Exploring The Practical Applications of Text-Guided Image Stylization: A Case Study of Diffstyler

Yuxuan Xie

School of Electrical Engineering and Artificial Intelligence, Xiamen University Malaysia, Sepang, Selangor, Malaysia

Abstract. Traditional image style transfer methods typically rely on fixed reference images, limiting user creativity and often resulting in the loss of fine image details. Although text-driven techniques have significantly improved the flexibility of style control, striking a balance between expressive style transformation and content fidelity remains a central challenge. This study explores the practical application of text-guided image stylization by revisiting and evaluating DiffStyler, a representative framework that integrates a dual-diffusion architecture with a learnable noise generation mechanism. Through this approach, this paper demonstrates the effectiveness of text-driven style transfer in both digital art creation and intelligent image processing. The model enables precise translation of textual style descriptions into visual modifications while preserving the structural integrity of the original image. Experimental results show that DiffStyler achieves a favorable trade-off between stylization quality and content preservation, particularly in artistic image translation scenarios. This work not only showcases the engineering potential of cross-modal generation techniques but also establishes a feasible pathway for extending such technologies beyond artistic domains.

1 Introduction

Image style transfer has attracted significant attention due to its wide applicability in digital art, film post-production, and game design. Early approaches primarily relied on low-level feature matching, but recent advances have leveraged deep neural networks to extract and recombine high-level semantic representations. For example, Gatys et al. pioneered a neural framework that separates and recomposes content and style features in the deep feature space, moving beyond traditional pixel-based operations and enabling more sophisticated style transformations [1]. Building on this, Liu et al. introduced AdaAttN, which adaptively modulates style features through joint learning of shallow and deep spatial attention maps [2]. In parallel, An et al. proposed ArtFlow, incorporating invertible neural flows to mitigate content leakage and enhance stylization fidelity [3].

Despite these advancements, most existing methods depend on reference style images, which constrain user creativity and limit the expressiveness of style transfer. Text-guided

style transfer has emerged as a promising alternative, offering more flexible and intuitive control. For instance, Bai et al. proposed ITstyler, a real-time framework that maps CLIP-based textual features into a pre-trained VGG style space, achieving fast and high-quality stylization without iterative optimization. However, ITstyler struggles with capturing detailed semantic content, such as object-specific or character-level descriptions [4]. Similarly, Zhang et al. highlighted that existing text-based methods tend to provide only coarse-grained stylizations (e.g., by genre or material) while failing to accurately convey fine-grained semantic elements like object shapes or artistic details [5].

To address these challenges, DiffStyler introduces a dual-diffusion architecture combined with a learnable noise generation mechanism, enabling fine-grained, text-aligned stylization while preserving the structural integrity of the source image [6]. This study revisits and evaluates DiffStyler from a practical perspective, aiming to validate its effectiveness in real-world scenarios and to demonstrate its potential as a powerful tool in intelligent image processing and digital art creation.

The remainder of this paper is organized as follows: Chapter 1 presents the research background, motivation, and objectives. Chapter 2 reviews the relevant literature on image style transfer and text-guided stylization. Chapter 3 provides an in-depth explanation of DiffStyler's architecture and technical innovations. Chapter 4 conducts experimental evaluations and case studies to assess its performance. Chapter 5 concludes the study, summarizes key findings, and discusses future research directions.

2 Related Work

2.1 Style Transfer

Traditional style transfer has evolved from traditional pixel-based techniques to advanced neural frameworks. Early methods relied on statistical analysis and patch-based approaches, such as texture synthesis via image quilting or histogram matching. These techniques transferred style by stitching patches from the style image onto the content image, aiming to replicate textures and color palettes. However, they struggled with global coherence, complex style integration, and structural preservation, often producing artifacts like visible seams or distorted details.

A paradigm shift emerged with Convolutional Neural Networks (CNNs). Gatys et al. pioneered neural style transfer (NST) by repurposing CNNs trained for object recognition. These networks hierarchically separate content and style: deeper layers capture high-level content (object arrangement), while style is encoded via Gram matrices measuring feature correlations across layers. By optimizing an image to match both content and style representations, NST achieved unprecedented quality, blending abstract artistic styles with content fidelity. Unlike patch-based methods, CNNs inherently disentangled style and content, leveraging learned invariances to preserve structural integrity while transferring textures [1].

