Training-free Fourier Phase Diffusion for Style Transfer

Siyuan Zhang[1] , **Wei Ma**[1*] , **Libin Liu**[1] , **Zheng Li**[1] and **Hongbin Zha**[2]

[1]College of Computer Science, Beijing University of Technology
[2]Key Laboratory of Machine Perception (MOE), School of IST, Peking University
s202374157@emails.bjut.edu.cn, mawei@bjut.edu.cn, {liulibin,lzlove}@emails.bjut.edu.cn, zha@cis.pku.edu.cn

Abstract

Diffusion models have shown significant potential for image style transfer tasks. However, achieving effective stylization while preserving content in a training-free setting remains a challenging issue due to the tightly coupled representation space and inherent randomness of the models. In this paper, we propose a Fourier phase diffusion model that addresses this challenge. Given that the Fourier phase spectrum encodes an image's edge structures, we propose modulating the intermediate diffusion samples with the Fourier phase of a content image to conditionally guide the diffusion process. This ensures content retention while fully utilizing the diffusion model's style generation capabilities. To implement this, we introduce a content phase spectrum incorporation method that aligns with the characteristics of the diffusion process, preventing interference with generative stylization. To further enhance content preservation, we integrate homomorphic semantic features extracted from the content image at each diffusion stage. Extensive experimental results demonstrate that our method outperforms state-of-the-art models in both content preservation and stylization. Code is available at https://github.com/zhang2002forwin/Fourier-Phase-Diffusion-for-Style-Transfer.

1 Introduction

Image style transfer, which aims to incorporate a specific style into an input image while preserving its core content, finds wide applications in fields such as digital art creation and augmented reality filters. Early deep learning-based style transfer methods, such as AdaIN [Huang and Belongie, 2017], Gram-based approaches [Gatys *et al.*, 2016], and WCT [Li *et al.*, 2017], achieved this by aligning the statistical properties of content and style features extracted from pre-trained VGG networks [Simonyan and Zisserman, 2014]. Subsequently, GAN-based methods [Zhu *et al.*, 2017] were introduced, employing adversarial training to transform content into a styled domain via CNNs. However, these CNN-based methods simply optimize the features extracted by a basic network to obtain stylized results, often struggling to achieve a harmonious integration of style and content.

Figure 1: Qualitative comparison with state-of-the-art diffusion-based style transfer methods, including FreeStyle [He *et al.*, 2024] and StyleID [Chung *et al.*, 2024]. All three methods are training-free, with FreeStyle and ours taking a textual style prompt as input, while StyleID uses an example image of the style as input (more details are given in Section 5.2). The results show that FreeStyle struggles with content preservation, while StyleID produces unsatisfactory stylization effects. In contrast, our model effectively stylizes the images while preserving their content.

Recently, diffusion models have achieved significant breakthroughs in tasks such as text-to-image generation, image editing, and super-resolution. Many studies [Zhang *et al.*, 2023c] [Wang *et al.*, 2023] [Li *et al.*, 2025] [Hamazaspyan and Navasardyan, 2023] have explored the generative capabilities of these models in style transfer, particularly focusing on decoupling and re-fusing style and content. Some methods [Chen *et al.*, 2024] [Qi *et al.*, 2024] require both a content prompt and a style image as input. These methods train additional adapters to extract style information from the style image and then integrate the extracted style into the generated content using the cross-attention mechanism or LoRA layers [Hu *et al.*, 2021]. However, these techniques demand substantial computational resources during training, and many rely on content descriptions, which, even when complex, often lead to results that deviate from the user's expectations,

* Corresponding author.

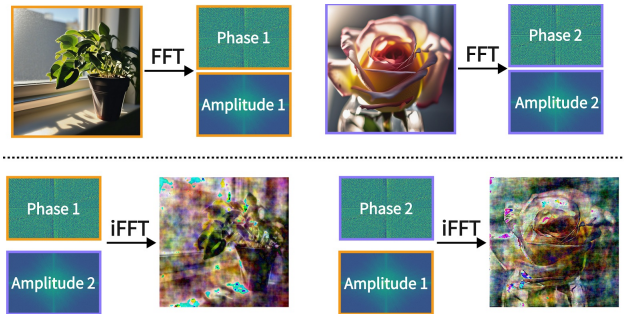

Figure 2: Visualizing what Fourier phase reveals. By combining the amplitude spectrum from one image with the phase spectrum from another image, followed by inverse Fourier transform, we obtain a new image that inherits structural information from the phase spectrum while preserving color characteristics from the amplitude spectrum. However, direct combination introduces noticeable artifacts.

making them unsuitable for style transfer on photographs.

Training-free methods leverage the inherent capabilities of pre-trained large diffusion models for style transfer, requiring much fewer computational resources and offering easy extensibility to various versions of diffusion models. However, such methods face challenges due to the tightly coupled representation space and the inherent randomness of the diffusion models. StyleID [Chung *et al.*, 2024] replaces the key and value of the content with those from the style image, obtained via DDIM inversion, to inject the style using the cross-attention mechanism during the generation process. This integration, which relies on the similarity between the style and content features, often leads to unsatisfactory output, as shown in Figure 1. In contrast, FreeStyle [He *et al.*, 2024] combines the low-frequency components of the content features with the high-frequency components of the style features for stylization. Although FreeStyle produces more intense style effects than StyleID, it struggles with content preservation, particularly in fine-grained structures. For example, the face of the old man is transformed into that of a child after stylization, as shown in Figure 1.

