
From Injection to Editing: Diffusion-Based Style Transfer

Yinuo Wang, Xinyue Meng, Xianfei Yang, Yixiang Wang
Carnegie Mellon University
Pittsburgh, PA 15213

1 Abstract

Recent advances in diffusion models have demonstrated remarkable potential for artistic style transfer. However, existing methods often require computationally expensive optimization or suffer from difficulties in balancing content preservation and style expression, particularly in text-guided scenarios. In this work, we propose a novel framework that combines cross-attention-based feature manipulation and dual diffusion processing to enable efficient and flexible style transfer. Specifically, we achieve training-free style injection and style blending by directly substituting the key and value features of self-attention layers during the generation process. Furthermore, we introduce a Dual Diffusion architecture to realize prompt-based style editing, controlling the trade-off between content preservation and style expression. Additionally, we incorporate cross-modal textual guidance and introduce a content-based learnable noise initialization to better preserve structural details. Extensive experiments demonstrate that our method achieves competitive content preservation, style faithfulness, and flexibility compared to both traditional and diffusion-based baselines. Code and implementation details will be made publicly available.

Code: <https://github.com/Serenii02/11785-Final-Project>

Video: <https://youtu.be/A22bpiJiSXY>

2 Related Work and Background

2.1 Task Background

Image Style Transfer. Image style transfer aims to render the content of a target image in the artistic style of a reference image and can be regarded as a specialized form of texture transfer. Early approaches relied on non-parametric texture synthesis(1), resampling pixels from a source texture while preserving structural features of the target image(2). However, these methods mainly used low-level features and struggled to separate semantic content from texture details.

With the rise of deep convolutional neural networks (CNNs), style transfer shifted toward learning feature representations that disentangle content and style(3). Neural style transfer algorithms formulated the process as matching deep feature statistics through optimization. Subsequent methods introduced feed-forward architectures, adaptive normalization, and feature transformation to enable arbitrary style transfer more efficiently(5). Recent advances further explore patch-based attention and transformer-based architecture(6) to improve content preservation and stylization control(7)(8).

Despite these developments, most style transfer methods still primarily address color tones and texture patterns, while complex artistic elements such as object shapes and decorative structures remain challenging to capture.

Text-Guided Style Transfer. Text-guided style transfer manipulates or generates images based on textual descriptions of the desired style. Compared to using style images directly, textual prompts offer a more abstract and flexible way to express aesthetic preferences. Early methods, such as

CLIPDraw(9) and VQGAN-CLIP(10), synthesize artistic images from text using vision-language models like CLIP. Diffusion-based models further improve quality by generating images in latent spaces conditioned on text.

Later approaches introduced combined image and text prompts to enhance stylization control. Methods like CLIPstyler(11) and StyleGAN-NADA(12) use CLIP-guided losses or adaptive tuning to better align the generated outputs with text descriptions. Recent works such as LDATA(13) further refine text-guided style transfer by strengthening the correlation between outputs and text prompts, enabling flexible stylization without retraining for each style.

Despite these advances, capturing complex artistic characteristics solely through text remains challenging. Our work addresses this gap by extracting style information directly from reference images, achieving more precise and controllable stylization guided by multimodal prompts.

2.2 Technical Foundations

Diffusion Model-based Methods. Diffusion models learn data synthesis through iterative noise reversal, with architectures like DDPMs surpassing traditional generative models (VAEs, GANs, autoregressive) in output quality across image/text generation tasks. In style transfer, their inversion techniques enable latent space manipulation but face content distortion and computational inefficiency. Example-guided approaches employ textual inversion (InST, StyleDiffusion) or internal component adjustment (DiffStyle) to separate style/content, though semantic preservation remains challenging. Recent advancements directly modulate diffusion models’ attention mechanisms for precise, optimization-free style transfer control while maintaining structural integrity.

Pre-trained large-scale Model-based Methods. Pre-trained large-scale diffusion models (e.g., Stable Diffusion) revolutionize text-conditioned image synthesis and style transfer through extensive dataset training. Existing style transfer methods using these models often rely on style-content disentanglement or textual inversion techniques, yet require per-style optimization/fine-tuning for latent space mapping, hindering scalability. While training-free approaches like DiffStyle exist, their incompatibility with latent diffusion architectures limits large-model utilization. Addressing these limitations, we propose an attention modulation mechanism that directly manipulates diffusion models’ spatial-aware attention layers to achieve optimization-free style transfer, maintaining compatibility with pre-trained architectures while ensuring efficient control over stylization outcomes.

