StyleSSP: Sampling StartPoint Enhancement for Training-free Diffusion-based Method for Style Transfer

Supplementary Material

1. Appendix

1.1. Startpoint Impact Analysis

Given that StyleSSP is specifically designed to enhance the sampling startpoint, we place primary emphasis on the importance of the startpoint in style transfer. We demonstrate how minor modifications to the startpoint can significantly influence style transfer results. As shown in Fig. 1, we present several style transfer results. The titles in the figure — "wi Inversion," "wo Inversion," "Noised Latent," "Shifted Latent," and "Scaled Latent" — correspond to the startpoints z_T, z_r, z_T^n, z_T^{sh}, and z_T^{sa}, respectively. Their formulations are as follows:

$$\begin{aligned} z_r &\sim \mathcal{N}(0, \mathbf{I}) \\ z_T^n &= z_T + \mathcal{N}(0, \mathbf{I}) \\ z_T^{sh} &= z_T + \mathbf{U}(-0.5, 0.5) \\ z_T^{sa} &= z_T \times \mathbf{U}(0.5, 1) \end{aligned} \tag{1}$$

where z_T is the DDIM latent of the content image, $\mathcal{N}$ represents a Gaussian distribution, and $U(-0.5, 0.5)$ and $U(0.5, 1)$ indicate uniformly random values selected within the ranges -0.5 to 0.5 and 0.5 to 1.0, respectively.

As illustrated in Fig. 1, manipulations of the sampling startpoint make a significant impact on the results of style transfer, resulting in notable changes in both the image hue and the content representation. Note that the following results are all conducted with ControlNet as an additional content controller. Several key observations can be made from this figure.

First, referring to the 3rd and 4th columns in this figure, using the DDIM latent z_T extracted from the content image as the sampling startpoint results in remarkably better content preservation compared to using random Gaussian noise as the startpoint. This finding motivates us to adopt DDIM inversion as the first step in our method, as is done in many inversion-based methods [1, 7, 9].

Second, we attempted minor modifications to the DDIM latent z_T. Referring to the 3rd, 5th, and 6th columns in this figure, we observe that these simple manipulations produce significant changes in image tone, and since color variation is a crucial aspect of style transfer, this finding further drives our focus on startpoint enhancement.

Third, by examining the results in the 3^{rd} and 5^{th} rows, we notice that the startpoint not only affects the tone of generated images but can also influence the content of generated images to some extent, such as the facial outline of the woman in the 3^{rd} row and the background in the 5^{th} row. This effect has been largely overlooked in previous works, yet it is undeniably critical for style transfer tasks.

In summary, through simple adjustments to the startpoint, we have discovered its substantial impact on style transfer results — affecting content preservation, content modification, and tonal changes. These insights have driven us to pursue sampling startpoint enhancement for style transfer research. Therefore, our method, StyleSSP, emphasizes guidance during the inversion step and manipulation of the inversion latent space to achieve a more effective sampling startpoint in style transfer issues.

1.2. Principle of Negative Guidance

In this section, we provide a detailed introduction to the principles of negative prompt guidance, starting with conditional generation. For conditional generation, that is, to sample samples from the conditional distribution $p(x|y)$. According to the Bayes formula, we can obtain:

$$\begin{aligned} p(x|y) &= \frac{p(y|x)p(x)}{p(y)}, \\ log\, p(x|y) &= log\, p(y|x) + log\, p(x) - log\, p(y), \\ \Rightarrow \nabla_x log\, p(x|y) &= \nabla_x log\, p(y|x) + \nabla_x log\, p(x). \end{aligned} \tag{2}$$

In the classifier-guided task, the score-based model with unconditional input is an estimation of $\nabla_x \log p(x)$, so in order to obtain $\nabla_x \log p(x|y)$, an additional classifier needs to be trained to estimate $\nabla_x \log p(y|x)$. At the same time, to control the strength of condition, the guidance scale ω is introduced:

$$\nabla_x log\, p(x|y) := \omega \nabla_x log p(y|x) + \nabla_x log\, p(x). \tag{3}$$

In classify-free guidance (CFG) tasks, they simultaneously train two score-based models, $\nabla_x log\ p(x)$ and $\nabla_x log\ p(y|x)$. Since $\nabla_x log\ p(y|x) = \nabla_x log\ p(x|y) - \nabla_x log\, p(x)$, it follows that:

$$\nabla_x log\, p(x|y) := \omega(\nabla_x log\, p(x|y) - \nabla_x log\, p(x)) + \nabla_x log\, p(x), \tag{4}$$

When negative prompt serves as a condition, the conditions for diffusion model contain two items, one is positive prompt condition y, and the other is negative prompt condition not $\tilde{y}$. Since re-training a score-based model to esti-