In this paper, we introduce a novel, training-free Fourier phase diffusion model for style transfer, where the phase spectrum of the content image is used to conditionally guide generation. This approach outperforms existing models in both content preservation and style transfer, as shown in Figure 1. Our motivation for utilizing the Fourier phase spectrum lies in its ability to capture the structures of an image, including object shapes and inner edges [Lv *et al.*, 2024] [Yao *et al.*, 2024] [Yang and Soatto, 2020]. As illustrated in our simple test in Figure 2, replacing the phase spectrum of an image results in a complete alteration of the content. This demonstrates the potential of the phase spectrum to decouple and conditionally guide the highly randomized generative process, thereby maintaining the content while enabling stylistic generation.

However, integrating the Fourier phase spectrum of the content image into the diffusion model without training is challenging, as it may interfere with the generation of natural-looking stylized images. As shown in Figure 2, combining the phase and amplitude from different images can result in noticeable artifacts. To address this challenge, we propose a phase spectrum fusion module to modulate intermediate stylization results with the content phase, along with a strategy to integrate the module into the generation process, aligning with the stage-specific generative characteristics of the diffusion model. Furthermore, to enhance content preservation, inspired by ControlNet [Zhang *et al.*, 2023a], we utilize the U-Net architecture of the diffusion model to encode the content image, providing semantic guidance throughout the entire generation process.

Our main contributions are summarized as follows:

- We propose a novel training-free Fourier phase diffusion model for style transfer, which leverages the phase spectrum of a content image to guide the style generation process, achieving outstanding performance in both content preservation and style transfer quality.
- We present a phase spectrum integration method that incorporates the phase spectrum of the content image into the generation process, preserving structural details while generating visually appealing stylized images.
- We propose a semantic injection mechanism that extracts the content image's semantics using the U-Net within the diffusion model and subsequently incorporates these semantics into the generation process, further enhancing content preservation.
- Extensive qualitative and quantitative evaluations demonstrate that the proposed model outperforms state-of-the-art methods in style transfer tasks.

2 Related Work

2.1 CNN-based Style Transfer

Gatys et al. [Gatys *et al.*, 2016] pioneered CNN-based neural style transfer by using VGG features and Gram matrices to capture style characteristics, and generating stylized results by optimizing both content and style objectives. Subsequent methods like AdaIN [Huang and Belongie, 2017] and WCT [Li *et al.*, 2017] leverage the VGG network [Simonyan and Zisserman, 2014] for feature extraction and align feature statistics for stylization. CycleGAN [Zhu *et al.*, 2017] approaches style transfer through adversarial learning to establish cross-domain mappings. Recently, CLIPStyler [Kwon and Ye, 2022] combines a CNN encoder-decoder architecture with patch-wise CLIP loss to capture hierarchical content features and refine stylization results. Despite their innovations, these methods often struggle to achieve a satisfactory integration of content from a content image and style from a reference image, due to the inherent differences between the two.

2.2 Diffusion-based Style Transfer

Diffusion models have surpassed GANs in image generation and have been successfully applied to various tasks, including style transfer. Leveraging CLIP [Radford *et al.*, 2021], the Latent Diffusion Model (LDM) [Rombach *et al.*, 2022] improves sampling efficiency by encoding images into a latent space using a pretrained auto-encoder [Esser *et al.*, 2021], enabling text-guided image generation through large-scale

pretraining. Recent works like DEADiff [Qi *et al.*, 2024] and ArtAdapter [Chen *et al.*, 2024] achieve style transfer by integrating style information into diffusion models through trained style embeddings or adapters. These methods rely on text prompts for content specification, limiting their applicability to photographs. While works like Artbank [Zhang *et al.*, 2024b] take content images and style descriptions as input, they still require substantial computational resources for training. Training-free image stylization methods have attracted much attention recently [Jiang and Chen, 2024] [He *et al.*, 2024], but maintaining content fidelity while achieving the desired stylization remains an open challenge.

2.3 Controllable Text-to-Image Diffusion

Diffusion models need to focus on controlling the content of generated images, which is also a crucial aspect of diffusion-based style transfer. Several approaches have been proposed to control the content of output generated by diffusion models [Zhang *et al.*, 2023a] [Ye *et al.*, 2023] [Zheng *et al.*, 2023] [Yu *et al.*, 2023]. For instance, ControlNet [Zhang *et al.*, 2023a] allows users to control the structures of the generated images by conditioning them on various input types, such as depth maps, Canny edges, or human poses. FreeControl [Mo *et al.*, 2024], a training-free method for controllable text-to-image generation, enables control over the generated images based on their semantic representations. Layout-diffusion [Zheng *et al.*, 2023] adopts a layout fusion module to control the generated images. Although significant progress has been made in controlling the output images of diffusion models, these methods are not directly applicable to image stylization. However, their conditional guidance based on various clues inspires us to address content preservation during stylization.