NST offered superior flexibility and perceptual quality. It abstracted complex patterns and color distributions, enabling nuanced stylization control. However, early NST faced limitations: generated images often exhibited artifacts, such as edge distortion or semantic mismatches where styles bled into unrelated regions (e.g., sky textures applied to buildings). These issues stemmed from global style application, neglecting scene semantics [7].

Semantic segmentation addressed these challenges by introducing region-aware processing. Networks like DeepLabv3+ labeled semantic regions (e.g., sky, trees) in content and style images, enabling targeted style transfer. Methods such as BS-NST (semantic segmentation-based NST) applied styles to corresponding semantic areas, preventing cross-

region contamination. This approach improved coherence in structured scenes (e.g., landscapes), yielding smoother results with fewer artifacts. Evaluations using metrics like SSIM and PSNR confirmed enhanced structural alignment and realism compared to non-semantic methods [8].

Despite progress, challenges persist. Segmentation-guided NST may still distort fine details (e.g., intricate textures) or require balancing content preservation and stylization intensity. Computational costs rise for high-resolution images or complex segmentation models. Achieving photorealism across all texture types remains elusive, necessitating further research into detail preservation and efficiency optimization.

In summary, image style transfer has transitioned from manual patch manipulation to AI-driven semantic frameworks. CNNs revolutionized the field by disentangling content and style, while semantic segmentation refined localization and realism. Though computational demands and fine-detail handling persist as hurdles, deep learning methods—particularly those integrating semantic context—represent the state of the art, offering unparalleled stylistic control and visual quality [1, 8]. Future advancements will likely focus on optimizing efficiency and bridging the gap between artistic abstraction and photorealistic fidelity.

2.2 Text-Driven Image Generation

Text-driven image generation leverages natural language descriptions to guide generative models, enabling cross-modal translation from text to image. In recent years, this paradigm has advanced rapidly with the development of models such as DALL·E and Stable Diffusion [9,10], which are primarily based on diffusion processes and generative adversarial networks (GANs). Compared to traditional style transfer methods that rely on reference images, text-driven approaches offer greater flexibility, allowing users to generate diverse and expressive images directly from textual prompts. This not only enhances creative freedom but also improves the richness and variety of style representation.

Early text-driven methods began to decouple style transfer from the necessity of explicit style reference images. CLIPstyler, for example, demonstrated the feasibility of modulating image style using only a single text condition by leveraging the powerful text-image embedding capabilities of CLIP. This approach utilized a patch-wise text-image matching loss combined with multiview augmentations to translate semantic descriptions into realistic textures, effectively enabling users to transfer styles based purely on textual imagination without needing a visual counterpart [11]. This marked a significant shift, addressing limitations where users might not possess a reference image for the desired aesthetic.

However, as text-driven methods matured, new challenges emerged, particularly concerning the ambiguity inherent in stylistic descriptions and the tendency of models to overfit to reference styles when available. Overfitting could lead to generated images overly mirroring the reference, potentially conflicting with textual instructions and limiting nuanced control over specific style elements like color or texture. Furthermore, maintaining consistent spatial layouts and avoiding artifacts became critical. StyleStudio addressed these issues by proposing complementary strategies: a cross-modal AdaIN mechanism to better integrate style and text features, a Style-based Classifier-Free Guidance (SCFG) technique for selective control over style elements, and the incorporation of a teacher model during generation to enhance layout stability [12,13]. These techniques aimed to provide finer control and improve alignment between the generated image, the text prompt, and the optional style reference, tackling overfitting and layout inconsistencies.

Alongside improvements in control and fidelity, the efficiency of text-driven stylization became a focal point. Initial text-guided methods often required substantial training iterations and computational resources for each stylization task. To address this, StyleMamba introduced a novel framework incorporating a conditional State Space Model (Mamba)

within an AutoEncoder system. This architecture, combined with proposed masked and second-order directional losses, aimed to sequentially align image features with text prompts more efficiently, significantly reducing training iterations and inference time compared to previous approaches. By leveraging models like SigLIP and pre-trained VAEs, StyleMamba demonstrated a path towards faster convergence and enhanced local and global style consistency, making complex, text-driven stylization more practical for real-world applications [14].

3 Core Innovations and Mechanisms of DiffStyler

3.1 DiffStyler System Architecture

DiffStyler employs a dual-diffusion model that seamlessly integrates traditional diffusion techniques with a text-driven module. The system begins by using free-guided diffusion to generate learnable noise closely related to the original image structure, effectively preserving image details. Next, a dual-channel U-Net model is utilized to separately handle content preservation and style rendering. These two components are then linearly combined during the sampling process to achieve the final stylized output.

