StyleFM: Frequency Manipulation Empowered by Recursive Attention on Diffusion Models for Arbitrary Style Transfer

Yingnan Ma[1], **Zhenye Liu**[1], **Siying Liu**[1], **Anup Basu**[1*]

[1] University of Alberta, Edmonton, Canada
{ma4, zhenye2, siying13, basu}@ualberta.ca

Abstract

Given the remarkable performance of diffusion models in image generation, recent research has been exploring their adaptation to style transfer. However, current diffusion-based approaches encounter persistent challenges, such as style distortions and the reliance on textual prompts for content preservation. To address these limitations, we introduce **StyleFM**, a novel training-free diffusion-based style transfer approach that incorporates optimization strategies into both the frequency and temporal domains. The proposed method provides two core innovations: (1) **Tripartite Frequency Manipulation**: To more precisely tailor frequency manipulation, StyleFM introduces a tripartite frequency design with a buffer band accounting for the overlap of content and style representations. In addition, StyleFM designs a frequency superposition editing method to achieve frequency enhancement. (2) **Recursive Attention**: StyleFM proposes the recursive attention strategy within the diffusion process, which facilitates the progressive and consistent injection of style information throughout the temporal process without reliance on text guidance. Experiments demonstrate that StyleFM outperforms state-of-the-art methods. It effectively preserves content fidelity while achieving sufficient style embedding.

Code — https://github.com/YingnanMa/StyleFM

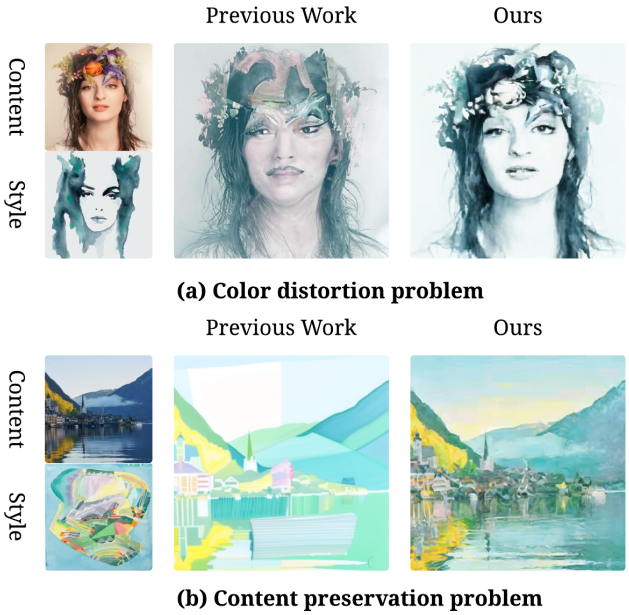

Figure 1: Illustration of existing challenges of diffusion-based approaches.

Introduction

Arbitrary style transfer (Park and Lee 2019; Deng et al. 2021) focuses on the capability of applying an arbitrary style to a target image without the restriction of style categories. It was originally proposed by Gatys et al. (Gatys, Ecker, and Bethge 2016), utilizing convolutional neural networks for iterative style embedding. AdaIN (Huang and Belongie 2017) further proposed adaptive instance normalization to solve style transfer by aligning the statistical features. Later on, transformer-based approaches (Wu et al. 2021; Deng et al. 2022) were proposed utilizing self-attention mechanisms. In addition, flow-based (Ho et al. 2019; An et al. 2021) methods were proposed to solve the content leakage problem.

In recent years, with the success of diffusion models in image generation tasks (Koo et al. 2024; Gao et al. 2024), diffusion-based approaches (Kwon and Ye 2022; Xu et al. 2025) have been explored for style transfer and have encountered persistent challenges. As shown in Figure 1, due to residual color artifacts from the content image and insufficient style embedding, diffusion-based approaches suffer from the style distortion problem. In addition, diffusion-based approaches rely on textual prompts to preserve content integrity. When the image information is complex and the text prompt fails to describe the information accurately, the content protection can be seriously affected.

To overcome the limitations, we propose StyleFM, a novel training-free diffusion-based style transfer approach driven by frequency manipulation. In the frequency domain, low and high frequencies are generally associated with style and content features, respectively (Koo et al. 2024). However, due to the inherent overlap between content and style information, directly applying frequency-based editing may compromise accuracy. To address this issue, we introduce

*Corresponding author

a tripartite frequency design, where a mid-frequency buffer band explicitly accounts for the overlap between style and content, enabling more precise and effective frequency-based manipulation for arbitrary style transfer. In addition, StyleFM further advances frequency editing strategies. In addition to conventional frequency filtering, we propose a frequency enhancement mechanism based on the superposition of target frequency bands, which strengthens both content preservation and style fidelity. Finally, StyleFM integrates temporal-domain optimization to refine the style injection process. We optimize the reverse procedure through a recursive attention mechanism, ensuring that the effects of frequency-domain manipulation are effectively and progressively propagated throughout the diffusion process.