3 Preliminaries of Diffusion Models

The diffusion model consists of two key processes: the diffusion process and the sampling process. In the diffusion process, noise is added to a given image x_0 to obtain a noisy image x_t at time step t according to the equation:

$$x_t = \sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\epsilon \tag{1}$$

where ϵ represents the Gaussian noise, and $\overline{\alpha_t} = \prod_{i=0}^{t} \alpha_i$. α_i is a hyperparameter that controls the signal strength at time step i.

The generation process of diffusion models typically starts with the noisy image, and through iterative denoising steps, the model gradually refines the content, eventually achieving fine textures and details. In detail, given a text prompt y, diffusion models use noise-predictor ϵ_θ (U-Net in our method) to predict noise $\epsilon_\theta(x_t, t, C)$ from x_t, where θ denotes parameters of the noise-predictor [Rombach *et al.*, 2022] and C represents the text embedding of the prompt y. The down-block and mid-block of ϵ_θ encode semantic information. In DDIM [Song *et al.*, 2020], $x_{0,t}$ at timestep t is given by

$$x_{0,t} = \frac{x_t - \sqrt{1-\overline{\alpha_t}}\epsilon_\theta(x_t, t, C)}{\sqrt{\overline{\alpha_t}}} \tag{2}$$

As t approaches zero, $x_{0,t}$ becomes increasingly similar to the generated result. $x_{0,t}$ is used to denoise x_t to obtain x_{t-1}. Mathematically,

$$x_{t-1} = \sqrt{\overline{\alpha_{t-1}}}x_{0,t} + \sqrt{1-\overline{\alpha_{t-1}}-\sigma_t^2}\epsilon_\theta(x_t, t, C) + \sigma_t^2\epsilon \tag{3}$$

where σ_t controls the stochasticity of the sampling process. DDIM sets σ_t to 0 in Eq. (3), making the sampling process deterministic.

4 Method

4.1 Overview

The overall framework is shown in Figure 3. Given a content image I_c and a style text description y_s, we aim to transfer the style from y_s to I_c, resulting in a stylized image I_{cs} that maintains the content of I_c. Firstly, we add noise to the input image I_c using Eq. (1), setting $t = T$, to obtain the noisy image x_T, $T = 1000$. x_T subsequently serves as the starting point for the sampling process to generate the stylized image I_{cs}. The style text description y_s is encoded by the CLIP text encoder [Radford *et al.*, 2021] into style embeddings, which influence the style of the generated result through cross-attention. We design a phase spectrum fusion module (see Section 4.2) and explore its integration into a specific stage of the diffusion model, based on the stage-specific generation characteristics (see Section 4.3). Additionally, to further enhance content preservation, we use the U-Net of the diffusion model to extract semantics from the input content image and inject them into the generation process (see Section 4.4).

4.2 Phase Spectrum Fusion

Our phase fusion module is illustrated in Figure 4. At timestep t, we obtain the intermediate stylized result $x_{0,t}$ using Eq. (2). Then we transform both $x_{0,t}$ and the content image I_c into Fourier space, yielding I_x^{freq} and I_c^{freq}. The transformation is given by

$$I^{freq}(u,v) = \sum_{m=0}^{M-1}\sum_{m=0}^{N-1} I(m,n)e^{-j2\pi(um/M+vn/N)} \tag{4}$$

where M and N are the length and width of the images, and (m, n) and (u, v) represent the coordinates in the RGB image space and its frequency space, respectively.

$$I^{freq}(u,v) = R(u,v) + jQ(u,v) \tag{5}$$

where $R(u, v)$ and $Q(u, v)$ are the real components and imaginary components, and j is the imaginary unit. We can obtain the amplitude spectrum of $x_{0,t}$ and I_c, denoted as $\left|I_x^{freq}\right|$ and $\left|I_c^{freq}\right|$, respectively, by using

$$\left|I^{freq}(u,v)\right| = \sqrt{R^2(u,v) + Q^2(u,v)} \tag{6}$$

Then we add them linearly and empirically set the weight to 0.5, to obtain $\left|I_{sum}^{freq}\right|$. On the other hand, the phase spectrum ϕ_c^{freq} of I_c is computed as