3 Methodology

3.1 Model Overview

Our proposed framework consists of two core modules: **Style Injection and Blending** and **Prompt-Guided Stylization**. The former enables visual texture control via attention-based style injection from an image reference, while the latter supports high-level semantic editing via natural language prompts. Both components operate in a fully training-free manner by modifying internal features of a pre-trained diffusion model at inference time.

3.1.1 Style Injection and Blending

Our approach aims to achieve artistic style transfer by leveraging the generative capabilities of a pre-trained large-scale text-to-image diffusion model. Given a content image I^c and a reference style image I^s , the objective is to generate a stylized output I^{cs} that preserves the structural and semantic details of I^c while adapting the texture and visual appearance from I^s .

Rather than retraining or fine-tuning the model, we perform training-free style transfer by directly manipulating the internal self-attention mechanisms of the diffusion model during the reverse denoising process. Specifically, we modify the attention features, including query, key, and value representations, to inject style characteristics into the content latent. This feature-space intervention enables flexible and efficient style transfer without altering the model’s parameters. The overall framework of our style injection mechanism is illustrated in Figure 1.

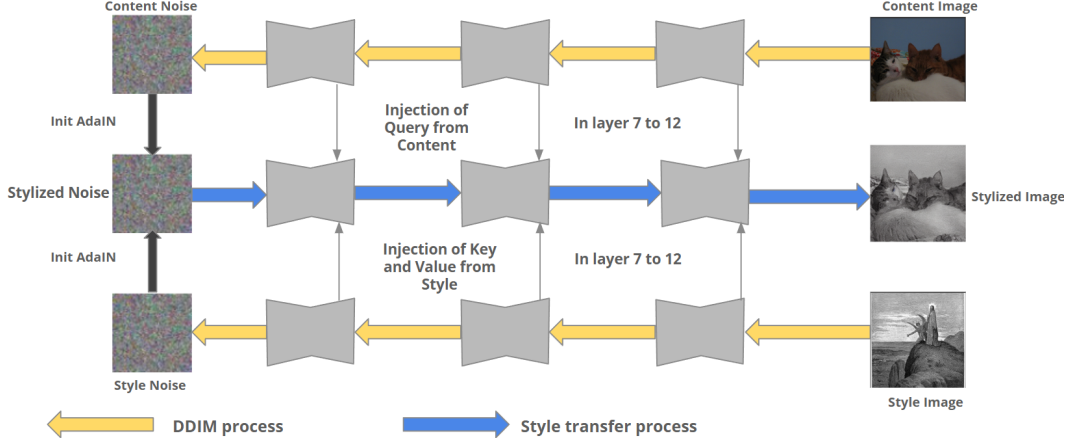


Figure 1: Overview of the Style Injection Framework. Key and value features from the style image are injected into the self-attention layers of the diffusion model, while content queries are preserved. Injection is selectively applied to the middle and high-level layers (layers 7 to 12) during the reverse diffusion process.

Feature Map Shapes and Parameters. Throughout the style transfer process, we operate on the latent space of the diffusion model, where the latent feature maps have a shape of $[4, 64, 64]$ for a 512×512 resolution input, corresponding to 4 channels at a $\times 8$ downsampling factor. The self-attention modifications occur at the output blocks of the U-Net decoder, where each attention map has a spatial resolution of 64×64 and a hidden dimension of 320, 640, or 1280 depending on the block.

Importantly, our method does not introduce any new learnable parameters. All modifications are applied directly to the internal features of the pre-trained Stable Diffusion U-Net, which has approximately 859M parameters.

Building upon the overall framework, we implement style transfer by manipulating the internal self-attention features of the diffusion model. Before style injection, we first modulate the initial latent state using **Adaptive Instance Normalization (AdaIN)** to enhance the adaptation of global color tones from the style image. While attention-based feature injection successfully transfers local textures and patterns, it often struggles to fully transfer the overall color distribution. We observe that the initial latent, sampled from a zero-mean, unit-variance Gaussian, tends to bias the generation toward a median color space, limiting color stylization.

