SigStyle: Signature Style Transfer via Personalized Text-to-Image Models

Ye Wang[1], **Tongyuan Bai**[1], **Xuping Xie**[2], **Zili Yi**[3], **Yilin Wang**[4*], **Rui Ma**[1,5*]
[1] School of Artificial Intelligence, Jilin University
[2] College of Computer Science and Technology, Jilin University
[3] School of Intelligence Science and Technology, Nanjing University
[4] Adobe
[5] Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MOE, China
{yewang22, baity23, xiexp21}@mails.jlu.edu.cn, yi@nju.edu.cn, yilwang@adobe.com, ruim@jlu.edu.cn

Abstract

Style transfer enables the seamless integration of artistic styles from a style image into a content image, resulting in visually striking and aesthetically enriched outputs. Despite numerous advances in this field, existing methods did not explicitly focus on the *signature style*, which represents the distinct and recognizable visual traits of the image such as geometric and structural patterns, color palettes and brush strokes etc. In this paper, we introduce SigStyle, a framework that leverages the semantic priors that embedded in a personalized text-to-image diffusion model to capture the signature style representation. This style capture process is powered by a hypernetwork that efficiently fine-tunes the diffusion model for any given single style image. Style transfer then is conceptualized as the reconstruction process of content image through learned style tokens from the personalized diffusion model. Additionally, to ensure the content consistency throughout the style transfer process, we introduce a time-aware attention swapping technique that incorporates content information from the original image into the early denoising steps of target image generation. Beyond enabling high-quality signature style transfer across a wide range of styles, SigStyle supports multiple interesting applications, such as local style transfer, texture transfer, style fusion and style-guided text-to-image generation. Quantitative and qualitative evaluations demonstrate our approach outperforms existing style transfer methods for recognizing and transferring the signature styles.

1 Introduction

Style transfer technology (Gatys, Ecker, and Bethge 2016), which seamlessly incorporates stylistic elements into content images to produce visually impactful results, has gained widespread attention in recent years due to its extensive applications in art design, photography, fashion and other fields. A critical aspect of style transfer is the preservation of the original style during the transfer process. Ensuring that these intricate details and artistic expressiveness are essential for achieving high-quality results, especially when dealing with complex styles such as the shape and layout of artistic elements, and the patterns of brush strokes and lines.

Early style transfer methods primarily include techniques relying on local region matching (Zhang et al. 2013; Wang et al. 2004), or use convolutional neural network (CNN) (Gatys, Ecker, and Bethge 2016; Gatys et al. 2017; Kolkin, Salavon, and Shakhnarovich 2019) or feed-forward network (Deng et al. 2020; Huang and Belongie 2017; Liao et al. 2017; Zhang et al. 2022) to achieve style transfer. With the rapid advancement of diffusion models, diffusion-based style transfer has significantly progressed. These methods can be categorized into two types: tuning-based and tuning-free. A representative tuning-based method is InST (Zhang et al. 2023b), which introduces a textual inversion-based approach that maps a given single reference style image into a corresponding textual embedding. This textual embedding is then used as a condition to achieve style transfer for the content image. In contrast, tuning-free methods such as StyleID (Chung, Hyun, and Heo 2024), DiffStyle (Jeong, Kwon, and Uh 2023), InstantStyle (Wang et al. 2024a), InstantStyle-Plus (Wang et al. 2024b), and FreeTuner (Xu et al. 2024) merge style and content features extracted from the attention layers of Stable Diffusion (Rombach et al. 2022) to achieve style transfer. These methods offer superior computational efficiency and only require a single forward pass without the need for additional tuning.

Despite the significant progress of the aforementioned methods, signature style transfer remains underexplored. Signature style refers to the unique and recognizable visual traits that defines a particular artistic style, such as geometric and structural patterns, color palettes, and brush strokes. For example, as illustrated in the first row of Figure 2, the signature style of this image is defined by the structural arrangement and composition of numerous small images that together form the figure of a person. Additionally, the signature style of the image in the second row is characterized by geometric and structural patterns, as well as distinctive color palettes. Although existing methods often succeed in transferring basic color information, they fail to capture and retain the essential artistic style from the reference images, including small image blocks, colorful ribbon-shaped lines and other intricate characteristics as shown in Figure 2. This highlights a critical limitation: current methods struggle to achieve signature style transfer.