$$\phi_c^{freq}(u,v) = \arctan\left[\frac{Q_c(u,v)}{R_c(u,v)}\right] \tag{7}$$

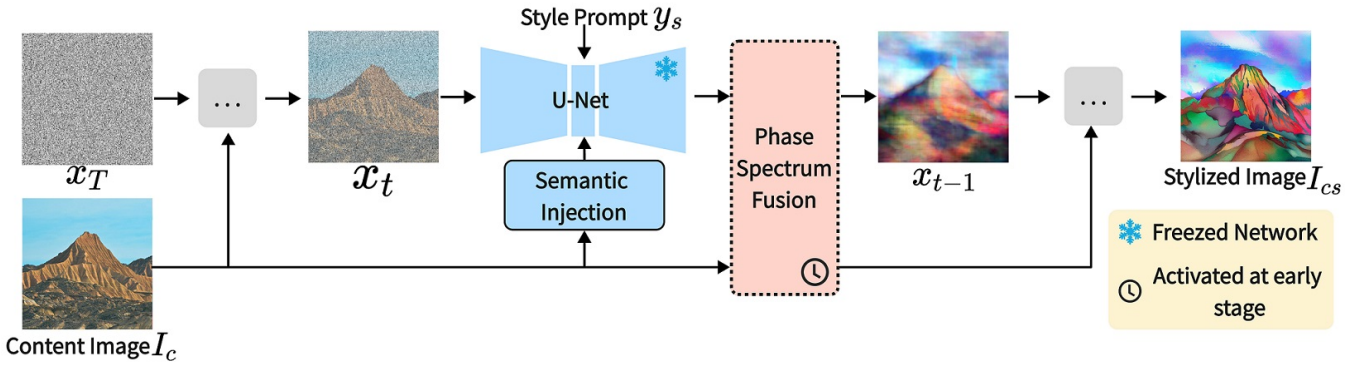

Figure 3: Overall pipeline of the proposed model. We use the phase spectrum fusion module (Section 4.2) to enhance the content structures in the stylized image. The phase fusion module is applied stepwise at a specifically selected stage of the generation process (Section 4.3). The semantics of the content image are injected, providing content information throughout the generation process (Section 4.4).

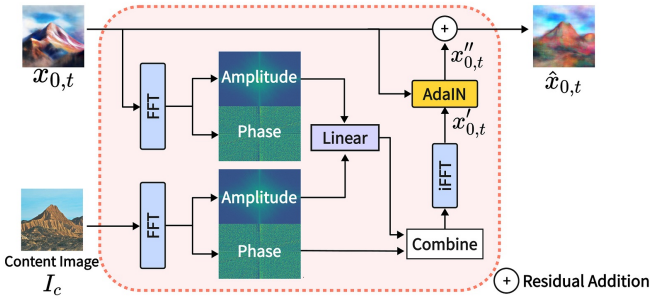

Figure 4: Our phase fusion module. The phase fusion module introduces the phase spectrum of the content image into $x_{0,t}$, influencing $x_{0,t}$ in a residual manner.

Then we recombine the phase spectrum ϕ_c^{freq} and the amplitude spectrum $\left|I_{sum}^{freq}\right|$ to obtain the new frequency data I_{comb}^{freq}, as follows

$$I_{comb}^{freq} = \left|I_{sum}^{freq}\right| e^{j\phi_c^{freq}} \tag{8}$$

Next, we perform inverse Fourier transform to obtain $x'_{0,t}$, given by:

$$x'_{0,t}(m,n) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} I_{comb}^{freq}(u,v) e^{j2\pi(um/M+vn/N)} \tag{9}$$

Note that directly mixing the phase and amplitude spectra of different images leads to artifacts, as explained in Figure 2. AdaIN [Huang and Belongie, 2017] is known for effectively aligning feature statistics in style transfer, while Style Spectroscope [Jin *et al.*, 2024] has shown that AdaIN can only affect the amplitude spectrum, not the phase spectrum. Therefore, we apply AdaIN to align the statistics of $x'_{0,t}$ with those of $x_{0,t}$, aiming to mitigate the color artifacts caused by the amplitude and phase mismatch in Eq. 9.

$$x''_{0,t} \leftarrow \mathrm{AdaIN}(x'_{0,t}, x_{0,t}) \tag{10}$$

Finally, we introduce $x''_{0,t}$ into $x_{0,t}$ in the form of residuals, namely

$$\hat{x}_{0,t} = \alpha x''_{0,t} + \beta x_{0,t} \tag{11}$$

where α and β are hyperparameters. α controls the amount of phase from the content image I_c, while β controls how much of the intermediate stylized result $x_{0,t}$ is retained. $\hat{x}_{0,t}$ then replaces $x_{0,t}$ in the subsequent denoising process.

4.3 Phase Spectrum Fusion in Diffusion

Prospect [Zhang *et al.*, 2023b], FreeU [Si *et al.*, 2024], and other works [Guo and Lin, 2024][Zhang *et al.*, 2024a] have demonstrated that the diffusion model generates images through a *layout - content - fine texture* process. Accordingly, we divide the generation process into three stages: early, middle, and late. We incorporate the content image via the phase spectrum fusion module in the early stage, where the edge structures exert the greatest influence on the final result's structure. Additionally, incorporating the content phase causes the intermediate results of the diffusion process to deviate from the learned data distribution. Several diffusion steps are needed to bring these results back into the expected distribution of the diffusion model. Therefore, we apply the phase spectrum fusion module intermittently during the early stage.

To support our design, we conduct contrastive experiments. The results are shown in Figure 5. Groups A, B, and C apply the phase fusion module continuously during the early, middle, and late stages, respectively, while groups D, E, and F use the module intermittently across these stages. The results of groups D, E, and F are better than those of groups A, B, and C, indicating that continuous application of the phase fusion module is suboptimal. In contrast, intermittent use of the module helps maintain content structures while achieving higher-quality stylization. Among these, group D, which applies the phase fusion module intermittently in the early stage, produces the most satisfying results, as expected.