In parallel, the pre-trained CLIP model is employed to convert textual descriptions into semantic embeddings, which guide the generation process. To further enhance the efficiency of the diffusion process, a pseudo-numerical method is applied, optimizing the diffusion steps and reducing the number of sampling iterations, thus improving the overall generation speed. A schematic representation of the architecture is shown in Figure 1 [6].

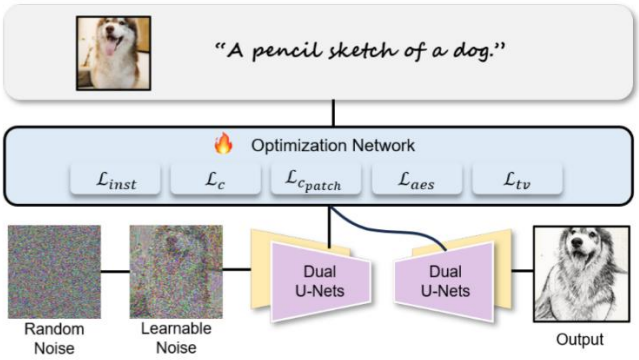

Figure 1. An illustration of a DiffStyler is showcased, where in the model generates a stylized image by leveraging a text-based description style and the content image as input [6].

3.2 Text-Guided Stylization with CLIP

To enable text-driven stylization, DiffStyler incorporates a pre-trained CLIP model to encode user-provided textual descriptions into semantic embeddings. These embeddings are integrated into the sampling process and fused with image features, ensuring that the generated outputs align closely with the intended textual guidance.

Furthermore, the system allows users to adjust the strength of text guidance, offering flexible control over the degree of stylistic influence. This design enhances the adaptability of the model across diverse artistic and application scenarios [6].

3.3 Learnable Noise Generation Mechanism

Unlike conventional approaches that initialize the diffusion process with purely random noise, DiffStyler introduces a learnable noise generation mechanism conditioned on the input image. This technique extracts structural cues from the original image through a multi-step, text-guided forward diffusion process, resulting in a noise map that encodes content-relevant features. The generated noise, which serves as the initial state for the reverse diffusion process, is thus highly correlated with the source image, as illustrated in Table 1 [6].

By preserving image-specific structural information in the initialization stage, this mechanism significantly enhances content fidelity and detail preservation during stylization. Moreover, it operates in conjunction with a multi-objective loss function designed to balance three key objectives: accurate style transfer, content consistency, and overall aesthetic quality. This synergy contributes to producing visually coherent and semantically meaningful stylized outputs, even for complex textual prompts.

Table 1. Learnable Noise Initialization for Guided Diffusion

Algorithm 1 Learnable noise generating, given a diffusion model $\left(\boldsymbol{\mu_\theta}(\mathbf{x_t}), \boldsymbol{\Sigma_\theta}(\mathbf{x_t})\right)$.

Require: The input image $\mathbf{x}_0$, free-guided noising steps T_1, denoising steps T.
Ensure: Learnable noise $\hat{\mathbf{x}}_{\boldsymbol{T}}$.

$\boldsymbol{t = T_1}$
$\mathbf{x}_{\boldsymbol{T_1}} \leftarrow$ sample from $\boldsymbol{\mathcal{N}(\mathbf{0}, \mathbf{I})}$
$\hat{\mathbf{x}}_{\boldsymbol{T_1}} = \mathbf{x}_{\boldsymbol{T_1}}$
repeat
  $\boldsymbol{t - 1 \leftarrow t}$
  $\boldsymbol{\mu_\theta(\hat{\mathbf{x}}_t) \leftarrow \epsilon_\theta(\hat{\mathbf{x}}_t, t \mid \emptyset)}$
  $\boldsymbol{\mu, \Sigma \leftarrow \mu_\theta(\hat{\mathbf{x}}_t), \Sigma_\theta(\hat{\mathbf{x}}_t)}$
  $\boldsymbol{\hat{\mathbf{x}}_{t-1} \sim \mathcal{N}(\mu, \Sigma)}$
until $\boldsymbol{t = T}$

3.4 Dual-Diffusion Collaborative Control Mechanism

To achieve effective decoupling of content preservation and style generation, DiffStyler introduces a dual-diffusion collaborative control mechanism. As illustrated in Figure 2, the system utilizes two parallel diffusion pathways: one dedicated to maintaining the structural integrity of the input image, and the other focused on synthesizing the desired artistic style. In Table 2, it can be observed that the linear fusion of the two pathways.