The main reason for the aforementioned issues is that existing methods insufficiently consider the distinct visual traits of style images. Meanwhile, personalized text-to-image models such as Dreambooth (Ruiz et al. 2023a)

*Corresponding authors

(a) Global Style Transfer

(b) Local Style Transfer

(c) Style-Guided Text-to-Image Generation

(d) Texture Transfer

Figure 1: Our method can achieve high-quality global style transfer (a) while keeping the signature style such as distinct and recognizable visual traits like geometric and structural patterns, color palettes and brush strokes etc. Also, our method is flexible and supports local style transfer (b), style-guided text-to-image generation (c), and texture transfer (d). Best viewed in color.

can capture rich information about style concepts including structures, pose, texture, lines, and more. Inspired by this, we propose SigStyle, a novel framework that leverages rich priors embedded in a personalized text-to-image model to capture complex style for facilitating signature style transfer. However, existing customized text-to-image models often require multiple reference images for fine-tuning, making them unsuitable for style transfer tasks that depend on a single reference image. To address this limitation, we propose a hypernetwork-powered, style-aware fine-tuning mechanism that enables precise concept learning and accurate inversion of style attributes using just one style reference image.

Specifically, we employ a lightweight hypernetwork to modulate and refine diffusion UNet weights. This strategy can not only ensure smoother updates of the parameters and reduces the likelihood of overfitting, but also effectively capture and represent the signature style attributes from a single reference image. Furthermore, unlike existing fine-tuning approaches (Zhang et al. 2023b; Ruiz et al. 2023a; Zhang et al. 2023b), our method focuses on fine-tuning only the modules related to style attributes, instead of the entire diffusion network. Such fine-tuning mechanism not only enhances tuning efficiency, but also improves the style inversion accuracy. Based on this fine-tuning mechanism, we can represent a signature style as a special token *. Subsequently, we define style transfer as the reconstruction process of the content image based on the style token learned from the customized diffusion model. Additionally, to maintain content consistency, we propose a time-aware attention swapping technique inspired by PhotoSwap (Gu et al. 2024a) and SwapAnything (Gu et al. 2024b). This technique transfers content-related attention from the original image generation process to the target image generation during the early denoising timesteps, ensuring content consistency throughout the style transfer process.

Figure 2: Signature style transfer comparison with SOTA methods on two complex style references.

As shown in Figure 1, our method achieves high-quality signature style transfer results across various complex style references. Moreover, our framework not only supports global style transfer, but also supports a broad range of applications, including local style transfer, texture transfer, style fusion, and style-guided text-to-image generation. Extensive experiments and evaluations further demonstrate the versatility and effectiveness of our method.

Our contributions can be summarized as follows:

- We propose SigStyle, a novel framework that is the first to explicitly focusing on the challenging signature style transfer via utilizing the personalized text-to-image diffusion models.
- We introduce a hypernetwork-powered style-aware fine-tuning mechanism that can learn the signature style attributes from only a single style image. This approach overcomes the limitation of customization methods that need multiple reference images, making it more suitable for style transfer tasks.
- Extensive experiments show that our method outperforms existing state-of-the-art methods on style trans-

fer. Notably, our approach excels in preserving signature style details, highlighting its superior capability in signature style transfer. Moreover, the diverse applications emphasize the generalizability and versatility of our method.

2 Related Work

Style Transfer. Style transfer (Zhang et al. 2019; Wang et al. 2020b,a; Park and Lee 2019; Lu et al. 2019; Li et al. 2018, 2017; Lai et al. 2017; Gatys, Ecker, and Bethge 2016) applies the artistic style of one image to another while preserving the latter's content and structure. Early methods (Zhang et al. 2013; Wang et al. 2004) relied on handcrafted features, while CNN-based approaches (Gatys, Ecker, and Bethge 2016; Gatys et al. 2017; Kolkin, Salavon, and Shakhnarovich 2019) leveraged pre-trained networks to capture style pattern. Arbitrary style transfer methods (Deng et al. 2020; Huang and Belongie 2017; Liao et al. 2017; Zhang et al. 2022) further improved this by using unified feed-forward models for flexible inputs.