4.4 Semantic Injection

Although phase spectrum fusion and its incorporation are effective, subsequent generation steps may introduce elements unrelated to the semantic content of the input image. To address this, inspired by ControlNet [Zhang *et al.*, 2023a], Zecon [Yang *et al.*, 2023], and other works [Choi *et al.*, 2025][Lin *et al.*, 2024][Li *et al.*, 2024] in the virtual try-on

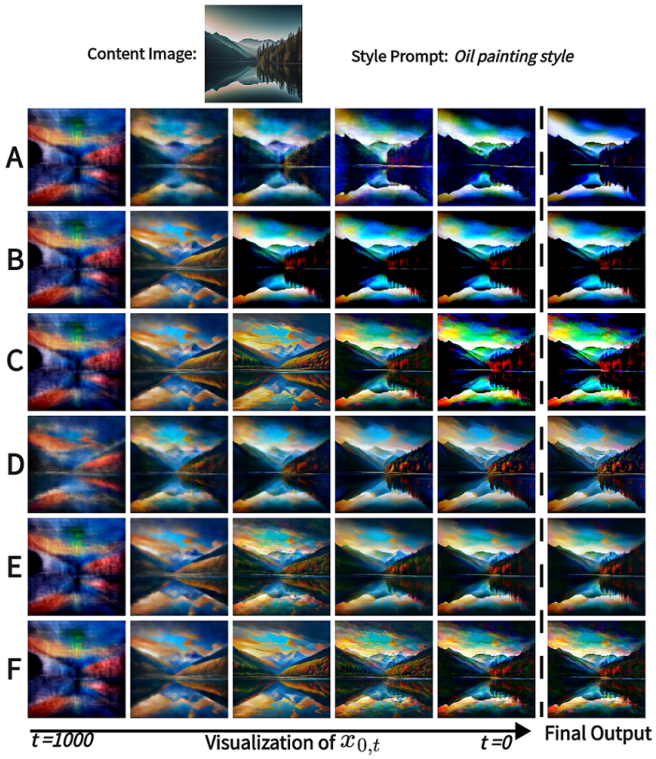

Figure 5: Contrastive experiments on integrating phase spectrum fusion in the diffusion model. We conducted six sets of experiments: A, B, C, D, E, and F. The first five columns show the visualizations of $x_{0,t}$, while the last column displays the generated images. We divide the generation process into three stages: early, middle, and late. Groups A, B, and C use the phase fusion module continuously within the early, middle, and late stages, respectively, while groups D, E, and F use it intermittently within the corresponding stages.

task, which demonstrate that the U-Net in diffusion models can extract semantic information, we extract and incorporate the semantics of the content image as guidance during generation. We utilize the down-blocks and mid-blocks of the U-Net in diffusion model to extract semantic features from the content image, which are then injected into the corresponding positions of the U-Net in the generative stylization model (as illustrated in Figure 3) in a residual manner.

5 Experiments

5.1 Implementation Details

We utilize Stable Diffusion XL 1.0 as the foundational text-to-image model and employ the DDIM sampler with 30 sampling steps for each stylized image generation. The phase fusion module is applied during the early sampling stage, specifically at the 2nd, 5th, and 8th timesteps, with the parameters set to $\alpha = 0.5$ and $\beta = 0.7$. Notably, our method is training-free and does not require fine-tuning. All experiments are conducted on a single RTX 3090 GPU.

5.2 Qualitative Comparison

We compare our method with state-of-the-art approaches, including Zecon [Yang *et al.*, 2023], CLIPStyler [Kwon and Ye, 2022], FreeStyle [He *et al.*, 2024], and StyleID [Chung *et al.*, 2024]. Unlike the other methods, which use a style text prompt and a content image as input, StyleID requires both a content image and a style image. To generate the necessary style image for StyleID, we employed the Stable Diffusion XL 1.0 model based on the provided style description. Figure 6 illustrates the comparison results. FreeStyle demonstrates excellent style transfer but struggles to preserve content. Conversely, CLIPStyler maintains content structures but delivers weaker style transfer. Zecon also exhibits limited style quality, likely due to the CLIP model's emphasis on semantic content rather than style. The Artist method produces results nearly identical to the input content images, while StyleID shows relatively weak style transfer and suboptimal structural preservation. In contrast, our method achieves superior content preservation while delivering style quality comparable to FreeStyle.

5.3 User Study

We conducted a user survey with 50 participants to compare our method against five other approaches: FreeStyle, Artist, Zecon, CLIPStyler, and StyleID. For these methods, all experimental results were produced using their publicly available default parameter settings. Participants were presented with 25 groups of results, each paired with the corresponding content image and style text description. For unconventional styles, a style image generated by the Stable Diffusion model was also included as a reference. Participants assessed both the content preservation and style quality of the generated results, selecting at least two results they favored the most in each group. The results of the user survey, presented in Figure 7, indicate that our method was more favored.