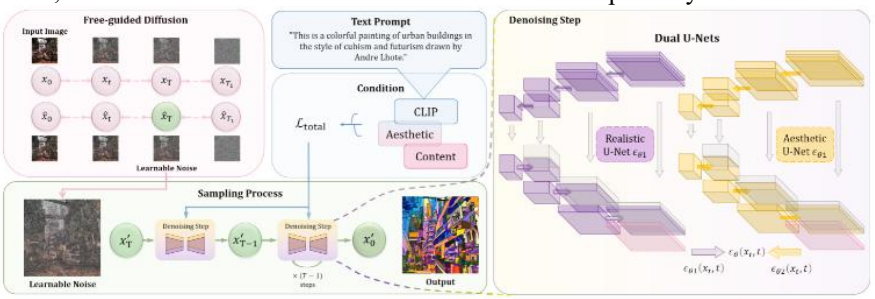

Figure 2. The free-guided diffusion framework employs dual U-Nets guided by text prompts and learnable noise to iteratively denoise and produce high-quality, multi-constrained stylized images. [6]

Table 2. Text-Guided Stylized Sampling via Learnable Diffusion

Algorithm 2 Text guided stylized diffusion sampling, given a diffusion model $\left(\boldsymbol{\mu_\theta}(\mathbf{x_t}), \boldsymbol{\Sigma_\theta}(\mathbf{x_t})\right)$.

Require: The input image $\mathbf{x}_0$, text prompt $\boldsymbol{\mathcal{T}}$, gradient scale w, free-guided noising steps T_1, denoising steps T.
Ensure: Stylized result $\mathbf{x}'_\mathbf{0}$ guided by $\boldsymbol{\mathcal{T}}$.

$\boldsymbol{t = T}$
$\hat{\mathbf{x}}_{\boldsymbol{T}} =$ learnable noise $(\mathbf{x}'_{\boldsymbol{T}})$
$\mathbf{x}'_{\boldsymbol{T}} = \hat{\mathbf{x}}_{\boldsymbol{T}}$
for t = T, ... , 1 **do**
  $\boldsymbol{\mu, \Sigma} \leftarrow \boldsymbol{\mu_\theta}(\mathbf{x}'_\mathbf{t}), \boldsymbol{\Sigma_\theta}(\mathbf{x}'_\mathbf{t})$
  $\tilde{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{x}'_\mathbf{t}, \boldsymbol{t} \mid \boldsymbol{\mathcal{T}}) \leftarrow \boldsymbol{w} \cdot \boldsymbol{\epsilon}_{\boldsymbol{\theta_1}}(\mathbf{x}'_{\mathbf{t-1}}, \boldsymbol{t} \mid \boldsymbol{\mathcal{T}}) +$
    $(\mathbf{1} - \boldsymbol{w}) \cdot \boldsymbol{\epsilon}_{\boldsymbol{\theta_2}}(\mathbf{x}'_\mathbf{t}, \boldsymbol{t} \mid \boldsymbol{\mathcal{T}})$

  $\boldsymbol{\mu_\theta}(\mathbf{x}'_\mathbf{t}) \leftarrow \tilde{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{x}'_\mathbf{t}, \boldsymbol{t} \mid \boldsymbol{\mathcal{T}})$
  $\boldsymbol{\mathcal{L}_{total}} \leftarrow \boldsymbol{\lambda_d \mathcal{L}_{inst}} + \boldsymbol{\lambda_{c1} \mathcal{L}_c} + \boldsymbol{\lambda_{c2} \mathcal{L}_{c_{patch}}} +$
    $\boldsymbol{\lambda_{aes} \mathcal{L}_{aes}} + \boldsymbol{\lambda_{tv} \mathcal{L}_{tv}}$
  $\mathbf{x}'_{\mathbf{t-1}} \leftarrow$ sample from $\boldsymbol{\mathcal{N}(\mu_\theta(\mathbf{x}'_\mathbf{t}) + \Sigma \mathcal{L}_{total}, \Sigma)}$
end for
return $\mathbf{x}'_\mathbf{0}$

During each step of the reverse diffusion process, the outputs from the two channels are linearly fused using a set of pre-defined weights. This fusion strategy allows the model to simultaneously preserve photorealistic details and render stylistic attributes. Importantly, the relative weights of the content and style channels can be dynamically adjusted, enabling users to fine-tune the output based on specific aesthetic or semantic requirements.

By leveraging this cooperative mechanism, DiffStyler enhances the controllability and flexibility of text-guided style transfer, allowing for a more personalized and balanced stylization process across diverse application scenarios.