In recent years, diffusion models have increasingly been employed for style transfer. These methods can be broadly categorized into two types: tuning-based methods and tuning-free methods. A representative example of the former is InST (Zhang et al. 2023b), which introduces a text inversion-based approach, aiming to map a given style to its corresponding text token. StyleDiffusion (Wang, Zhao, and Xing 2023) refines diffusion models by introducing a CLIP-based style decoupling loss, effectively separating style from content. Tuning-free methods achieve style transfer in a single forward process without model fine-tuning. DEADiff (Qi et al. 2024) utilizes the Q-Former (Li et al. 2023) and paired datasets to extract decoupled representations of content and style, facilitating style transfer. InstantStyle (Wang et al. 2024a) uses specific block injection techniques to implicitly decouple content and style for effective style transfer. Building on this, InstantStyle-Plus (Wang et al. 2024b) introduces ControlNet (Zhang, Rao, and Agrawala 2023) to further maintain the integrity of image content. StyleID (Chung, Hyun, and Heo 2024) adjusts self-attention layers and introduces novel techniques such as query preservation and initial latent AdaIN to maintain content integrity.

Nevertheless, these approaches struggle with achieving signature style transfer. In this paper, we address this challenge by utilizing a personalized text-to-image model to perform signature style transfer, generating visually compelling and aesthetically enhanced outputs.

Personalized Text-to-Image Generation. Recent studies (Ruiz et al. 2023a; Gal et al. 2022; Ruiz et al. 2023b; Kumari et al. 2023) have increasingly focused on using visual exemplars as a foundation for image generation to address the inherent ambiguity and unpredictability of text-based prompts. This approach involves collecting multiple reference images to fine-tune diffusion models. However, for style transfer tasks that rely on a single style image, the aforementioned methods often cause severe overfitting when fine-tuning on this single image, significantly diminishing the effectiveness of the style transfer. In contrast, our proposed hypernetwork-powered style aware fine-tuning method can overcome the limitations associated with fine-tuning on a single style image. It enables precise inversion of style attributes without the drawbacks of overfitting

Parameter Efficient Fine Tuning (PEFT). PEFT represents an innovative approach in the refinement of deep learning models, emphasizing the adjustment of a subset of parameters rather than the entire model. These parameters are identified as either specific subsets from the originally trained model or a minimal number of newly introduced parameters during the fine-tuning phase. PEFT has been applied in text-to-image diffusion models (Saharia et al. 2022; Rombach et al. 2022) through techniques such as LoRA (Ryu 2023) and adapter tuning (Mou et al. 2023; Ye et al. 2023; Wei et al. 2023; Chen et al. 2024; Ma et al. 2023). In this paper, we leverage a hypernetwork to adjust and refine a unique subset of pre-trained parameters.

3 Method

3.1 Preliminaries

Stable Diffusion. Stable Diffusion (Rombach et al. 2022) is a state-of-the-art text-to-image model that operates in a low-dimensional latent space. It encodes an image x into a latent z via a VAE encoder, adds noise ϵ at time step t to obtain a noisy latent z_t, and uses a CLIP text encoder τ to incorporate textual prompts c through cross-attention layers. A conditional UNet ϵ_θ is then trained to predict the noise ϵ, guided by the following training objective:

$$L_{SD}(\theta) := E_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau(c))\|^2\right]. \quad (1)$$

Attention Mechanisms in Diffusion. The UNet-based diffusion model comprises self-attention and cross-attention modules. As demonstrated in (Hertz et al. 2022; Tumanyan et al. 2023; Gu et al. 2024a,b), self-attention maps capture image structure and identity-related information. The computation of self-attention maps is defined as follows:

$$SA = \text{Softmax}\left(\frac{Q_s K_s^T}{\sqrt{d}}\right), \quad (2)$$

where Q_s and K_s represent different projections of visual features, and d is the dimension of feature used for scaling.

3.2 SigStyle

Pipeline Overview. Given a style image I_s and a content image I_c, SigStyle can seamlessly transfer the signature style to the content image while preserving the original content. The SigStyle pipeline is illustrated in Figure 3. First, we use a hypernetwork to perform style-aware fine-tuning on diffusion model with a single style image and represent the target style using a special token $*$ (see Figure 3.a). Next, we employ DDIM Inversion to obtain the noise z_t that can reconstruct the content image. Then, we extract the required self-attention maps SA_c from the UNet, which preserve the structure and content information of the content image. Finally, during the generation of the target image that conditioned on the noise z_t and the target text prompt P_t, we use the obtained self-attention maps to replace the corresponding ones in the first k steps to maintain the content information (see Figure 3.b).