Experiments prove that StyleFM achieves superior performance compared to state-of-the-art methods, as measured by both qualitative and quantitative assessments. StyleFM maintains content fidelity without textual prompts, and achieves sufficient style embedding while eliminating color distortions. The key contributions of this work are:

- We introduce a tripartite frequency design that enables more precise frequency manipulation for arbitrary style transfer, by accounting for the overlap between content and style representations.
- We propose a frequency enhancement mechanism, based on the superposition of target frequency bands, to improve both content preservation and style consistency.
- We propose a novel recursive attention strategy so that the effect of frequency manipulation can be effectively brought to the temporal domain for style injection.
- We conduct extensive qualitative and quantitative evaluations to validate the effectiveness of StyleFM, including the design of a texture-sensitive evaluation metric.

Related Work

Attention-based Style Transfer

Attention-based methods have made significant strides in enhancing style fidelity and content preservation. SANet (Park and Lee 2019) introduced self-attention for feature adaptation, while IE (Chen et al. 2021) improved content protection using contrastive and internal-external learning. StyleFormer (Wu et al. 2021) integrated a style bank to improve style representation. StyTr^2 (Deng et al. 2022) employed both spatial and temporal attention to model global dependencies. In addition, restoration-based approaches such as CAST (Zhang et al. 2022), RAST 1.0 (Ma et al. 2023), and RAST 2.0 (Ma et al. 2024) were explored to mitigate content leakage. By contrast, AesPA (Hong et al. 2023) proposed the pattern-aware transformer through pattern repeatability. S2WAT (Zhang et al. 2024) adopted a hierarchical vision transformer to extract multi-scale features and proposed strip window attention for dependency modeling. SCSA (Shang et al. 2025) introduced semantic continuous sparse attention to maintain semantic alignment.

Diffusion-based Style Transfer

With the success of Denoising Diffusion Probabilistic Models (DDPM) (Sohl-Dickstein et al. 2015; Nichol and Dhariwal 2021) in image generation, Denoising Diffusion Implicit Models (DDIM) (Song, Meng, and Ermon 2021) were introduced to reduce the stochastic nature of DDPM by employing a deterministic, non-Markovian sampling process. Building on these foundations, Latent Diffusion Models (LDM) (Rombach et al. 2022) improved the diffusion process in a compressed latent space. Leveraging LDM, pretrained Stable Diffusion Models (SDM) (Rombach et al. 2022) have been widely adopted for image generation and style transfer in a training-free manner.

Guided by textual prompts, DiffuseIT (Kwon and Ye 2022) employed the CLIP model (Gal et al. 2022) to perform style transfer within the DDPM sampling process. InST (Zhang et al. 2023) further enhanced semantic alignment by incorporating the inversion process, enabling style transfer without relying on detailed textual descriptions. In the frequency domain, FCNet (Gao et al. 2024) introduced frequency editing by designing four frequency filters to facilitate content generation under text guidance. FlexiEdit (Koo et al. 2024) proposed a low-pass filtering approach, which designs a frequency boundary to separate content and style information. Inspired by FlexiEdit, SSP (Xu et al. 2025) introduced a Negative Guidance strategy to achieve text-guided style transfer. It integrated ControlNet (Zhang, Rao, and Agrawala 2023) to mitigate content leakage.

Beyond text-guided approaches, diffusion models have also been employed without text prompts. StyleDiffusion (Wang, Zhao, and Xing 2023) introduced a disentanglement loss to facilitate the separation of content and style representations. BBDM (Li et al. 2023) proposed a Brownian Bridge diffusion process that enables domain translation through a bidirectional diffusion process. Additionally, StyleID (Chung, Hyun, and Heo 2024) enhanced diffusion-based style transfer by incorporating cross-attention mechanisms. It fused the stylized query with the content query to preserve structural details, and introduced temperature scaling to adaptively modulate attention weights.

Differences

While prior works have explored frequency manipulation and attention mechanisms to enhance diffusion models, they remain fundamentally different from our proposed approach. For instance, FlexiEdit and SSP employed a single frequency boundary to separate content and style features through a $highpass = 1 - lowpass$ design. In contrast, StyleFM introduces a tripartite frequency design that includes a mid-frequency buffer band, explicitly accounting for the overlap between content and style. Moreover, StyleFM extends beyond conventional frequency filtering by introducing a frequency superposition editing strategy, which enhances content and style consistency through additive manipulation of targeted frequency bands. Beyond frequency manipulation, attention-based strategies have also been integrated into diffusion models. Unlike StyleID, which uses stylized queries to preserve content structure, StyleFM proposes a recursive attention mechanism that effectively propagates the effect of frequency-domain modifications into the temporal domain, ensuring that their impact persists throughout the injection process.