4 Experiments

4.1 Experimental Implementation

To evaluate the practical performance of the DiffStyler model, a two-stage experimental design was developed based on the resources publicly available on GitHub, which include the core model and an art-style generator trained on the WikiArt dataset.

In the first stage, the experiment investigates the model's ability to handle different categories of style prompts. Specifically, it compares the stylized outputs generated from the same input image under artistic versus non-artistic textual prompts. This comparison provides an initial assessment of DiffStyler's style transfer adaptability across varied stylistic domains.

Based on prior analysis of the model's training data, the experiment hypothesize that DiffStyler will exhibit stronger performance in tasks involving artistic styles, while its effectiveness may decline when confronted with non-artistic descriptions. To further validate the model's performance in artistic stylization, the second stage focuses on mainstream artistic style transfer tasks. Key control variables, such as the freedom scale parameter (-fs) and the WikiArt-specific scale parameter (-ws), are systematically adjusted under otherwise

fixed conditions to generate a variety of style-oriented outputs. This enables us to examine how each parameter influences the model's ability to express stylistic features.

For evaluation, the experiment adopts a hybrid assessment framework. Quantitative metrics include CLIP-based similarity scores and the Learned Perceptual Image Patch Similarity (LPIPS), measuring both stylistic accuracy and content preservation. These objective indicators are complemented by subjective visual assessments, ensuring a holistic understanding of DiffStyler's performance in balancing artistic fidelity with structural coherence across multiple style transfer scenarios.

4.2 Evaluation Metrics

To comprehensively evaluate the style transfer performance of DiffStyler across various style types and image complexities, the paper adopts a set of quantitative metrics focused on both style fidelity and content preservation.

For style alignment, the metrics employ the CLIP-Style Score, which measures the cosine similarity between the CLIP embeddings of the generated image and the corresponding style prompt. A higher score indicates stronger alignment between the stylized output and the intended textual style description.

To assess content preservation, the metrics use the LPIPS-Content Score, which calculates the perceptual distance between the generated image and the original input using the LPIPS (Learned Perceptual Image Patch Similarity) metric. A lower LPIPS score reflects better retention of image structure and semantic content.

To balance these two aspects, the experiment introduces a composite evaluation metric, termed the Style-Content Balance Index (SCB Index). The SCB Index integrates the normalized CLIP-Style Score and LPIPS-Content Score as follows:

$$\mathrm{SCB-Index} = \alpha \cdot S_{\mathrm{norm}} + (1-\alpha) \cdot (1 - C_{\mathrm{norm}}) \tag{1}$$

$$S_{\mathrm{norm}} = \frac{clipscore + 1}{2} \tag{2}$$

where S_{norm} denotes the normalized CLIP-Style Score, and C_{norm} is the normalized LPIPS-Content Score, representing content distortion. The term $1 - C_{\mathrm{norm}}$ is used to reflect content preservation, with higher values indicating less distortion. The weight coefficient α serves as a dynamic balance factor that adjusts the emphasis between style fidelity and content retention depending on the task context.

Based on empirical observations and the practical objectives of DiffStyler, the balance factor was set $\alpha = 0.6$, reflecting a moderate preference for style consistency while maintaining reasonable content integrity.

4.3 Experimental Results

4.3.1 Stage One

In the first stage of experimentation, the experiment systematically explored the parameter space of the base model and the WikiArt-pretrained style model to identify an optimal configuration for artistic style transfer tasks. Experimental results demonstrate that setting the sampling steps to 50 offers a favourable trade-off between image quality and computational efficiency.

Figure 3. Observed differences in output images under different parameter ratios (Picture credit: Original)

Moreover, the combination of Free Scale (fs) and WikiArt Scale (ws) parameters within the range of [0.7, 0.3] to [0.9, 0.1] yielded consistent performance in balancing style expression and content preservation. This outcome can be attributed to the inherent denoising stability of diffusion models. Through manual observation, the experiment analyzed the parameter combinations ranging from [0.7, 0.3] to [0.5, 0.5]. As shown in Figure 3, there is a substantial increase in the absence of subject features in the output images. Concurrently, the degree of stylization exhibits a comparable rise, suggesting a potential enhancement in its applicability to a broader range of abstract styles and images characterized by ambiguous subjects. Additionally, small proportion of stylistic guidance prevents semantic drift during high-dimensional sampling, thereby enhancing structural fidelity and artistic coherence. As illustrated in Figure 4, configurations such as fs=0.8 and ws=0.2 produced stylized outputs that closely align with textual prompts while effectively retaining the local structure and fine details of the original input.