To address this, instead of directly using the style latent z_T^s , which would distort the content structure, we apply AdaIN to the content latent z_T^c , aligning its channel-wise mean and variance with those of the style latent:

$$z_T^{cs} = \sigma(z_T^s) \left(\frac{z_T^c - \mu(z_T^c)}{\sigma(z_T^c)} \right) + \mu(z_T^s)$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation operations across channels. This produces a modulated latent z_T^{cs} that preserves the structural information of the content while inheriting the global color and illumination properties from the style image.

After the AdaIN modulation, we perform style transfer through two core mechanisms: **Style Injection** and **Style Blending**.

Style Injection. During the reverse diffusion process, we inject the style information into the content latent by replacing the key and value features of the self-attention layers with those extracted from the style image, while preserving the content queries. Specifically, the injection is applied only to the middle and high-level layers, corresponding to layers 7 to 12 in the U-Net decoder. This

enables the model to borrow local textures and patterns from the style image without disrupting the spatial structure of the content.

However, the mismatch between content queries and style keys can weaken the attention distribution, making the stylized outputs blurry. To address this, we apply **Attention Temperature Scaling**, where the query-key dot product is multiplied by a scaling factor $\tau > 1$ before softmax normalization:

$$\Phi_t^{style} = \text{softmax} \left(\frac{\tau Q_t K_t^T}{\sqrt{d}} \right) V_t$$

where d is the feature dimensionality. This sharpening operation helps maintain strong attention focus and better preserves the content’s structure.

Style Blending. To enable flexible multi-style transfer, we further extend our approach by introducing style blending at the self-attention feature level. Given two reference style images, we extract their respective key and value features, denoted as $K_t^{(1)}, V_t^{(1)}$ and $K_t^{(2)}, V_t^{(2)}$, and perform a linear interpolation controlled by a style blending coefficient α :

$$\hat{K}_t = \alpha K_t^{(1)} + (1 - \alpha) K_t^{(2)}, \quad \hat{V}_t = \alpha V_t^{(1)} + (1 - \alpha) V_t^{(2)}$$

Similarly, to balance between preserving content structure and allowing stylistic adaptation, we interpolate between the content query $Q_t^{content}$ and the fused style query Q_t^{style} , controlled by a parameter γ :

$$\hat{Q}_t = \gamma Q_t^{content} + (1 - \gamma) Q_t^{style}$$

The final stylized feature map is then computed via a sharpened self-attention operation:

$$\Phi_t^{style} = \text{softmax} \left(\frac{\tau \hat{Q}_t \hat{K}_t^T}{\sqrt{d}} \right) \hat{V}_t$$

where τ is the attention temperature scaling factor to enhance attention sharpness.

This mechanism allows continuous control over the influence of each style source and the strength of content preservation, providing a smooth transition between different artistic styles.

The overall style blending workflow is illustrated in Figure 2.

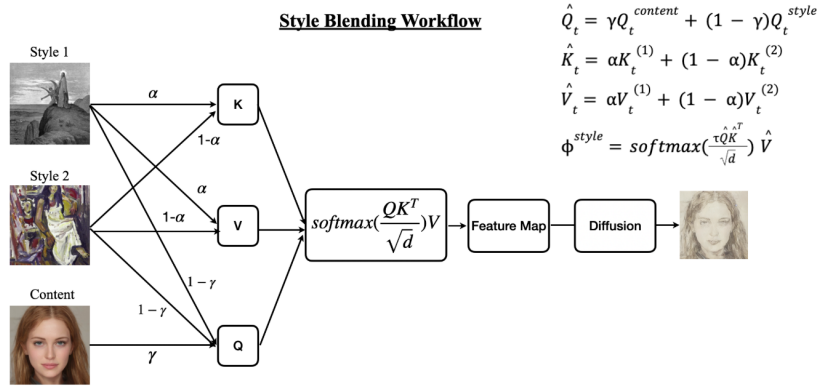


Figure 2: Style Blending Workflow. Key and value features from two styles are interpolated with blending factor α , and the query is interpolated with factor γ between content and style queries. The resulting feature map guides the diffusion process to generate the final stylized image.

