that subject across all comic-style panels. Beyond comics, our method also extends to typography, leveraging the spelling capabilities of recent models [2, 48]. As shown in the second row of Fig. 8, this flexibility allows users to effortlessly generate a broad range of customized textual elements in unique styles.

5. Related Works

5.1. Diffusion Models

Diffusion models encapsulate score matching [20, 46, 47] and flow matching [27, 28]. The performance of diffusion models is significantly influenced by the noise level distribution during training. Prior studies have improved model quality by carefully adjusting noise schedules [15, 29] or weights [5, 20, 21], achieving strong results on established benchmarks [10, 13]. However, these prior methods predominantly target improving generic image quality and object-centric fidelity. In contrast, our work provides the empirical validation demonstrating that a style-specific SNR distribution has been overlooked.

5.2. Style-Driven Generation

With advancements in text-to-image models, practitioners have increasingly sought to generate images featuring personal styles [24, 37, 43, 52]. Fine-tuning methods [24, 38, 43] have been particularly prominent in this area. StyleDrop [43], a study closely related to our work, utilizes a masked generative model [4] and involves human data selection through multi-stage training. Some works focus on learning multiple concepts simultaneously [17, 23] or merging several fine-tuned models [24, 41], while others analyze the diffusion model's U-Net [36] layers to identify those most effective for learning styles [8]. However, several existing methods have not yet been validated or applied to recently released large-scale models [2, 7]. In contrast, our method provides an adaptable fine-tuning strategy, which we validate on up-to-date diffusion models.

As an alternative approach for style-driven generation, zero-shot approaches have been proposed [12, 37, 49, 52], but these methods still fall short in style alignment compared to fine-tuning and are often limited to specific domains [11, 39]. Furthermore, some methods [12, 37] require extra inference-time gradient guidance [37] or inversion [12], increasing inference cost. Due to these limitations, we focus on fine-tuning in our work, aiming to provide insights into the behavior of diffusion objectives to make fine-tuning more accessible and effective. While fine-tuning is not the only option, our results show that it is a promising approach.

6. Conclusion

In this paper,we analyzed that stylistic cues in diffusion models primarily arise at lower SNR ranges, which clarifies why conventional schedules in tuning-based style-driven generation often struggle to capture stylistic cues. Motivated by this insight, we proposed the *Style-friendly SNR sampler*, which biases the SNR distribution towards style-emerging SNR regions. Extensive experiments across diverse backbones and fine-tuning methods demonstrate that our approach consistently provides promising improvements in style alignment and fidelity. We hope this work will serve as a stepping stone toward using diffusion models as digital art previewers.

Acknowledgements

This work was partly supported by the Sovereign AI Foundation Model Project (Data Track), organized by the Ministry of Science and ICT (MSIT) and supported by the National Information Society Agency (NIA), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [No. 2022R1A3B1077720; 2022R1A5A7083908], Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2022-II220959; No.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2025.

References

[1] aim uofa. Styledrop-pytorch. https://github.com/aim-uofa/StyleDrop-PyTorch, 2023. 4, 8

[2] BlackForestLabs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 3, 4, 5, 8

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5, 7

[4] Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *International Conference on Machine Learning*, pages 4055–4075. PMLR, 2023. 8

[5] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 2, 8

[6] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 571–576. 2023. 2

[7] 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. 2, 4, 5, 7, 8

[8] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. *arXiv preprint arXiv:2403.14572*, 2024. 8

[9] 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

[10] 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, 2024. 2, 4, 8

[11] Zecheng He, Bo Sun, Felix Juefei-Xu, Haoyu Ma, Ankit Ramchandani, Vincent Cheung, Siddharth Shah, Anmol Kalia, Harihar Subramanyam, Alireza Zareian, et al. Imagine yourself: Tuning-free personalized image generation. *arXiv preprint arXiv:2409.13346*, 2024. 8

[12] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 2, 5, 7, 8

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3, 8

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[15] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023. 2, 8

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 4, 5

[17] Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. Instruct-imagen: Image generation with multi-modal instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4754–4763, 2024. 8

[18] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv: 2307.06350*, 2023. 2, 4

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5

[20] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 2, 8

[21] Diederik Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 8

[22] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 2

[23] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 1931–1941, 2023. 8

[24] Kyungmin Lee, Sangkyung Kwak, Kihyuk Sohn, and Jinwoo Shin. Direct consistency optimization for compositional text-to-image personalization. *arXiv preprint arXiv:2402.12004*, 2024. 2, 4, 7, 8

[25] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in neural information processing systems*, pages 386–396, 2017. 2

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 3

[27] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 4, 8

[28] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 4, 8

[29] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 8

[30] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[31] Pablo Pernias, Dominic Rampas, Mats L. Richter, Christopher J. Pal, and Marc Aubreville. Wuerstchen: An efficient architecture for large-scale text-to-image diffusion models, 2023. 7

[32] 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. In *The Twelfth International Conference on Learning Representations*, 2024. 7

[33] 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. 3, 5, 7

[34] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 8

[37] Litu Rout, Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Rb-modulation: Training-free personalization of diffusion models using stochastic optimal control. *arXiv preprint arXiv:2405.17401*, 2024. 2, 5, 7, 8

[38] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, pages 22500–22510, 2023. 2, 3, 7, 8

[39] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6527–6536, 2024. 8

[40] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. `https://github.com/cloneofsimo/lora`. 2

[41] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, pages 422–438. Springer, 2025. 8

[42] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[43] Kihyuk Sohn, Lu Jiang, Jarred Barber, Kimin Lee, Nataniel Ruiz, Dilip Krishnan, Huiwen Chang, Yuanzhen Li, Irfan Essa, Michael Rubinstein, et al. Styledrop: Text-to-image synthesis of any style. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4, 5, 7, 8

[44] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024. 5, 7

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 5

[46] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 8

[47] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2, 8

[48] Stability. stable-diffusion-3.5-large. https://huggingface.co/stabilityai/stable-diffusion-3.5-large, 2024. 2, 4, 5, 8

[49] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 8

[50] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *The Thirteenth International Conference on Learning Representations*, 2025. 2, 4, 5, 8

[51] XLabs-AI. flux-ip-adapter. https://huggingface.co/XLabs-AI/flux-ip-adapter, 2024. 2, 7

[52] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 5, 7, 8