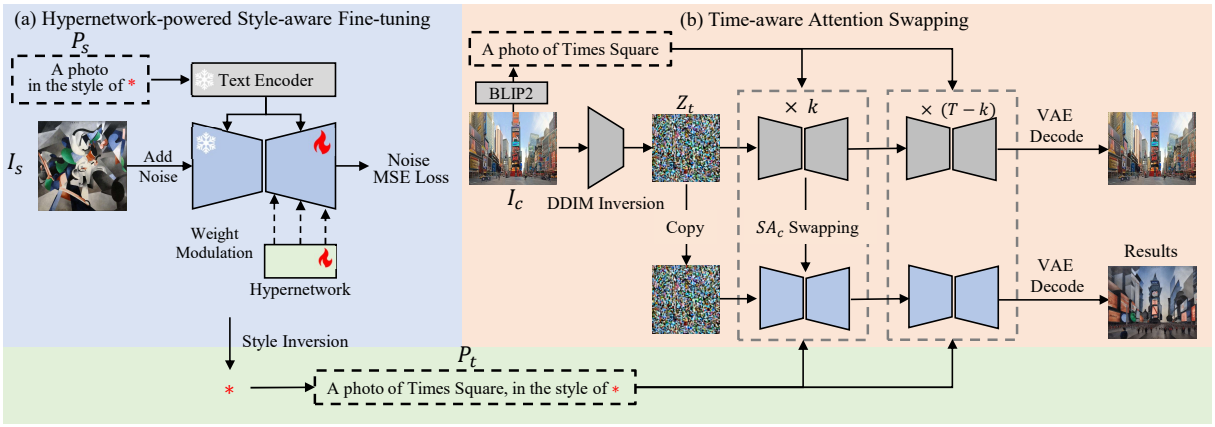

Figure 3: The SigStyle framework. First, given a style image, we perform hypernetwork-powered style-aware fine-tuning for style inversion and represent the reference style as a special token * (see Figure 3.a). In Figure 3.b, the upper branch represents the reconstruction process of the content image, while the lower branch represents the generation process of the target image. When generating the target image using a pre-trained model and target text, we first use DDIM Inversion to map the content image into noise latents, which are then copied as the initial noise for generating the target image. Then, we adopt time-aware attention swapping to inject structural and content information during the first k steps of the denoising process (see Figure b). In the subsequent $T - k$ steps, we proceed with the usual denoising process without any swapping. Finally, by decoding with VAE, we obtain the style-transferred image.

Hypernetwork-powered Style-aware Fine-tuning. Object-level inversion (Ruiz et al. 2023a) typically requires fine-tuning the entire UNet, but such an approach is not suitable for style inversion. We hypothesize that style, as an attribute of the image, should be learned and understood by specific modules within the network. To verify this hypothesis, we conducted an in-depth analysis of the style attribute learning preferences of UNet.

Style Learning Preferences Analysis of UNet. Previous work (Voynov et al. 2023; Zhang et al. 2023a) has also attempted to analyze the learning preferences of different layers, focusing mainly on shape and color. However, these studies did not accurately identify the specific layers responsible for learning style attributes. Therefore, we conducted a simple experiment using Stable Diffusion to analyze the learning preferences of different modules of UNet. As shown in Figure 4, we first divided the UNet into two parts: the encoder and the decoder. we selected a style reference image, for example, *The Starry Night*, and inputted default text prompts such as "a photo in the style of *." The * is learnable token embedding. Subsequently, we separately fine-tuned the encoder and decoder modules of the UNet and utilized the fine-tuned model to generate the corresponding images. During fine-tuning, * is progressively refined to encapsulate the distinctive style patterns in the style image. This enables * to function as an abstract, high-level representation of the style, effectively acting as a trigger for the fine-tuned T2I model to generate stylized outputs. Importantly, the original style image is not required as input once the model has been fine-tuned, as * captures the essential elements of the style. During inference, The text prompt used for inference is "a dog in the style of *".

We observe significant differences in the generated results when separately fine-tuning the encoder and decoder. Fine-tuning the encoder fails to produce images that match the style of the reference image, while fine-tuning the decoder results in images with similar styles. This experiment further validates our hypothesis that style attributes are learned by specific network modules, namely the decoder module.

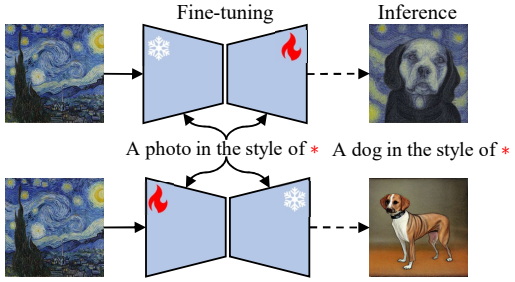

Figure 4: Style learning preferences analysis of UNet's encoder and decoder.