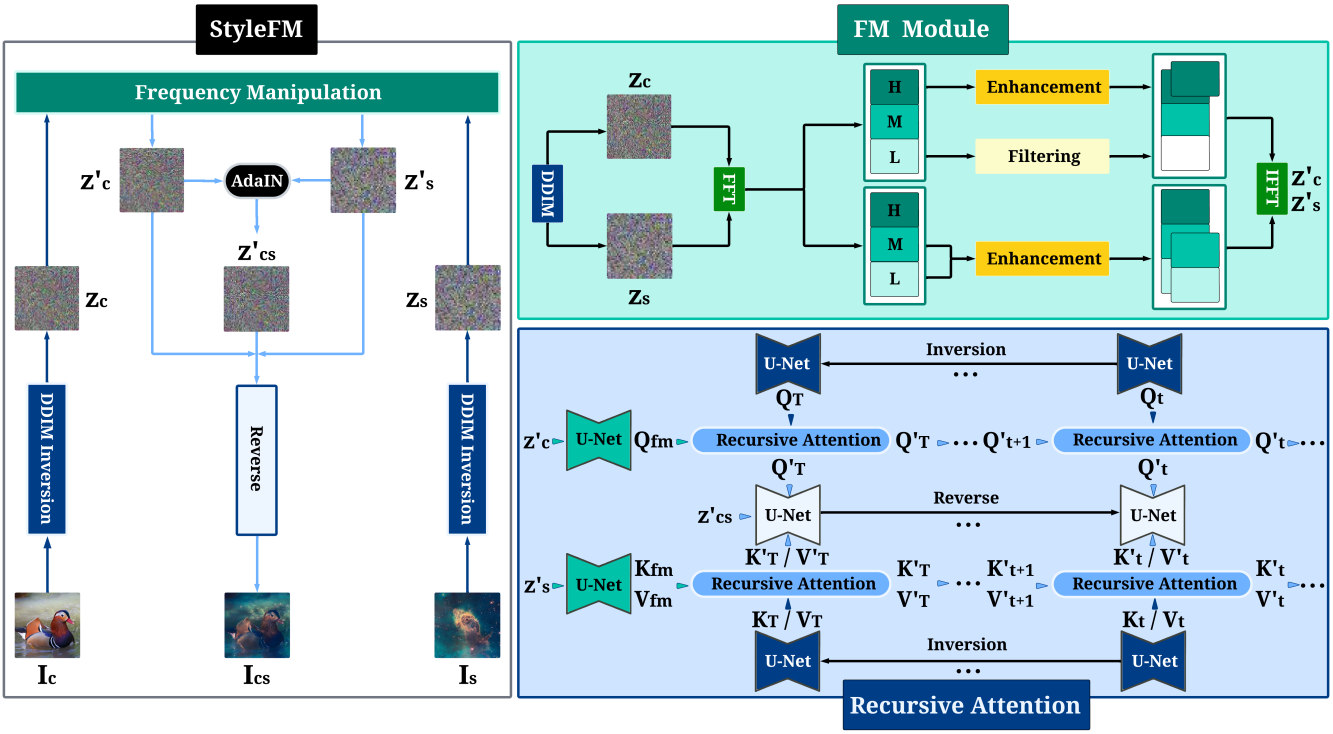

Figure 2: Overall architecture of StyleFM with frequency manipulation module and recursive attention details.

Method

Diffusion-based style transfer methods often suffer from color distortion and rely on text prompts for content preservation. To address these issues, we propose StyleFM, a novel training-free diffusion-based style transfer method that introduces improvements in both the frequency and temporal domains. StyleFM leverages a tripartite frequency manipulation design, incorporating a buffer band to handle the overlap between content and style frequencies, thereby enhancing the precision of frequency editing. In the temporal domain, a recursive attention mechanism is introduced to propagate the benefits of frequency manipulation throughout the diffusion process, achieving a more effective balance between content preservation and style embedding.

Diffusion-based Style Embedding

The overall workflow of StyleFM is illustrated in Figure 2. We adopt the Stable Diffusion Model (SDM) (Rombach et al. 2022) with DDIM sampling (Song, Meng, and Ermon 2021) to accelerate the diffusion process. SDM is trained in the latent space as described below:

$$\mathcal{L}_{SDM} := \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t\sim(0,T)}\left[\left\|\epsilon - \epsilon_\theta(z^t, t)\right\|_2^2\right] \quad (1)$$

where z denotes the latent representation, ϵ is the Gaussian noise, and t denotes the discrete timesteps ranging from 0 to T. The noise prediction $\epsilon_\theta(z^t, t)$ is generated by a U-Net model. In StyleFM, we adopt pre-trained SDM to perform arbitrary style transfer under a training-free strategy.

In the initial stage of the DDIM process, the content image I_c and style image I_s are encoded into latent representations using a pre-trained VAE encoder. Through DDIM forward sampling, the initial latents z_c^0 and z_s^0 are progressively diffused, yielding z_c and z_s at timestep T. We then apply tripartite frequency manipulation in the frequency domain to edit content and style representations, resulting in the modified latents z_c' and z_s'. The frequency manipulation module is described in detail in the following subsection.

To obtain the initial stylized latent z_{cs}', we adopt AdaIN (Huang and Belongie 2017; Chung, Hyun, and Heo 2024) to embed the modified latent z_s' on z_c' as below:

$$z_{cs}' = \sigma(z_s')\left(\frac{z_c' - \mu(z_c')}{\sigma(z_c')}\right) + \mu(z_s') \quad (2)$$

where σ and μ represent standard deviation and mean functions. During the reverse process, we introduce a recursive attention mechanism to progressively embed the style latent on the content latent, producing the final stylized latent z_{cs}^0. The stylized image I_{cs} is obtained using the VAE decoder. The full reverse process is outlined in the final subsection.