3.1.2 Prompt-Guided Stylization

To support fine-grained stylization control conditioned on rich natural language descriptions, we design a two-stage diffusion framework as illustrated in Figure 3. This process enables detailed editing of texture, color, lighting, and artistic rendering based on user prompts, while preserving the structural content of the input image.

In the first stage, we perform *free-guided forward diffusion* on the input image $\mathbf{x}_0$ for T_1 steps, without any text prompt. At each step t , we apply the forward process defined as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}),$$

where $\bar{\alpha}_t$ is the cumulative noise level up to time t . This produces a latent representation $\hat{\mathbf{x}}_T$, termed *learnable noise*, which encapsulates the core structural features of the original image in the diffusion space. The use of a zero embedding during this stage ensures that no semantic style is introduced prematurely. Formally, the noise predictor is conditioned as:

$$\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t \mid \emptyset).$$

In the second stage, the system initiates *text-guided reverse denoising*, using the learnable noise $\hat{\mathbf{x}}_T$ as the starting point. A user-provided text prompt $\mathcal{T}$ is fed into a pre-trained CLIP encoder (ViT-B/16) to provide semantic supervision through multiple losses: *instruction loss* for alignment with textual intent, *content loss* to preserve original details, and *aesthetic loss* to encourage stylistic fidelity.

During each reverse step, we estimate the denoising direction using a linear interpolation of two specialized U-Nets:

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t \mid \mathcal{T}) = w \cdot \boldsymbol{\epsilon}_{\theta_1}(\mathbf{x}_t, t \mid \mathcal{T}) + (1 - w) \cdot \boldsymbol{\epsilon}_{\theta_2}(\mathbf{x}_t, t \mid \mathcal{T}),$$

where $\boldsymbol{\epsilon}_{\theta_1}$ and $\boldsymbol{\epsilon}_{\theta_2}$ are trained on realistic and artistic domains, respectively. The predicted noise is then used to compute the posterior mean:

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t \mid \mathcal{T}) \right).$$

The final sample $\mathbf{x}'_0$ is obtained after T denoising steps, supervised by a weighted sum of guidance objectives:

$$\mathcal{L}_{\text{total}} = \lambda_d \mathcal{L}_{\text{inst}} + \lambda_{c1} \mathcal{L}_c + \lambda_{c2} \mathcal{L}_{\text{cpatch}} + \lambda_{\text{aes}} \mathcal{L}_{\text{aes}} + \lambda_{\text{tv}} \mathcal{L}_{\text{tv}}.$$

This reverse process iteratively refines the output, ultimately generating a stylized image $\mathbf{x}'_0$ that reflects the prompt while maintaining fidelity to the original structure.

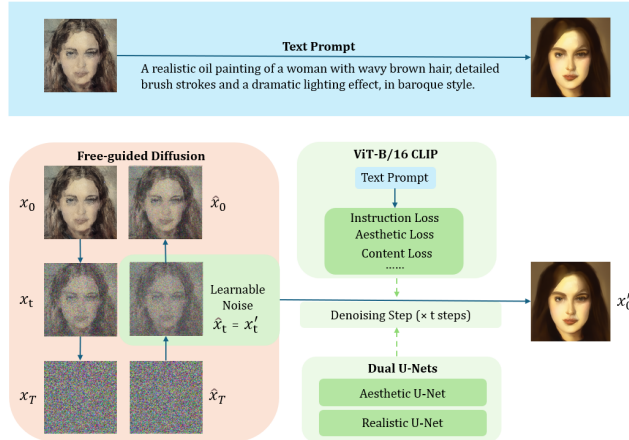


Figure 3: Overview of the prompt editing workflow. The framework consists of two stages: a free-guided forward diffusion stage that extracts learnable noise $\hat{\mathbf{x}}_T$ from the input image $\mathbf{x}_0$, and a text-guided reverse denoising stage that stylizes the image according to the input prompt using CLIP supervision and dual U-Nets.

3.2 Dataset

Our method is entirely training-free and relies solely on pre-trained models such as Stable Diffusion and CLIP. Therefore, no part of our dataset is used for model training or fine-tuning; instead, the dataset is exclusively employed for evaluation purposes.

We create a collection of style and content images to comprehensively assess the effectiveness of our framework. The style images are sourced from WikiArt and encompass diverse painting styles and artistic genres. The content images consist of real-world photographs depicting landscapes, architecture, and other natural scenes, obtained from open-source datasets and online resources. In total, our dataset includes approximately 200 style images and 500 content images, which are evaluated in various combinations to test both single-style and multi-style transfer capabilities.