Based on this insight, we propose a style-aware fine-tuning mechanism that tunes only the decoder module of the UNet to learn style attributes. However, this approach may suffer from severe overfitting issues, particularly when fine-tuning with a single image. To address this challenge, we use a hypernetwork to refine and modulate the network parameters, achieving smoother updates and thereby reducing the risk of overfitting.

Hypernetwork. The hypernetwork architecture is shown in the Figure 5, we draw inspiration from E4T (Gal et al. 2023) to design our hypernetwork. The module takes as input a learnable constant $cons$ (default-initialized to 1) and the dimension information $[dim_r, dim_c]$ of the target weight parameters. It is then trained to predict weight offsets in

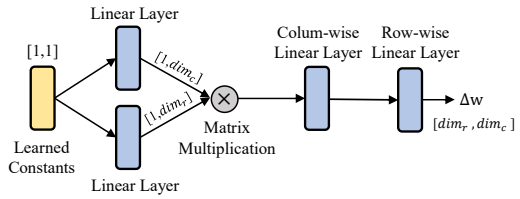

Figure 5: The architecture of hypernetwork.

the same dimensions as the target weight parameters. Here, dim_r represents the number of rows of the target weight parameters, and dim_c represents the number of columns. In detail, the learnable constant passes through two linear layers, yielding outputs that are multiplied to derive the initial weight offset matrix. Row and column transformations are then applied to this matrix to obtain the final weight offset matrix Δw. As discussed in the literatures (Gal et al. 2023; Kumari et al. 2023; Wei et al. 2023), the weights of self-attention and cross-attention play a crucial role in the process of image customization. Therefore, we utilize the hypernetwork as a weight offsets prediction module to modulate and guide the updates of attention-related weights within the decoder. The high-level parameter update process is defined as follows:

$$\Delta w = hypernetwork(cons, dim_r, dim_c), \quad (3)$$

$$w^*_{attn} = w_{attn} + \lambda * \Delta w, \quad (4)$$

where w_{attn} denotes the general term for attention-related parameters, including the query matrix, key matrix and value matrix for self-attention and cross-attention layers; λ is a weight coefficient that is used to regulate the updating strength of parameters. During training, λ is set to 1.0.

Loss Function. To guide the style inversion, we adapt the original noise prediction loss function to work with the text prompt for style learning:

$$Loss(\theta) := E_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, \tau(P_s))\|^2\right]. \quad (5)$$

Note that θ denotes the parameters of the decoder and hypernetwork, t denotes the current time step, ϵ denotes the noise, x_t represents the noise latent of style image at time t, $\tau(P_s)$ represents the CLIP embedding of the input text prompt, i.e., "a photo in the style of *."

Time-aware Attention Swapping. To achieve style transfer, a straightforward approach could be using the noise of a content image as input and performing denoising with the pre-trained personalized diffusion model. While this method captures the target style, it often alters the original image's content. Previous work (Gu et al. 2024a,b) has shown that the self-attention feature maps in the early stages of denoising encode the image's content and structural information. Therefore, we first store the self-attention maps from the UNet during the content image reconstruction, treating them as content priors. During the generation of the target image, we replace the self-attention maps with these content priors. Since content and structural information are typically established in the early stages of denoising, this attention swapping operation is performed only within the first k steps.

Implementation Details. We employ Stable Diffusion 1.4 (Rombach et al. 2022) as our base model. We use BLIP-2 (Li et al. 2023) to generate the text prompt for the content image. We apply random cropping and horizontal flipping to the input image to improve the robustness for style learning. Our model is trained on a single NVIDIA A6000 GPU with a batch size of 1 and a learning rate of 1e-6. The number of fine-tuning steps and time may vary slightly for each reference image, but on average, approximately 1500 steps are sufficient. We set k = 25 for time-aware attention swapping. It takes about 10 seconds to generate a style-transferred image during the inference stage.

4 Experiments

4.1 Qualitative Comparison

We compared multiple state-of-the-art methods to demonstrate the effectiveness of our approach. As shown in the third column of Figure 6, our method can well preserve and transfer the reference signature style. Specifically, our approach effectively transfers and preserves key signature style elements, such as the moon in the background in the first row and the small picture details in the second row of the style image, onto the content image. In contrast, other methods typically only transfer simple colors and fail to achieve high-quality signature style transfer. Furthermore, for complex style images, our method generates more natural images compared to other methods, which often produce many artifacts (see the last four columns of the second row). Our method also maintains the structure of the content image well, whereas other tuning-based methods, such as InST (Zhang et al. 2023b) and DiffuseIT (Kwon and Ye 2022), alter the original content image's structure. This further demonstrates that our hypernetwork-powered style-aware tuning method can more precisely inverse style attributes, while the time-aware attention injection effectively preserves content information.