Tripartite Frequency Design

In arbitrary style transfer, content and style information often overlap across frequency bands, making precise separation challenging. As illustrated in Figure 3, filtering out high-frequency components leads to significant content degradation, including severe blurring and edge loss, along with substantial loss of texture and color. Removing mid-frequency components results in moderate edge alteration

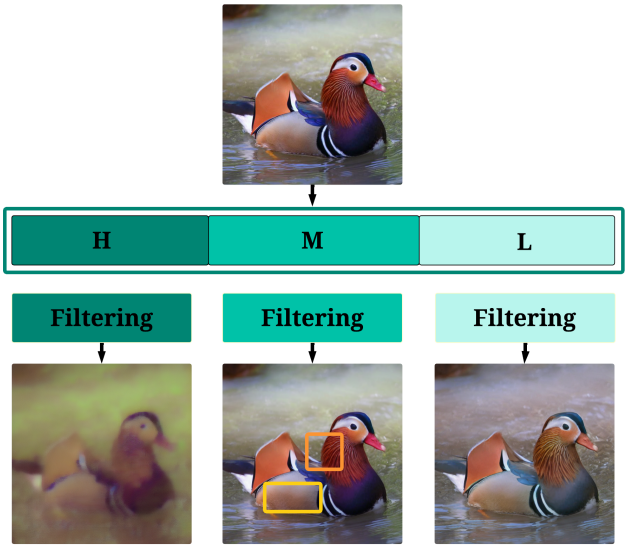

Figure 3: Filtering results on different frequency bands. The yellow box highlights the texture changes and the orange box highlights the edge changes.

(orange box) and mild reduction in texture and color (yellow box). Filtering low frequencies primarily affects color, while texture is slightly diminished and content structure remains relatively unaffected.

These observations indicate that content features predominantly reside in high frequencies, with fine details extending into mid and low frequency bands. In contrast, color features span the entire spectrum, while texture is concentrated in the high and mid frequencies. Given this overlap, a single frequency boundary is insufficient for clean separation. To address this, we propose a tripartite frequency design, introducing a mid-frequency buffer band to better isolate content and style, thus minimizing information loss during frequency manipulation.

Frequency Manipulation

Given the content latent z_c and style latent z_s, we apply the 2D Fast Fourier Transform (FFT) (Koo et al. 2024) to convert them into the frequency domain, followed by the inverse FFT (IFFT) to reconstruct the modified latent representations. The transformations are defined as follows:

$$f = FFT(z), \quad z = IFFT(f). \tag{3}$$

Building on the tripartite frequency design, we introduce different frequency manipulations for content and style latents. As shown in Figure 2, we apply low-frequency filtering to the content frequency f_c to suppress residual color, thereby enhancing style embedding effectiveness. In style transfer, content distortion often stems not from added color, but from matches between low-frequency content and high-frequency texture. Thus, removing low-frequency content helps stabilize content structure. To preserve content features, we avoid filtering in the high-frequency and mid-frequency band to prevent loss from overlapping information. For the style latent, no filtering is applied, since color spans all frequencies.

To achieve low-frequency filtering on f_c, we design a low-band filter F_L based on the Gaussian low-pass filter F_{LP} (Koo et al. 2024). The equation of the low-band filter is given below:

$$F_L = 1 - F_{LP}. \tag{4}$$

The equation for low-band filtered content frequency f_c^F can be represented as:

$$f_c^F = f_c \odot F_L, \tag{5}$$

where the $\odot$ represents the element-wise multiplication.

Beyond frequency filtering, we introduce a frequency enhancement technique to further improve content preservation and style consistency. For the content latent, we selectively amplify the high-frequency band, where content details are primarily concentrated, while avoiding enhancement of the mid-frequency buffer to prevent the color residual. For the style latent, we enhance both low and mid frequency bands to reinforce color and moderate texture features. The high-frequency band is left unchanged to avoid the content changes resulting from excessive textures.

To perform frequency enhancement, we introduce a superposition editing method using content and style superposition parameters, denoted by α and β, respectively. A Gaussian low-pass filter F_{LP} is adapted to form a low-mid-pass filter F_{LMP}, defined by the mid-high frequency boundary. For content enhancement, we derive a low-mid-band filter F_{LM} from F_{LMP} to get the high-frequency band.

$$F_{LM} = 1 - F_{LMP} \tag{6}$$

The equation for high-band enhanced content frequency f_c^E is given below:

$$f_c^E = f_c + \alpha * f_c \odot F_{LM}. \tag{7}$$

For style enhancement, the equation for low-mid-band enhanced style frequency f_s^E is:

$$f_s^E = f_s + \beta * f_s \odot F_{LMP}. \tag{8}$$

The final modified content frequency f_c' and modified style frequency f_s' can be expressed as:

$$f_c' = f_c^F + f_c^E - f_c, \quad f_s' = f_s^E. \tag{9}$$

After the IFFT function, the modified frequency can be tranformed into the modified latents z_c' and z_s'.

Recursive Attention

In StyleFM, the reverse process is tailored for style embedding. After applying tripartite frequency manipulation, we obtain the modified latent z_{cs}'. However, the features used in the reverse phase originate from the inversion steps and are not optimized by frequency editing. To effectively transfer the benefits of frequency manipulation into the temporal domain, we introduce a recursive attention mechanism within the reverse process.