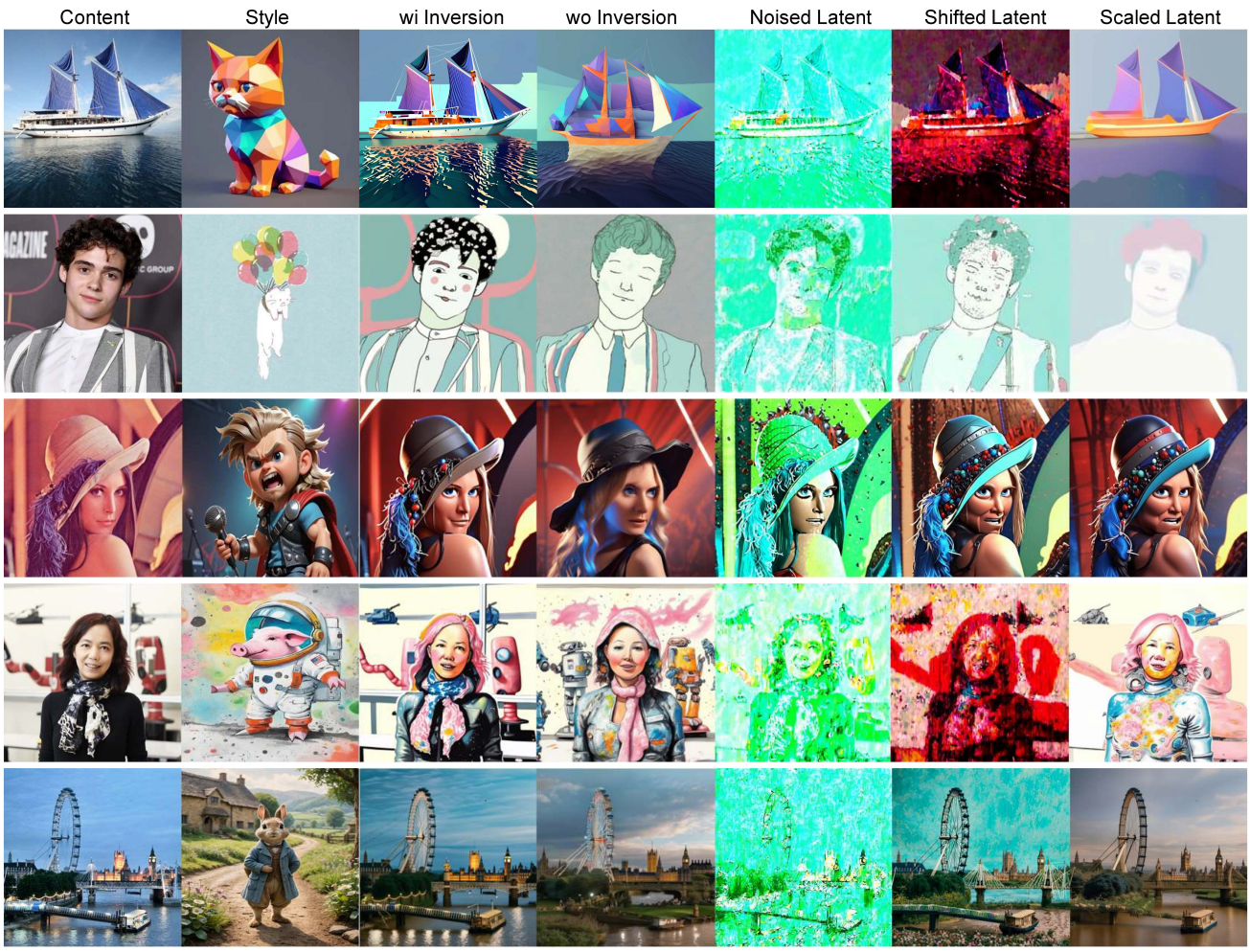

Figure 1. Illustrations of style transfer results based on various startpoints. As shown in this figure, startpoint manipulations yield significant changes in both image hue and content representation, underscoring the crucial role of the sampling startpoint in style transfer. All results are generated with ControlNet as an additional content controller.

mate $\nabla_x p(x|y, \text{not } \tilde{y})$ is costly, the following simplification is made:

$$\begin{aligned} p(x|y, \text{not } \tilde{y}) &= \frac{p(x, y, \text{not } \tilde{y})}{p(y, \text{not } \tilde{y})} \\ &= \frac{p(y|x)p(\text{not } \tilde{y}|x)p(x)}{p(y, \text{not } \tilde{y})} \\ &\propto \frac{p(x)}{p(y, \text{not } \tilde{y})} \frac{p(y|x)}{p(\tilde{y}|x)}, \end{aligned} \tag{5}$$

so that:

$$\begin{aligned} \nabla_x log\, p(x|y, \text{not } \tilde{y}) \propto \nabla_x log\, p(x) \\ + \nabla_x log\, p(y|x) - \nabla_x log p(\tilde{y}|x). \end{aligned} \tag{6}$$

The Eq. 5 and Eq. 6 assume that x, y and not $\tilde{y}$ are mutually independent. Letting ω^+ be the guidance scale of positive condition and ω^- be the guidance scale of negative condition, we have:

$$\begin{aligned} \nabla_x p(x|y, \text{not } \tilde{y}) := \nabla_x p(x) + \omega^+(\nabla_x p(x|y) - \nabla_x p(x)) \\ - \omega^-(\nabla_x p(x|\tilde{y}) - \nabla_x p(x)). \end{aligned} \tag{7}$$

Thus, we can estimate $\nabla_x p(x|y, \text{not } \tilde{y})$ only by calculating $\nabla_x p(x), \nabla_x p(x|y), \nabla_x p(x|\tilde{y})$, and all of these can be obtained through the pre-trained diffusion model.