5.4 Quantitative Comparison

Evaluation Metrics

CLIP Aesthetic Score. It is a metric for assessing aesthetic scores. The underlying model is a neural network that uses CLIP embeddings as input and is trained to predict an image's aesthetic score based on its visual appeal.

HPS. The Human Preference Score (HPS) [Wu *et al.*, 2023] is a scoring model developed to predict human preferences for generated images. Trained on a Human Preference Dataset, this model assigns higher scores to images that are more likely to be favored by people.

CLIP Score. We calculate the cosine similarity between the CLIP features of a stylized image and the corresponding style text description. A higher CLIP Score reflects a more accurate alignment with the intended style.

SSIM. Structural Similarity Index (SSIM) is a metric used to measure the similarity between two images. It evaluates image similarity by analyzing three key dimensions: luminance, contrast, and structure within local regions of the images. An SSIM value close to 1 indicates high similarity between the images, while a value near 0 suggests significant differences.

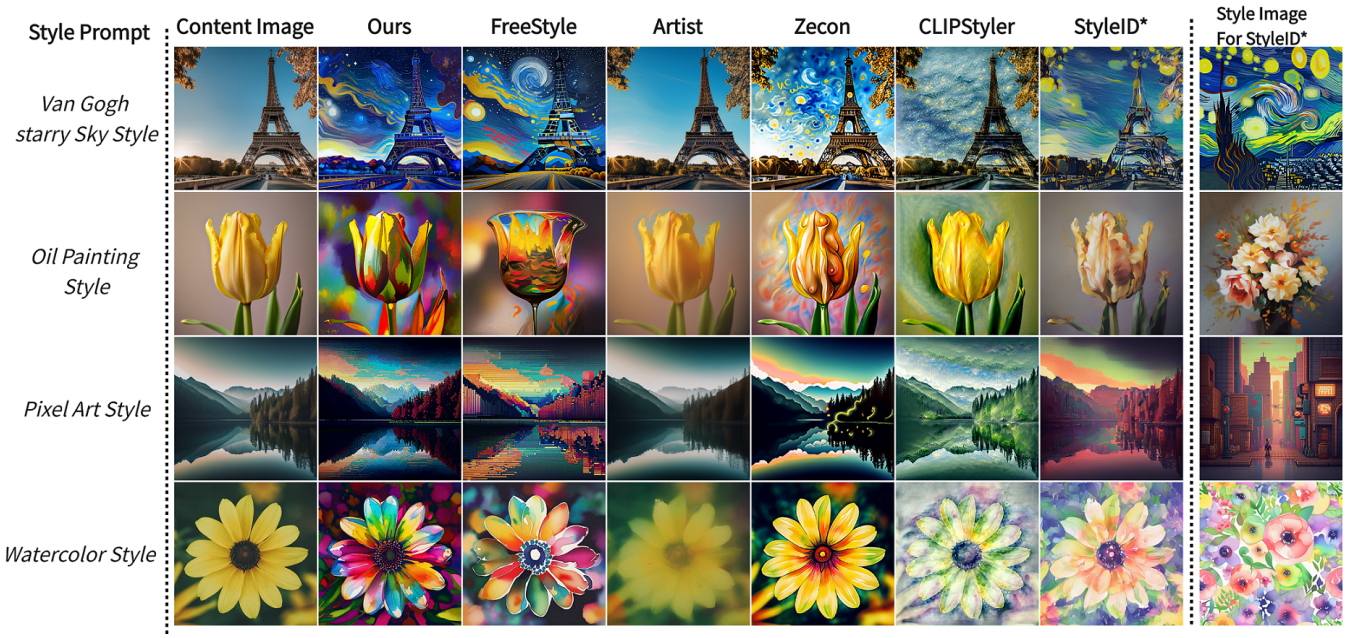

Figure 6: Qualitative comparison with state-of-the-art methods. We mainly compared the methods that use text descriptions as style references. While these methods typically take content images along with textual style prompts as input, StyleID uniquely requires both a content image and a style image. For StyleID, we first generate the style image (shown in the last column) via text-to-image generation based on each style prompt.

Methods	CLIP Aesthetic Score ↑	HPS ↑	CLIP Score ↑	SSIM ↑
StyleID [Chung *et al.*, 2024]	5.863	0.206	26.542	0.477
FreeStyle [He *et al.*, 2024]	<u>6.220</u>	<u>0.239</u>	27.142	0.469
Artist [Jiang and Chen, 2024]	5.823	0.182	23.228	**0.575**
Zecon [Yang *et al.*, 2023]	5.738	0.221	25.857	0.498
CLIPStyler [Kwon and Ye, 2022]	5.992	0.232	**31.085**	0.447
Ours	**6.385**	**0.246**	<u>27.228</u>	<u>0.515</u>

Table 1: Quantitative Comparison. For each metric, the **bolded data** represents the best result, while <u>the underlined data</u> indicates the second-best result.