We design the recursive attention based on the self-attention mechanism. During the inversion process, we obtain the query map Q_t from the self-attention layer of the U-Net by using content latent at time t. Similarly, we get the

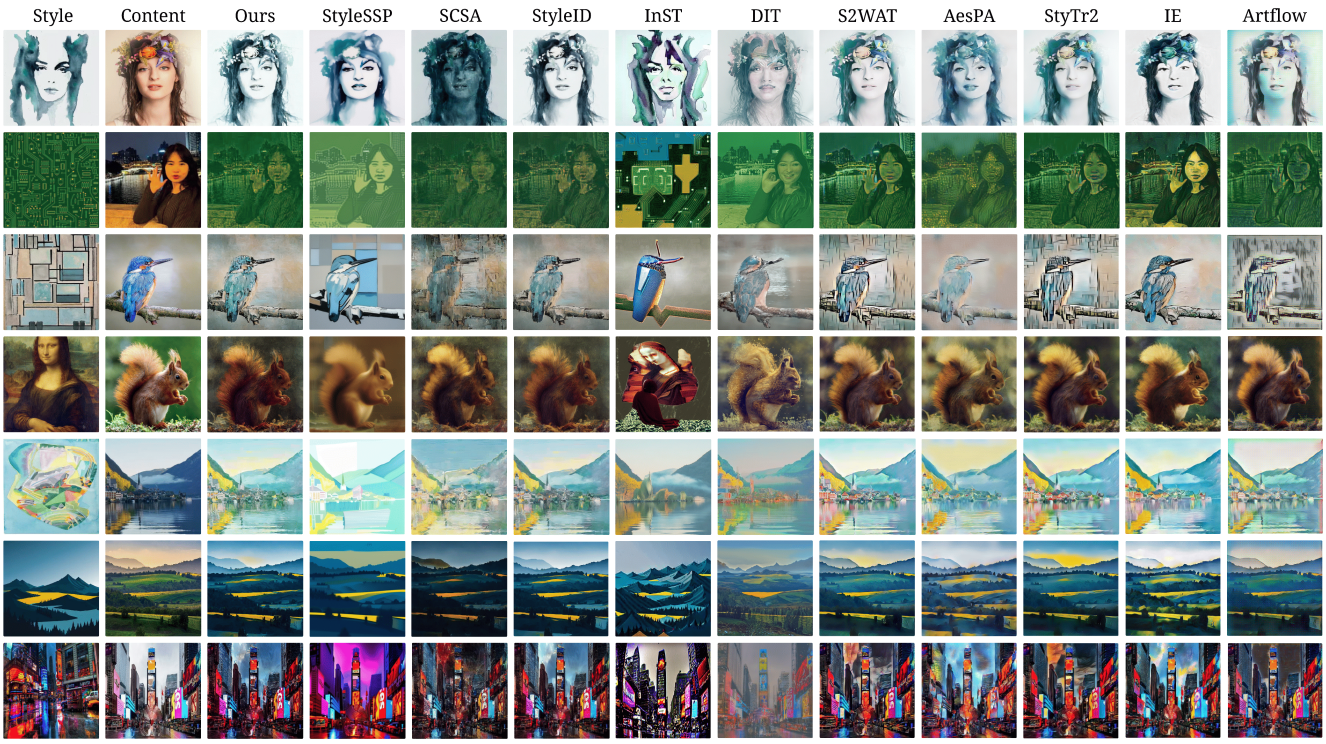

Figure 4: Visualization of stylized results for comparisons with state-of-the-art approaches. Artistic style transfer results are illustrated in the 1^{st}-6^{th} rows and photo-realistic style transfer result is depicted in the 7^{th} row.

key map K_t and value map V_t from the style latent. To involve the optimization effect from frequency manipulation, we obtain modified query map Q_{fm} from the modified content latent z_c'. Similarly, we get modified key map K_{fm} and modified value map V_{fm} from the modified style latent z_s'.

In order to avoid large feature gaps, we do not directly add Q_{fm}, K_{fm}, V_{fm} to Q_t, K_t, V_t at each timestep. Instead, we combine Q_{fm}, K_{fm}, V_{fm} with Q_T, K_T, V_T through a recursion parameter λ in the first step of the reverse process. In the subsequent timesteps, we combine Q_t, K_t, V_t with Q'_{t+1}, K'_{t+1}, V'_{t+1} through the recursion parameter, which can indirectly propagate the frequency editing effects into the subsequent timesteps. The equations of recursive maps Q'_t, K'_t, V'_t are defined below:

$$Q'_t = \begin{cases} Q_T + \lambda Q_{fm}, & t = T \\ Q_t + \lambda Q'_{t+1}, & t \neq T \end{cases} \tag{10}$$

$$K'_t = \begin{cases} K_T + \lambda K_{fm}, & t = T \\ K_t + \lambda K'_{t+1}, & t \neq T \end{cases} \tag{11}$$

$$V'_t = \begin{cases} V_T + \lambda V_{fm}, & t = T \\ V_t + \lambda V'_{t+1}, & t \neq T. \end{cases} \tag{12}$$

Based on the self-attention mechanism, the equation for the recursive attention is shown below:

$$Attn(Q'_t, K'_t, V'_t) = \text{softmax}\left(\frac{Q'_t K'_t}{\sqrt{d}}\right) \cdot V'_t. \tag{13}$$

Experiments

Experimental Settings

We use the pre-trained Stable Diffusion 1.4 model (Rombach et al. 2022) with DDIM sampling (Song, Meng, and Ermon 2021) for 50 timesteps. For StyleFM, the tripartite frequency boundaries are set to 0.5 (low-mid) and 0.7 (mid-high). The content and style superposition parameters, α and β, are set to 0.5 and 0.3, respectively, while the recursion parameter λ is set to 0.1. All experiments are performed on a single Nvidia V100 GPU.