To ensure compatibility with the latent diffusion backbone, all images are resized to a resolution of 512×512 . For the prompt-guided stylization module, we also compute CLIP embeddings from corresponding textual descriptions of style images, enabling semantic conditioning during the reverse diffusion stage.

3.3 Evaluation Metric

ArtFID: A metric specifically designed for style transfer evaluation. It evaluates both how well the style is transferred and how much content is preserved.

$$ArtFID = (1 + LPIPS) \times (1 + FID) \quad (1)$$

$$FID = \|\mu_r - \mu_g\|_2^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}) \quad (2)$$

LPIPS: Learned Perceptual Image Patch Similarity. It is a perceptual similarity based on deep neural network, and it measures perceptual differences between two images based on their feature representations rather than raw pixels. Lower LPIPS means higher content preservation.

$$LPIPS = \sum_i d(f_i(x), f_i(y)) \quad (3)$$

where $f_i(x)$ represents deep feature activations from a neural network.

CFSD: Content Feature Structural Distance. It is a newer metric than LPIPS, and it is only focus on structure preservation.

$$CFSD = \frac{1}{hw} \sum_{i=1}^{hw} KL(S_i^c || S_i^{cs}) \quad (4)$$

where S_i^c and S_i^{cs} are the probability maps of the content image and stylized image.

CLIP score: A cross-modal similarity score between an image and a text prompt, using the CLIP model. It measures how well an image matches a given text description. The higher the better.

$$CLIPScore(I, T) = 100 \times \cos(f_I(I), f_T(T)) = 100 \times \frac{f_I(I) \times f_T(T)}{\|f_I(I)\| \times \|f_T(T)\|} \quad (5)$$

Where I is the image, T is the text, and f is the output unit vector (512). The equation computes the cosine similarity between the image and the text embedding.

4 Baseline and Extensions

4.1 Baseline selection & Evaluation

For the Style Injection and Blending task, we select AdaAttn and DiffStyle as baselines. AdaAttn is a representative attention-based style transfer method that enables style injection without requiring

per-style retraining, making it a strong baseline for training-free comparisons. It introduces adaptive attention normalization, which computes attention maps between content and style features and transfers localized mean and variance statistics in a spatially adaptive manner. This mechanism allows fine-grained control of style patterns while preserving the structural layout of the content image. DiffStyle, on the other hand, is a diffusion-based approach that manipulates style representations during the denoising process and represents the state-of-the-art in training-free diffusion stylization. It enables high-quality style transfer by injecting style semantics at each diffusion step.

For the Prompt-Guided Stylization task, we compare against LD-AST, which performs language-driven style editing by aligning visual features to text embeddings, Stable Diffusion, a general-purpose generative model that supports prompt-controlled image synthesis, and DiffusionCLIP, which modifies image semantics via CLIP-guided gradients. These methods cover both traditional neural editing paradigms and recent advances in diffusion-based models, providing a comprehensive and diverse basis for evaluating the effectiveness of our prompt-guided style editing framework.

4.2 Implemented Extensions

We implement two core modules in our framework: Style Injection and Blending, and Prompt-Guided Stylization, each designed to evaluate the flexibility and robustness of our method under different style manipulation scenarios. The Style Injection module allows direct transfer of visual patterns from reference style images into the content image. Style Blending extends this by enabling smooth interpolation between multiple styles, showcasing the model’s capability to synthesize complex stylistic combinations. Prompt-Guided Stylization, meanwhile, focuses on semantic-level control, where image attributes are modified according to textual instructions.

All modules are implemented in a unified, training-free architecture, allowing consistent comparison across settings. We conduct comprehensive evaluations of each module using multiple baseline methods and benchmark metrics. The final results, along with both quantitative comparisons and qualitative visualizations, are presented in the following section to demonstrate the effectiveness of our approach across various style editing tasks.

5 Results and Analysis

5.1 Results

As our approach is entirely training-free, this project does not involve conventional training and validation procedures. Instead, we evaluate our framework through two core experiments: (1) style injection and blending, and (2) prompt-guided stylization. For each task, we provide both qualitative and quantitative comparisons against strong baselines, using multiple benchmark metrics to assess style fidelity, content preservation, and semantic alignment.