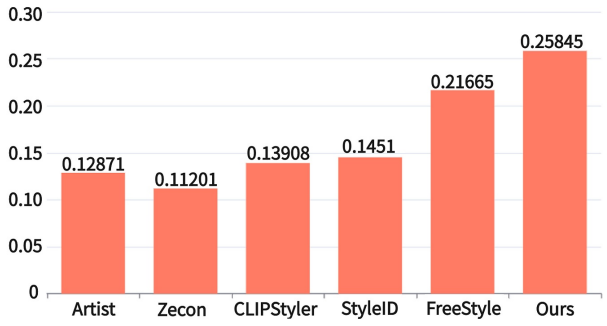

Figure 7: User study. We visualize the voting ratio for each method, indicating the popularity of each method's results. For each question in the questionnaire, participants were asked to select at least two answers they preferred, considering both content preservation and stylization effects.

Quantitative Comparison Results

Table 1 provides a quantitative comparison of our method with five competitive approaches. Our method achieves the highest scores in both the CLIP Aesthetic Score and HPS metrics, demonstrating that the results produced by our approach are more likely to be preferred by users. This aligns with the results of our user study. For the CLIP Score, our method ranks second, following CLIPStyler, which explicitly incorporates the CLIP Score as a constraint in its style transfer process. However, when evaluating performance, we prioritize user-centric metrics, such as the CLIP Aesthetic Score and HPS, as the CLIP Score is not considered the primary criterion for assessing quality of stylized results.

To compute the SSIM score, a fine-grained similarity metric, we used SAM [Kirillov *et al.*, 2023] to mask out blank regions in both the content images and the stylized ones. SSIM was then calculated only on areas containing distinct content structures, thereby minimizing the influence of stylistic elements in blank regions. While our method achieves the

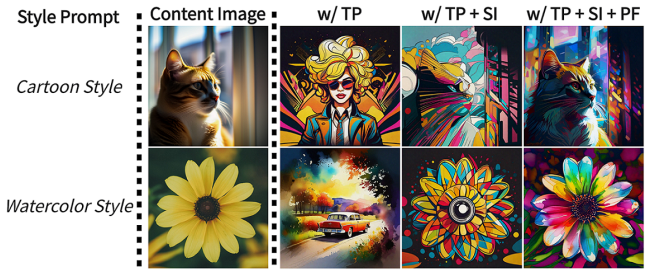

Figure 8: Ablation study of the modules. The third to fifth columns show results obtained with different settings: the basic diffusion model using only the style text prompt (TP) as input, the diffusion model with semantic injection (SI), and our complete model with both semantic injection (SI) and phase fusion (PF) modules.

second-highest SSIM score, Artist attains the highest score for this metric. However, Artist struggles to deliver high-quality style transfer results based on the given style descriptions (see Figure 6).

5.5 Ablation Study

Effect of Modules

We validate the effectiveness of the two key modules, namely the phase fusion module and the semantic injection module, by incorporating them incrementally. The results are presented in Figure 8. The baseline model (w/TP) refers to the original diffusion model that uses only the style text prompt as input. The model w/TP + SI integrates the semantic injection module, while w/TP + SI + PF represents our complete model, which includes both semantic injection and phase fusion modules. As shown in Figure 8, the addition of semantic injection helps to constrain the subjects, such as the cat and flower, but their structures remain inconsistent with the content images. By further incorporating phase fusion, our model achieves superior performance in both content preservation and stylization.

Effect of Hyperparameters α and β

The parameter α determines the extent to which the phase spectrum of the content image is introduced, while β controls the amount of style information retained. While a full grid search could be employed to determine these two hyperparameters, we simplified the process by initially setting both α and β to 0.5 and performing a localized grid search within specific ranges. First, we fixed β at 0.5 and adjusted α within the range of 0.3 to 0.7. We observed that increasing α further, which leads to a stronger influence from the content image's phase spectrum, introduces noticeable artifacts. In contrast, decreasing α results in weaker stylization. Based on this observation, we set α to 0.5. Next, with α fixed at 0.5, we tuned β within the range of 0.4 to 0.8. A smaller β resulted in strong artifacts, while a larger β reduced the effectiveness of style expression. Ultimately, we determined the optimal values as $\alpha = 0.5$ and $\beta = 0.7$, achieving a balance that ensured effectively stylized and artifact-free results. Once determined, these hyperparameters are fixed for all stylization tasks.

Figure 9: Ablation study of α and β. First, we fix the value of β and search for the optimal α. Then, we fix α and adjust β to find the best result with desirable color. Finally, we set α to 0.5 and β to 0.7. Once α and β are determined, they remain fixed for style transfer on any image.

6 Conclusion

In this work, we proposed a Fourier domain phase spectrum incorporation method to guide the diffusion model in preserving the content of an input image during style transfer. To further enhance content preservation, we introduced a mechanism that integrates semantic information from the content image. These two designs enable the diffusion model to achieve high-quality style transfer results while ensuring the generated image remains closely aligned with the input content image, in a completely training-free manner. Our method currently does not support transferring styles from image references. In the future, we plan to expand our model to accommodate various types of style references.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62176010), the Beijing Natural Science Foundation (No. 4252029) and the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. VRLAB 2024B02).

References

[Chen *et al.*, 2024] Dar-Yen Chen, Hamish Tennent, and Ching-Wen Hsu. Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8619–8628, June 2024.