Dataset: We utilize the publicly available test set from StyleID (Chung, Hyun, and Heo 2024), which consists of 800 image pairs formed by randomly selecting 20 content images from MS-COCO (Lin et al. 2014) and 40 style images from WikiArt (Karayev et al. 2013). All images are resized to 512×512 to ensure consistency in evaluation.

Metric: Following RAST 1.0 (Ma et al. 2023), we employ the Learned Perceptual Image Patch Similarity ($LPIPS_s$) (Zhang et al. 2018) to assess style consistency between the style and stylized images. As a texture-sensitive metric, $LPIPS_s$ is well-suited for evaluating artistic styles where texture and brushstrokes play a critical role. To assess content preservation, we use both $LPIPS_c$ and $CFSD$ (Chung, Hyun, and Heo 2024) to measure the similarity between the content and stylized images. Additionally, inspired by ArtFID (Wright and Ommer 2022), we propose a new metric, $LPIPS_{Art}$, to provide a comprehensive evalua-

Metric	**Ours**	**SSP**	**SCSA**	**StyleID**	**InST**	**DIT**	**S2WAT**	**AesPA**	**StyTr^2**	**IE**	**Artflow**
$LPIPS_s \downarrow$	**0.693**	0.831	0.735	0.708	0.744	0.759	0.706	0.720	0.702	0.737	0.700
$LPIPS_c \downarrow$	**0.496**	0.622	0.714	0.506	0.716	0.555	0.517	0.512	0.545	0.508	0.556
$CFSD \downarrow$	**0.219**	0.716	0.325	0.228	0.660	0.263	0.263	0.246	0.301	0.268	0.292
$LPIPS_{Art} \downarrow$	**2.532**	2.970	2.974	2.571	2.993	2.734	2.587	2.601	2.629	2.621	2.645

Table 1: Quantitative comparisons with diffusion-based (3^{rd}-7^{th} columns) and conventional (8^{th}-12^{th} columns) approaches.

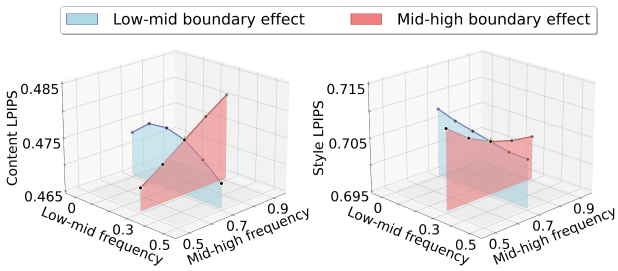

Figure 5: Effects of low-mid and mid-high frequency boundaries.

Figure 6: Visualization of the effects of content superposition parameter.

α	**0**	**0.1**	**0.3**	**0.5**
$LPIPS_s$	0.7038	0.7039	0.7040	0.7042
$LPIPS_c$	0.4717	0.4708	0.4696	0.4689

Table 2: Effects of content superposition parameter.

tion of overall style transfer performance. $LPIPS_{Art}$ leverages the $LPIPS$ metric for evaluating both content preservation and style consistency, thereby ensuring metric alignment and reducing bias introduced by using heterogeneous evaluation metrics. The formulation is as follows:

$$LPIPS_{Art} = (1 + LPIPS_c) * (1 + LPIPS_s). \quad (14)$$

Qualitative Comparisons

To demonstrate the effectiveness of StyleFM, we present qualitative comparisons with ten state-of-the-art style transfer methods including five diffusion-based approaches: StyleSSP (SSP) (Xu et al. 2025), SCSA (StyleID-base) (Shang et al. 2025), StyleID (Chung, Hyun, and Heo 2024), InST (Zhang et al. 2023), and DiffuseIT (DIT) (Kwon and Ye 2022). We also conduct comparisons with five conventional methods: S2WAT (Zhang et al. 2024), AesPA (Hong et al. 2023), StyTr^2 (Deng et al. 2022), IE (Chen et al. 2021), and ArtFlow (An et al. 2021). All experiments are conducted using the public implementations with their default settings.

Figure 4 illustrates that diffusion-based methods often struggle with content preservation. While SSP benefits from ControlNet for better contour protection, it still exhibits noticeable loss of fine details (rows 1, 3, 4, 5, and 6). SCSA fails to maintain a proper style-content balance, leading to significant content changes (rows 3 and 5). InST and DIT also show varying degrees of content leakage. In comparison, StyleFM and StyleID demonstrate superior content preservation. Regarding style consistency, SSP shows evident color distortions and insufficient texture embedding (rows 1–5), while SCSA introduces excessive style features, resulting in deviations from the original style (rows 1, 3, and 6). StyleID inadequately embeds color and texture details (rows 1 and 3), with InST and DIT performing even worse, retaining content color in the stylized results. In contrast, StyleFM achieves richer texture and color transfer, delivering notably better style fidelity.