Figure 4 provides a visual demonstration of the two core tasks in our framework: style injection/blending (left) and prompt-guided stylization (right). The examples highlight how our method successfully transfers stylistic patterns or interprets textual prompts to produce semantically consistent and visually coherent outputs. Meanwhile, Table 1, Table 2, and Table 3 present the corresponding quantitative results, evaluating each module as well as the integrated pipeline across multiple performance metrics.

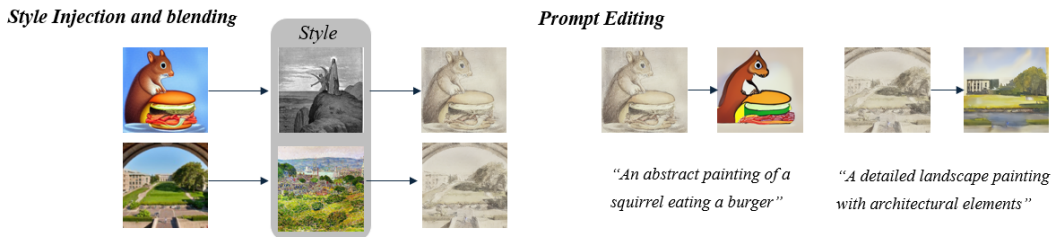


Figure 4: Visual Demonstration of Style Injection, Blending and Prompt-Guided Stylization

	Ours	AdaAttn(14)	DiffStyle(15)
ArtFID ↓	36.212	30.350	41.464
LPIPS ↓	0.5910	0.5439	0.8931
CFSD ↓	0.1637	0.2862	0.2819

Table 1: Style Injection and blending

	Ours	LDAST(16)	Stable diffusion(17)	DiffusionCLIP(18)
CLIP score ↑	0.2631	0.2235	0.2773	0.2389
LPIPS ↓	0.608	0.744	0.637	0.619

Table 2: Prompt-Guided Stylization

	Ours	AdaAttn(14)	DiffStyle(15)
ArtFID ↓	34.133	30.350	41.464
LPIPS ↓	0.5734	0.5439	0.8931
CFSD ↓	0.1512	0.2862	0.2819

Table 3: Final Integrated Model

As shown in **Table 1**, our method achieves a strong balance between stylization quality and structural preservation. While AdaAttn attains the lowest ArtFID score (30.350), suggesting slightly higher stylization fidelity, our approach significantly outperforms both AdaAttn and DiffStyle in terms of CFSD (0.1637 vs. 0.2862 and 0.2819), demonstrating superior spatial structure retention. Our LPIPS score (0.5910) is also competitive, confirming our model’s effectiveness in preserving perceptual similarity to the content image.

These findings highlight the strength of our self-attention-based style injection mechanism, which transfers high-frequency texture details without disrupting the original semantic layout. Furthermore, our support for style blending via Key/Value interpolation enables smooth composition of multiple styles while minimizing structure loss.

In **Table 2**, our method obtains the lowest LPIPS score (0.608), indicating the best preservation of content across all baselines during prompt-guided stylization. Although Stable Diffusion achieves a slightly higher CLIP score (0.2773), our result (0.2631) remains competitive and outperforms both LD-AST (0.2235) and DiffusionCLIP (0.2389), showing our ability to interpret prompts while maintaining visual and semantic coherence.

Table 3 summarizes the performance of our full pipeline, which combines both style injection and prompt-guided stylization. Our method further improves structural alignment with the best CFSD score (0.1512), and achieves a strong LPIPS score (0.5734), approaching AdaAttn (0.5439), confirming robust perceptual content preservation in the joint setting.

Overall, these results show that our dual-branch design, combining a realism-preserving path and a prompt-guided editing path, can effectively adjust image semantics while keeping both structure and content visually consistent. While our model does not outperform AdaAttn in terms of ArtFID and LPIPS, this is understandable since AdaAttn is specifically trained to enhance style intensity and perceptual similarity. In contrast, our method requires no training and focuses on flexible, easy-to-control, and structure-preserving editing, making it more practical for real-world applications where retraining is not feasible.