[Choi *et al.*, 2025] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *European Conference on Computer Vision (ECCV)*, pages 206–235, 2025.

[Chung *et al.*, 2024] 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, June 2024.

[Esser *et al.*, 2021] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, June 2021.

[Gatys *et al.*, 2016] 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 (CVPR)*, pages 1–10, June 2016.

[Guo and Lin, 2024] Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6986–6996, June 2024.

[Hamazaspyan and Navasardyan, 2023] Mark Hamazaspyan and Shant Navasardyan. Diffusion-enhanced patchmatch: A framework for arbitrary style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 797–805, June 2023.

[He *et al.*, 2024] Feihong He, Gang Li, Mengyuan Zhang, Leilei Yan, Lingyu Si, Fanzhang Li, and Li Shen. Freestyle: Free lunch for text-guided style transfer using diffusion models. *arXiv preprint arXiv:2401.15636*, pages 1–23, 2024.

[Hu *et al.*, 2021] 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. *arXiv preprint arXiv:2106.09685*, pages 1–26, 2021.

[Huang and Belongie, 2017] 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 (ICCV)*, pages 1–10, Oct 2017.

[Jiang and Chen, 2024] Ruixiang Jiang and Changwen Chen. Artist: Aesthetically controllable text-driven stylization without training. *arXiv preprint arXiv:2407.15842*, pages 1–21, 2024.

[Jin *et al.*, 2024] Zhiyu Jin, Xuli Shen, Bin Li, and Xiangyang Xue. Style spectroscope: Improve interpretability and controllability through fourier analysis. *Machine Learning*, 113(6):3485–3503, 2024.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, October 2023.

[Kwon and Ye, 2022] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18062–18071, June 2022.

[Li *et al.*, 2017] 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 (NeurIPS)*, pages 1–11, 2017.

[Li *et al.*, 2024] Xinghui Li, Qichao Sun, Pengze Zhang, Fulong Ye, Zhichao Liao, Wanquan Feng, Songtao Zhao, and Qian He. Anydressing: Customizable multi-garment virtual dressing via latent diffusion models. *arXiv preprint arXiv:2412.04146*, pages 1–23, 2024.

[Li *et al.*, 2025] 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 (ECCV)*, pages 110–126, 2025.

[Lin *et al.*, 2024] Ente Lin, Xujie Zhang, Fuwei Zhao, Yuxuan Luo, Xin Dong, Long Zeng, and Xiaodan Liang. Dreamfit: Garment-centric human generation via a lightweight anything-dressing encoder. *arXiv preprint arXiv:2412.17644*, pages 1–16, 2024.

[Lv *et al.*, 2024] Xiaoqian Lv, Shengping Zhang, Chenyang Wang, Yichen Zheng, Bineng Zhong, Chongyi Li, and Liqiang Nie. Fourier priors-guided diffusion for zero-shot joint low-light enhancement and deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25378–25388, June 2024.

[Mo *et al.*, 2024] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7465–7475, June 2024.

[Qi *et al.*, 2024] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8693–8702, June 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763, Jul 2021.

[Rombach *et al.*, 2022] 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 (CVPR)*, pages 10684–10695, June 2022.

[Si *et al.*, 2024] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4733–4743, June 2024.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, pages 1–14, 2014.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, pages 1–22, 2020.

[Wang *et al.*, 2023] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7677–7689, October 2023.

[Wu *et al.*, 2023] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, pages 1–24, 2023.

[Yang and Soatto, 2020] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–11, June 2020.

[Yang *et al.*, 2023] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. Zero-shot contrastive loss for text-guided diffusion image style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22873–22882, October 2023.

[Yao *et al.*, 2024] Zishu Yao, Guodong Fan, Jinfu Fan, Min Gan, and C. L. Philip Chen. Spatial–frequency dual-domain feature fusion network for low-light remote sensing image enhancement. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

[Ye *et al.*, 2023] 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*, pages 1–16, 2023.

[Yu *et al.*, 2023] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23174–23184, October 2023.

[Zhang *et al.*, 2023a] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, October 2023.

[Zhang *et al.*, 2023b] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, December 2023.

[Zhang *et al.*, 2023c] 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, June 2023.

[Zhang *et al.*, 2024a] Zhanjie Zhang, Quanwei Zhang, Huaizhong Lin, Wei Xing, Juncheng Mo, Shuaicheng Huang, Jinheng Xie, Guangyuan Li, Junsheng Luan, Lei Zhao, et al. Towards highly realistic artistic style transfer via stable diffusion with step-aware and layer-aware prompt. *arXiv preprint arXiv:2404.11474*, pages 1–9, 2024.

[Zhang *et al.*, 2024b] Zhanjie Zhang, Quanwei Zhang, Wei Xing, Guangyuan Li, Lei Zhao, Jiakai Sun, Zehua Lan, Junsheng Luan, Yiling Huang, and Huaizhong Lin. Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 38(7):7396–7404, Mar. 2024.

[Zheng *et al.*, 2023] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22490–22499, June 2023.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–10, Oct 2017.