Similar to diffusion-based approaches, conventional approaches also face the color distortion and content protection challenges. S2WAT, StyTr^2, and ArtFlow exhibit residual color from content images (row 1) and noticeable content leakage (row 3). Meanwhile, AesPA and IE struggle with insufficient style embedding (rows 2, 3, 6, and 7).

Quantitative Comparisons

Beyond qualitative evaluation, we also perform quantitative comparisons, as summarized in Table 1. StyleFM consistently outperforms both diffusion-based and conventional methods, achieving the lowest $LPIPS_s$ score, which reflects enhanced style consistency. It also reports the lowest $LPIPS_c$ and $CFSD$ values, indicating strong content preservation. Moreover, StyleFM ranks first in the overall performance metric $LPIPS_{Art}$. These results demonstrate that StyleFM achieves state-of-the-art performance across all evaluated criteria: $LPIPS_s$, $LPIPS_c$, $CFSD$, and $LPIPS_{Art}$, indicating superior style consistency and content preservation performance.

Analyzing the effects of frequency boundaries

In the tripartite frequency design, the mid-high boundary primarily influences frequency enhancement. As this boundary increases, the proportion of enhanced content frequency decreases while the enhanced style frequency expands. Consequently, $LPIPS_c$ increases due to reduced content enhancement, and $LPIPS_s$ decreases as style embedding be-

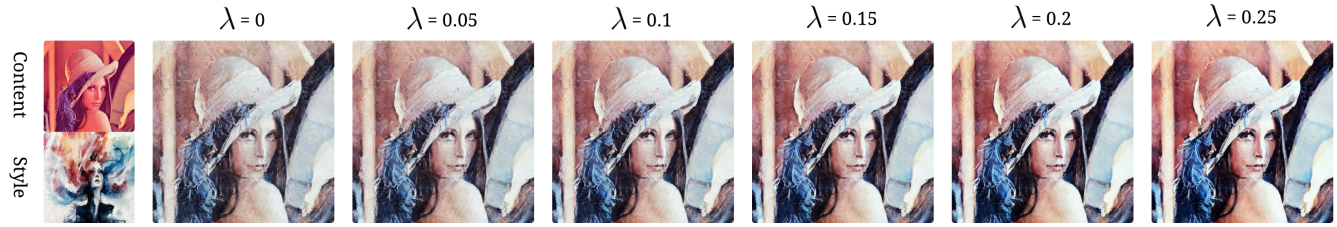

Figure 7: Visualization of the effects of recursion parameter.

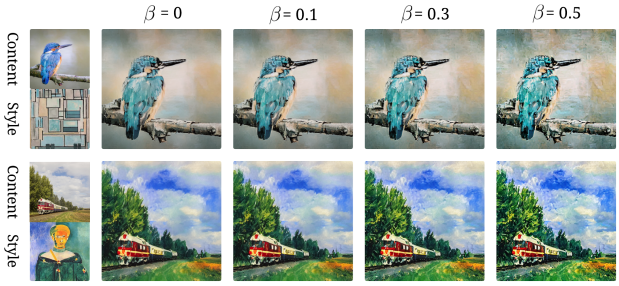

Figure 8: Visualization of the effects of style superposition parameter.

comes more prominent (Figure 5). In contrast, the low-mid boundary affects content filtering only. Initially, frequency filtering removes fine details slightly, causing a temporary rise in $LPIPS_c$. As the boundary shifts further, some low-frequency content information that is more prone to change is filtered, avoiding content changes caused by high-frequency texture matching, resulting in improved content stability and a decline in $LPIPS_c$. Simultaneously, more residual style elements are removed from the content representation, allowing for richer style incorporation and thus a continued decrease in $LPIPS_s$.

Analysis on the enhancement parameter α and β

As illustrated in Figure 6, content enhancement leads to clearer structural details, such as the eye socket and eyeball. As shown in Table 2, increasing the content superposition parameter α strengthens content preservation, evidenced by a decrease in $LPIPS_c$. Meanwhile, $LPIPS_s$ exhibits a slight increase with minimal compromise in style fidelity.

As shown in Figure 8, when the style superposition parameter β is low, texture is difficult to be embedded at the beginning due to the content preservation effects of recursive attention, while lower-frequency attributes like color are more easily embedded. As β increases, enhanced mid-frequency textures are prominently added.

Analysis on the recursion parameter λ

During the reverse diffusion process, the enhanced style and content from frequency editing are propagated by recursive attention. Without recursive attention, low-frequency content remains in the query maps while high-frequency content is underrepresented, which leads to greater content changes and more prominent style textures, as shown in Figure 7. In contrast, recursive attention progressively integrates the effects of content filtering and content enhancement into the query maps over time. The low-frequency content information that is prone to change is filtered, and the high-frequency content information is enhanced, which weakens the texture embedding and avoiding the changes in content information. Furthermore, increasing the recursion parameter λ also strengthens color embedding, as recursive attention updates the key and value maps along with the queries.