6 Future Directions

Fine-grained control via prompts. A key direction for future improvement is to enable more fine-grained and semantically rich control over the stylization process through prompts. Currently, prompts tend to operate at a global level, affecting the image holistically. By supporting more structured and expressive prompts, such as combining spatial references ("sky," "foreground") with stylistic adjectives ("surreal," "watercolor," "cubist"), users could specify different styles for different regions or elements within the same image. Additionally, parsing complex natural language descriptions into

multi-level visual transformations would enable controllable manipulation of both low-level textures (e.g., brushstroke detail) and high-level composition (e.g., lighting, perspective), opening the door to truly personalized and interpretable style transfer.

Extension to video and real-time applications. While our method performs well on still images, extending it to videos presents additional challenges and opportunities. A major concern is ensuring temporal consistency, applying style transformations across frames without introducing flickering or spatial drift. Future work could explore frame-to-frame feature alignment or temporal attention mechanisms to address this. Moreover, optimizing the model’s inference efficiency would make it viable for real-time applications, including AR filters, video conferencing effects, mobile content creation, and immersive artistic experiences. Real-time feedback and streaming-based control (e.g., prompt editing during playback) would further enhance the interactivity of the system.

Other training-free architectures. Although our framework is built on latent diffusion models, there is growing interest in other training-free or few-shot generative architectures that may complement or outperform diffusion in specific scenarios. For example, autoregressive models (e.g., masked transformers) may offer better sequence-level control, VAEs provide faster generation with structured latent spaces, GANs can deliver high-resolution results with low latency, and flow-based models ensure exact likelihood estimation. Exploring these architectures, individually or in hybrid designs, could yield better trade-offs between visual quality, speed, and controllability. This line of work would further enrich the space of plug-and-play stylization tools with minimal training or supervision cost.

7 Conclusion

Method Innovation. In this work, we propose a novel training-free framework that unifies two complementary mechanisms, style injection and prompt-guided stylization, to achieve controllable, dual-level style transfer. Our style injection module leverages self-attention mechanisms to apply high-frequency textures in a spatially coherent manner, while our prompt editing module enables semantic-level manipulation using natural language instructions. This dual-branch design allows users to precisely control both the visual style and the underlying content semantics without requiring any per-style training or prompt-specific tuning.

Metric Superiority. Extensive quantitative evaluations demonstrate that our method achieves strong performance across multiple metrics. We outperform or match state-of-the-art baselines in content preservation, as measured by LPIPS, and maintain competitive CLIP scores that reflect accurate semantic alignment with user-provided prompts. Our framework also attains the best CFSD scores in all evaluated settings, indicating superior structural consistency. These results confirm that our approach successfully balances stylization fidelity with perceptual and structural integrity.

Efficient and Training-Free. A key advantage of our method is its fully inference-time, plug-and-play nature. Unlike traditional approaches that require task-specific training or fine-tuning, our framework can operate directly on pre-trained models such as Stable Diffusion and CLIP without additional optimization. This enables rapid deployment in real-world applications and supports flexible use cases such as multi-style blending, prompt-guided edits, and future extensions to video or real-time scenarios. By eliminating the need for training, we reduce computational costs while enhancing generalizability and usability.

8 Team Contribution

Yinuo Wang:

- Conducted extensive research on related work and literature review.
- Proposed the initial idea of Style-Injection and Prompt-Editing.
- Designed high-level architecture of entire model and workflow.
- Implemented prompt-editing module and finished the final integration of the system.

Xinyue Meng:

- Implemented the prompt-editing module and conducted evaluations over different datasets.
- Helped integrate all modules into the final model and debugged the integration.
- Conducted testing and evaluations of the final integrated model on different datasets.

Yixiang Wang:

- Investigated the Style Injection method and implemented the method.
- Conduct baseline implementation for Style-Injection method.
- Implemented style blending method and conducted evaluations.

Xianfei Yang:

- Conducted baseline implementation for Style-Injection method.
- Conducted experiments and evaluations for different datasets.
- Tested different types of style transfer using the Style-Blending method.