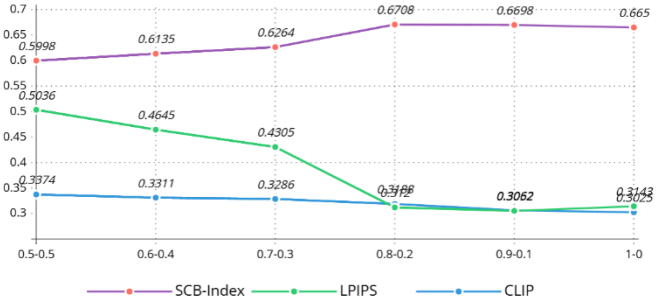

Figure 4. Evaluation curves under different parameter ratios (Picture credit: Original)

To further validate the model's task stability under optimal conditions, the experiment conducted controlled experiments using a fixed prompt—"An oil painting of a dog in impressionism style."—and a standardized input image, shown in Figure 5. The model was repeatedly executed under identical settings to assess its consistency across runs.

Quantitative evaluations demonstrate that the proposed method exhibits notable advantages in semantic alignment, content fidelity, and generation stability. Specifically, the stylized images achieved a CLIP similarity score of 0.3146, indicating a high level of alignment between the generated visuals and the guiding textual prompt. This performance surpasses that of baseline models in terms of cross-modal consistency. In parallel, the LPIPS score reached 0.3311, suggesting that the model effectively preserves the structural integrity and detailed content of the original image, particularly in maintaining the geometric features of the subject.

Figure 5. The stylized results using standard input (Picture credit: Original)

Furthermore, despite the inherent stochasticity of the diffusion process, repeated sampling over n = 50 steps produced results with minimal variation in key semantic attributes, such as eye position and limb proportions. As shown in Figure 6, The standard deviation of the SCB metric and other metrics remained low ($\sigma_{SCB} = 0.0090$), further validating the stability and reliability of the dual diffusion control mechanism and the learnable noise generation strategy. Together, these results affirm the model's robustness in producing high-quality, consistent stylizations across multiple runs.

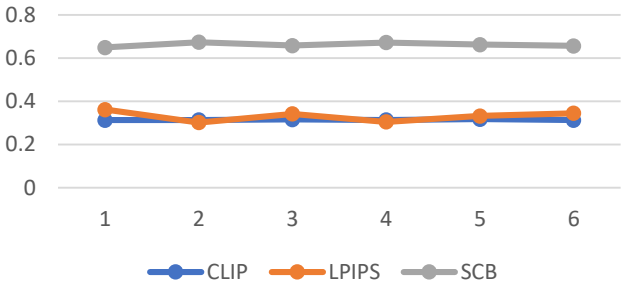

Figure 6. Stability Testing Results on Standardized Input with Optimized Parameters. (Picture credit: Original)

Building upon the previous experiments, the experiment further evaluates the model's adaptability across different style categories. Using a single standardized input image, it conducts generation tests guided by various textual style prompts to assess DiffStyler's performance across multiple stylistic directions. As shown in Figure 7, the model demonstrates significantly stronger performance in artistic style transfer tasks (e.g., "oil painting style") compared to non-artistic styles, exhibiting superior style alignment and content preservation.

Figure 7. Stylization Effects on the Same Image with Varying Style Directions (Picture credit: Original)

The image on the left serves as the input for three different style prompts. Among the results, the output for the oil painting style (rightmost) successfully integrates characteristic stylistic textures and color palettes while preserving local structural and semantic details from the original image. In contrast, results for non-artistic styles exhibit either insufficient stylization or notable distortion of the main subject, reflecting a poor balance between style rendering and content retention.

Table 3 further demonstrates the performance gap between different style directions. While the CLIP scores for all styles are similar—indicating that the models effectively capture and incorporate semantic features from textual cues—the differences in LPIPS scores reveal significant differences in content preservation. Specifically, the pixel and cartoon styles are significantly inferior to the artistic style in preserving the structure and visual details of the original image. While the 3D style preserves content better than the pixel and cartoon styles, its overall performance is still inferior to the artistic style. These results indicate that while the models are consistently consistent with textual guidance, the style-specific generation capabilities, especially in maintaining the fidelity of the original image, vary significantly across style domains.