Modules	Baseline	+ FM	+ RA	StyleFM
$LPIPS_s$	0.72	0.6891	0.7256	0.6928
$LPIPS_c$	0.5091	0.5192	0.4517	0.4955

Table 3: Ablation study on the effects of innovative modules.

Ablation Study

To assess the effectiveness of the proposed components, we conduct quantitative ablation studies. The baseline model excludes both the tripartite frequency manipulation (FM) and recursive attention (RA) modules. As presented in Table 3, incorporating the FM module significantly improves style embedding but leads to a reduction in content consistency. In contrast, the RA module notably enhances content preservation while suppressing the incorporation of style features. By integrating both modules, our StyleFM method achieves a more balanced trade-off between style and content, ultimately outperforming the baseline in both style consistency and content preservation metrics.

Conclusion

We presented **StyleFM**, a novel training-free style transfer method based on diffusion models, which integrates optimization strategies across both frequency and temporal domains. StyleFM introduces a tripartite frequency formulation, incorporating a buffer band to address the overlap between content and style features. To further alleviate style distortion, the method combined frequency filtering with targeted frequency enhancement. Moreover, a recursive attention mechanism was introduced during the reverse diffusion process, enhancing content preservation performance without reliance on text prompts. Extensive experiments illustrated that StyleFM surpasses existing state-of-the-art approaches, achieving superior content fidelity alongside effective style embedding.

References

An, J.; Huang, S.; Song, Y.; Dou, D.; Liu, W.; and Luo, J. 2021. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 862–871.

Chen, H.; Wang, Z.; Zhang, H.; Zuo, Z.; Li, A.; Xing, W.; Lu, D.; et al. 2021. Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems*, 34: 26561–26573.

Chung, J.; Hyun, S.; and Heo, J.-P. 2024. 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*, 8795–8805.

Deng, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; and Xu, C. 2021. Arbitrary video style transfer via multi-channel correlation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1210–1217.

Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Pan, X.; Wang, L.; and Xu, C. 2022. StyTr2: Image Style Transfer with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11326–11336.

Gal, R.; Patashnik, O.; Maron, H.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4): 1–13.

Gao, X.; Xu, Z.; Zhao, J.; and Liu, J. 2024. Frequency-controlled diffusion model for versatile text-guided image-to-image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1824–1832.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

Ho, J.; Chen, X.; Srinivas, A.; Duan, Y.; and Abbeel, P. 2019. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning*, 2722–2730. PMLR.

Hong, K.; Jeon, S.; Lee, J.; Ahn, N.; Kim, K.; Lee, P.; Kim, D.; Uh, Y.; and Byun, H. 2023. Aespa-net: Aesthetic pattern-aware style transfer networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22758–22767.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Karayev, S.; Trentacoste, M.; Han, H.; Agarwala, A.; Darrell, T.; Hertzmann, A.; and Winnemoeller, H. 2013. Recognizing image style. *arXiv preprint arXiv:1311.3715*.

Koo, G.; Yoon, S.; Hong, J. W.; and Yoo, C. D. 2024. Flexiedit: Frequency-aware latent refinement for enhanced non-rigid editing. In *European Conference on Computer Vision*, 363–379. Springer.

Kwon, G.; and Ye, J. C. 2022. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*.

Li, B.; Xue, K.; Liu, B.; and Lai, Y.-K. 2023. BBDM: Image-to-Image Translation With Brownian Bridge Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Ma, Y.; Zhao, C.; Huang, B.; Li, X.; and Basu, A. 2024. RAST: Restorable Arbitrary Style Transfer. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(5): 1–21.

Ma, Y.; Zhao, C.; Li, X.; and Basu, A. 2023. RAST: Restorable arbitrary style transfer via multi-restoration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 331–340.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, 8162–8171. PMLR.

Park, D. Y.; and Lee, K. H. 2019. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5880–5888.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shang, C.; Wang, Z.; Wang, H.; and Meng, X. 2025. SCSA: A Plug-and-Play Semantic Continuous-Sparse Attention for Arbitrary Semantic Style Transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13051–13060.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. pmlr.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Wang, Z.; Zhao, L.; and Xing, W. 2023. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7677–7689.

Wright, M.; and Ommer, B. 2022. Artfid: Quantitative evaluation of neural style transfer. In *DAGM German Conference on Pattern Recognition*, 560–576. Springer.

Wu, X.; Hu, Z.; Sheng, L.; and Xu, D. 2021. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14618–14627.

Xu, R.; Xi, W.; Wang, X.; Mao, Y.; and Cheng, Z. 2025. StyleSSP: Sampling StartPoint Enhancement for Training-free

Diffusion-based Method for Style Transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 18260–18269.

Zhang, C.; Xu, X.; Wang, L.; Dai, Z.; and Yang, J. 2024. S2wat: Image style transfer via hierarchical vision transformer using strips window attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7024–7032.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Y.; Huang, N.; Tang, F.; Huang, H.; Ma, C.; Dong, W.; and Xu, C. 2023. Inversion-Based Style Transfer with Diffusion Models. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Lee, T.-Y.; and Xu, C. 2022. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–8.