References

- [1] A. A. Efros and T. K. Leung, *Texture synthesis by non-parametric sampling*, in *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, vol. 2, pp. 1033–1038, 1999. DOI: 10.1109/ICCV.1999.790383.
- [2] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, *Image analogies*, in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*, pp. 327–340, ACM, New York, NY, USA, 2001. DOI: 10.1145/383259.383295.
- [3] L. A. Gatys, A. S. Ecker, and M. Bethge, *Image Style Transfer Using Convolutional Neural Networks*, in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016. DOI: 10.1109/CVPR.2016.265.
- [4] N. Huang, Y. Zhang, F. Tang, C. Ma, H. Huang, Y. Zhang, W. Dong, and C. Xu, *DiffStyler: Controllable Dual Diffusion for Text-Driven Image Stylization*, arXiv preprint arXiv:2211.10682, 2022.
- [5] X. Huang and S. Belongie, *Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization*, arXiv preprint arXiv:1703.06868, 2017.
- [6] Y. Deng, F. Tang, W. Dong, C. Ma, X. Pan, L. Wang, and C. Xu, *StyTr²: Image Style Transfer with Transformers*, arXiv preprint arXiv:2105.14576, 2022.
- [7] M. Zhu, X. He, N. Wang, X. Wang, and X. Gao, *All-to-key Attention for Arbitrary Style Transfer*, arXiv preprint arXiv:2212.04105, 2023.
- [8] C. Zhang, X. Xu, L. Wang, Z. Dai, and J. Yang, *S2WAT: Image Style Transfer via Hierarchical Vision Transformer using Strips Window Attention*, arXiv preprint arXiv:2210.12381, 2023.
- [9] K. Frans, L. B. Soros, and O. Witkowski, *CLIPDraw: Exploring Text-to-Drawing Synthesis through Language-Image Encoders*, arXiv preprint arXiv:2106.14843, 2021.
- [10] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, and E. Raff, *VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance*, arXiv preprint arXiv:2204.08583, 2022.
- [11] G. Kwon and J. C. Ye, *CLIPstyler: Image Style Transfer with a Single Text Condition*, arXiv preprint arXiv:2112.00374, 2022.
- [12] R. Gal, O. Patashnik, H. Maron, G. Chechik, and D. Cohen-Or, *StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators*, arXiv preprint arXiv:2108.00946, 2021.

- [13] T.-J. Fu, X. E. Wang, and W. Y. Wang, *Language-Driven Image Style Transfer*, arXiv preprint arXiv:2106.00178, 2022.
- [14] S. Liu, T. Lin, D. He, F. Li, M. Wang, X. Li, Z. Sun, Q. Li, and E. Ding, *AdaAttN: Revisit Attention Mechanism in Arbitrary Neural Style Transfer*, arXiv preprint arXiv:2108.03647, 2021. [Accepted by ICCV 2021].
- [15] J. Jeong, M. Kwon, and Y. Uh, *Training-free Content Injection using h-space in Diffusion Models*, arXiv preprint arXiv:2303.15403, 2023.
- [16] T.-J. Fu, X. E. Wang, and W. Y. Wang, *Language-Driven Image Style Transfer*, arXiv preprint arXiv:2106.00178, 2022. [ECCV 2022].
- [17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, *High-Resolution Image Synthesis with Latent Diffusion Models*, arXiv preprint arXiv:2112.10752, 2022. [CVPR 2022].
- [18] G. Kim, T. Kwon, and J. C. Ye, *DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation*, arXiv preprint arXiv:2110.02711, 2021. [Accepted to CVPR 2022].
- [19] Chung, J., Hyun, S., & Heo, J.-P. (2023). Style Injection in Diffusion: A Training-free Approach for Adapting Large-scale Diffusion Models for Style Transfer. *arXiv preprint arXiv:2312.09008*.
- [20] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. *Inversion-Based Style Transfer with Diffusion Models*. arXiv preprint arXiv:2211.13203, 2023.
- [21] Gihyun Kwon and Jong Chul Ye. *CLIPstyler: Image Style Transfer with a Single Text Condition*. arXiv preprint arXiv:2112.00374, 2022.
- [22] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. *Neural Style Transfer: A Review*. arXiv preprint arXiv:1705.04058, 2018.
- [23] Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.
- [24] Karras, T., Laine, S., & Aila, T. (2018). A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*.
- [25] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*.
- [26] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*.
- [27] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *arXiv preprint arXiv:1706.08500*.
- [28] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600-612.
- [29] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *arXiv preprint arXiv:1801.03924*.
- [30] Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.