Table 3. Evaluation Metrics on the Same Image with Varying Style Directions

Scores	Pixel	3D	Cartoon	Art
CLIP Score	0.3066	0.2983	0.3162	0.3132
LPIPS Score	0.7082	0.4176	0.7010	0.3621
SCB − Index	0.5087	0.6225	0.5145	0.6491

These findings align with expectations and suggest that the model excels in artistic style transfer tasks. This can be attributed to the style model's pre-training on the WikiArt dataset, which endows it with strong specialization in artistic representation. It is thus reasonable to infer that extending the training dataset to include a broader range of non-artistic styles (e.g., cartoon, comic, photographic filters) would likely improve the model's performance in those style domains as well.

4.3.2 Stage Two

Building on the findings of the first experimental phase, this stage further investigates the model's performance across different sub-categories of artistic styles. Specifically, the paper explore whether abstract and highly stylized prompts significantly impact generation quality and whether the distribution of optimal parameters remains consistent under such conditions.

To this end, four prompts with varying levels of abstraction were crafted based on the previously used dog image. Under fixed conditions (latent code strength lc=3, sampling steps = 50), the paper systematically adjusted the ratio between the free scale (fs) and WikiArt scale (ws) parameters to observe differences in model performance across styles.

For each prompt, multiple generations were performed to mitigate the influence of occasional sampling anomalies. As illustrated in Figure 8, using the optimal parameter configuration identified in Phase 1 (fs=0.8,ws=0.2), the prompts were arranged from left to right in increasing order of abstraction. A preliminary visual inspection reveals that the model effectively completes the style transfer tasks across all four cases. In the leftmost output—corresponding to the most realistic style—the model precisely stylizes the image while fully preserving the main subject's structural features. However, as the prompt abstraction increases toward the right, the generated outputs exhibit progressively more distortion of the

original subject, even though the model consistently demonstrates strong semantic alignment with the intended artistic style.

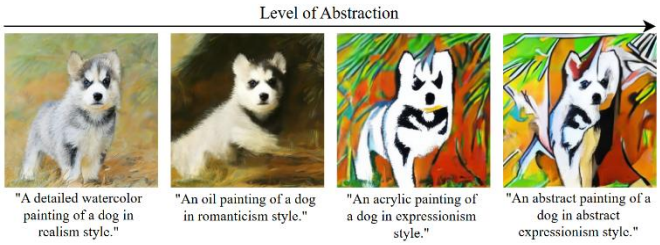

Figure 8. Stylization results of dog images under prompts with different levels of abstraction. (Picture credit: Original)

Table 4. Evaluation Metrics on the Same Image with different levels of abstraction.

Scores	1	2	3	4
CLIP Score	0.3162	0.3064	0.3381	0.3232
LPIPS Score	0.2796	0.4295	0.5829	0.5966
SCB − Index	0.6830	0.6201	0.5683	0.5583

Quantitative evaluation results (Table 4) further corroborate these observations. The SCB-Index across the four generations shows a declining trend, primarily driven by the continuous increase in LPIPS scores, while CLIP-Style scores remain stable and high. This suggests that although the model maintains robust style adherence even for abstract prompts, the preservation of fine-grained content features gradually deteriorates, leading to a slight decline in overall generation quality.

In Figure 8, the first and last images represent two extremes of output—realistic and abstract, respectively. The strong performance observed in the first image, consistent with the results in Figure 5, confirms that the previously determined optimal fs-ws parameter ratio is particularly well-suited for styles leaning toward realism. Based on this finding, the subsequent experiment fixed the abstract prompt corresponding to the last image and focused on exploring the effects of varying the fs-ws parameter ratios.

After multiple rounds of stable generation and comparative analysis, it was observed that the model exhibited particularly pronounced variations in abstraction when the ws ratio was high (between 0.5 and 1.0). Within the range of 0.3–0.7 ws proportion, the model demonstrated an ability to achieve a high degree of integration between original image details and the target style, with the specific optimal parameter setting influenced by the degree of abstraction in the target style. Furthermore, the strong stylization introduced notable instability in the generation process, resulting in increased fluctuations in output quality and evaluation metrics, thus presenting greater challenges for quantitative assessment.

Based on both human evaluation on Figure 9 and metric analysis, it was confirmed that the model maintained a consistently strong understanding of the target style throughout the experiment. Taking into account subjective elements commonly associated with abstract art, the experiment validated that, under highly abstract style prompts, the model was still capable of achieving high-quality style transfer within an appropriate range of parameters.