It should be noted that in the negative guidance method proposed in this paper, IP-Instruct merely exists as a style and content extractor, which can be replaced by any other extractor. Meanwhile, this CFG-based guidance can also be replaced by the gradient-based guidance like FreeTune [8] does. We emphasize that our prominent contribution lies in discovering that by guiding the startpoint of sampling stage

Figure 2. Qualitative comparison with with baselines(StyleID, InstantStyle plus). Zoom in for viewing details.

to distance from the style image's content, thereby preventing the content leakage from style image.

1.3. User Study

We conduct added a user study based on the setting of Deadiff [5]. We employed StyleID [1], StyleAlign [4], InstantStyle plus [7], InstantStyle [6], DiffuseIT [3], DiffStyle [2]. Additionally, InST [9] and StyleSSP to separately generate 4 stylized images. As shown in Fig. 3, 24 users from diverse backgrounds evaluate there images in terms of best content preservation (BCP), least style leakage (LSL), and overall performance (Overall). StyleSSP outperforms all state-of-the-art methods on three evaluation aspects with a big margin, which demonstractes the broad application prospects of our method.

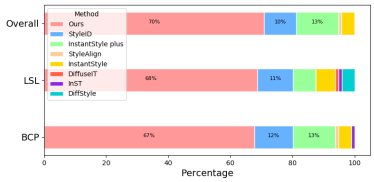

Figure 3. Results for the user study in percentages.

1.4. Parameter Selection

We conducted additional experiments to show how the frequency pass parameter σ affects the results. As shown in Fig. 4, σ performs best in the range of 0.3 to 0.5, performing the best background and facial line preservation. This is because a very small σ fails to emphasize high-frequency

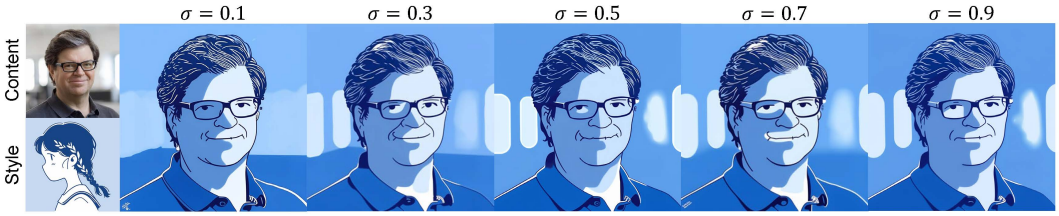

Figure 4. Visualization of frequency pass parameter σ's effect.

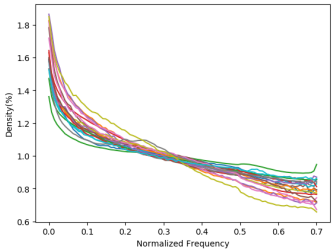

Figure 5. Frequency spectrum distribution of 20 random images.

information, while a very large σ suppresses too many valid components of images. Moreover, Fig. 5 shows that, although images differ in the spatial domain, their frequency distributions are quite similar, which supports us to use nearly the same σ for different images. Moreover, since the frequency distribution of images is similar, the frequency band related to contours is not significantly affected by the choice of diffusion model. Therefore, different diffusion models should share the same frequency pass parameter σ. In summary, we recommend choosing σ between 0.3 and 0.5, and this choice of value is not significantly related to the diffusion model.

1.5. Additional Results

We additionally compare the proposed method with the most recent baseline (StyleID) and the baseline with lowest ArtFID (InstantStyle plus). Fig. 2 shows the additionally qualitative comparison of ours with diffusion model baselines.

Also, in Fig. 6, we visualize the style transfer results of various pairs of content and style images, which further demonstrate StyleSSP's robustness and versatility in adapting to diverse content and style.

Figure 6. Style transfer results of style and content image pairs. Zoom in for viewing details.

References

[1] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8795–8805, 2024. 1, 3

[2] Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. Training-free content injection using h-space in diffusion models, 2024. 3

[3] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation, 2023. 3

[4] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Stylediffusion: Prompt-embedding inversion for text-based editing. *arXiv preprint arXiv:2303.15649*, 2023. 3

[5] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. *arXiv preprint arXiv:2403.06951*, 2024. 3

[6] Haofan Wang, 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. 3

[7] Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788*, 2024. 1, 3

[8] Youcan Xu, Zhen Wang, Jun Xiao, Wei Liu, and Long Chen. Freetuner: Any subject in any style with training-free diffusion, 2024. 2

[9] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10146–10156, 2023. 1, 3