Figure 9. Abstract stylization outputs under different parameter ratios (Picture credit: Original)

5 Conclusion

This study focuses on DiffStyler, a text-driven image stylization framework, and investigates its effectiveness and technical advantages in the task of artistic style transfer. Experimental results explore the optimal parameter range and demonstrate that, under suitable parameter configurations, the model can effectively preserve the structure and details of the original image while accurately achieving the intended stylistic expression. Further testing reveals that DiffStyler produces stable and outstanding results when applying different style prompts to the same image, particularly excelling in artistic style representation—a performance largely attributed to the targeted training of the WikiArt style model. The combined evaluation using CLIP scores, LPIPS metrics, and human perceptual assessments confirms that the selected parameters achieve a well-balanced trade-off between content preservation and stylistic fidelity.

More importantly, DiffStyler exhibits significant potential in real-world applications. The experimental findings show that the generated artworks not only align closely with the target textual descriptions in terms of style expression, but also maintain a high degree of content fidelity. Further experiments validate DiffStyler's ability to handle highly abstract style prompts while maintaining strong semantic alignment and visual coherence. Even under significant abstraction, the model successfully integrates the core features of the original image with the target style, demonstrating remarkable robustness and adaptability across varying degrees of stylistic abstraction.

These capabilities suggest that DiffStyler has considerable practical value in domains such as digital art, creative design, and cultural heritage preservation. Artists and designers can leverage this technology to rapidly generate stylized sketches or high-quality digital artworks, thereby lowering the barrier to creation, enhancing efficiency, and stimulating creative inspiration.

In addition, the experiments reveal the feasibility of extending DiffStyler to broader style transfer tasks. By integrating diverse style models and customized training weights, the system is expected to perform equally well in non-artistic applications—such as cartoon rendering, product advertising, user interface design, and personalized image processing—thereby meeting the varied needs of commercial and industrial environments.

Ultimately, the superior performance of DiffStyler in text-driven image stylization not only offers strong technical support for artistic image generation, but also lays a solid foundation for more complex cross-modal generation tasks such as video stylization and dynamic scene synthesis. The architecture and mechanisms proposed in this study provide new theoretical insights and engineering pathways for advancing artificial intelligence in the domain of visual content creation.

References

1. Gatys, L., Ecker, A., and Bethge, M.: 'A Neural Algorithm of Artistic Style,' Journal of Vision, vol. 16, no. 12, p. 326, Sep. 2016
2. Liu, S., et al.: 'AdaAttN: Revisit Attention Mechanism in Arbitrary Neural Style Transfer,' 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2021
3. An, J., Huang, S., Song, Y., Dou, D., Liu, W., and Luo, J.: 'ArtFlow: Unbiased Image Style Transfer via Reversible Neural Flows,' 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 862–871, Jun. 2021
4. Bai, Y., Liu, J., Dong, C., and Yuan, C.: 'ITstyler: Image-optimized Text-based Style Transfer,' arXiv (Cornell University), Jan. 2023
5. Zhang, Y., et al.: 'Inversion-based Style Transfer with Diffusion Models,' Jun. 2023
6. Huang, N., et al.: 'DiffStyler: Controllable Dual Diffusion for Text-Driven Image Stylization,' IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 2, pp. 3370–3383, Feb. 2025
7. Xu, X., et al.: 'Style Transfer: From Stitching to Neural Networks,' arXiv (Cornell University), pp. 526–530, Sep. 2024
8. Ye, H., Xue, L., Chen, X., and Liu, W.: 'Research on the Method of Landscape Image Style Transfer based on Semantic Segmentation,' 2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), vol. 34, pp. 1171–1175, Dec. 2021
9. Ramesh, A., et al.: 'Zero-Shot Text-to-Image Generation,' arXiv, Feb. 2021
10. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B.: 'High-Resolution Image Synthesis With Latent Diffusion Models,' Thecvf.com, pp. 10684–10695, 2022
11. Kwon, G., and Ye, J. C.: 'CLIPstyler: Image Style Transfer with a Single Text Condition,' 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2022
12. Huang, X., and Belongie, S.: 'Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization,' IEEE Xplore, Oct. 1, 2017
13. Lei, M., Song, X., Zhu, B., Wang, H., and Zhang, C.: 'StyleStudio: Text-Driven Style Transfer with Selective Control of Style Elements,' arXiv (Cornell University), Dec. 2024
14. Wang, Z., and Liu, Z.-S.: 'StyleMamba: State Space Model for Efficient Text-driven Image Style Transfer,' arXiv (Cornell University), May 2024