4.2 Quantitative Comparison

We randomly selected 10 content images and 15 style images, generating a total of 150 stylized images for each method by applying the Cartesian product to the content and style images. To evaluate content fidelity, we followed (Chung, Hyun, and Heo 2024) and employed the LPIPS metric (Zhang et al. 2018), which measures the similarity between the stylized image and the corresponding content image. For style similarity, we adopted the Style Loss (Gatys, Ecker, and Bethge 2016), measuring the alignment between the stylized image and the corresponding style image.

We compared our method with eight SOTA style transfer methods: StyleID, InstantStyle, Diffuse-IT, InST, AesPA-Net, CAST, AdaAttN, and ArtFlow. As presented in Table 1, our method achieves the second-lowest LPIPS score of 0.5191 and the lowest Style Loss of 0.7641, surpassing the majority of existing approaches. This result indicates that our approach effectively preserves content and style fidelity during signature style transfer. Furthermore, we also conducted a comprehensive user study. We randomly collected

Figure 6: Qualitative comparison with various SOTA image style transfer methods for global style transfer.

Metrics\Methods	Ours	StyleID	InstantStyle	Diffuse-IT	InST	AesPA-Net	CAST	AdaAttN	ArtFlow
Style Loss ↓	**0.7641**	1.0902	0.7653	1.2312	0.9875	0.8766	0.7842	1.0288	0.7693
LPIPS ↓	0.5191	0.5959	0.7226	0.7435	0.8199	0.5328	0.6485	**0.4727**	0.6922
Rank 1 (%) ↑	**9.1**	5.06	2.76	2.85	1.01	2.20	6.07	3.49	3.49
Top 3 (%) ↑	14.54	13.61	6.16	6.25	2.48	10.40	**17.75**	13.71	15.10

Table 1: Quantitative comparison with SOTA image style transfer methods. The best results are in **bold** while the second best results are marked with underline.

12 content-style pairs, each producing nine images generated by the methods mentioned above. Participants were asked to select and rank the top three results from these nine randomly-arranged images based on two criteria: 1) the preservation of signature or key styles from the style image, including distinct and recognizable visual traits such as geometric patterns, color palettes, and brush strokes; 2) the preservation of content from the content image, including overall shape, structure, and semantics of the main object. A total of 1,152 votes were collected from 32 participants. As demonstrated in Table 1, we computed the proportion of first-place rankings (Rank 1) and the overall percentage of top three votes (Top 3) for each method. Our approach achieved the highest Rank 1 score (9.1%) and a relatively high percentage of Top 3 votes (14.54%). These results highlight that the style-transferred images generated by our method were preferred by users, showcasing its superior performance in signature style transfer.

As shown in Table 2, we compare our method with tuning-based methods like InST and Diffuse-IT, which require 20 and 18 minutes for tuning, respectively. Our method is more efficient, with 10 minutes of tuning and 10 seconds of inference. Although tuning-free methods like StyleID, InstantStyle, and CAST are faster (5-8 seconds for inference), they do not preserve style details well. Our method achieves a better balance between speed and style quality.

Methods	Tuning	Inference	Methods	Tuning	Inference
InST	20 mins	24 s	StyleID	no need	7 s
Diffuse-IT	18 mins	-	InstantStyle	no need	5 s
Ours	10 mins	10 s	CAST	no need	8 s

Table 2: Comparison of time consumption in the fine-tuning and inference phases between ours and other methods.

Figure 7: The ablation study on hypernetwork.

4.3 Ablation Study

Hypernetwork. As shown in Figure 7, our hypernetwork effectively facilitates the precise learning and inversion of the style. In contrast, when fine-tuning without the hypernetwork, the generated images match the target text but fail to retain the reference style. This further demonstrates the effectiveness of our proposed hypernetwork.

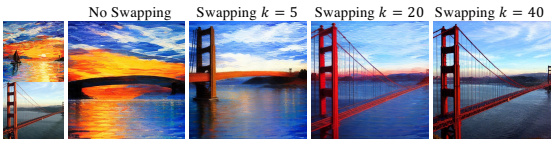

Figure 8: The ablation study on attention swapping.

Figure 9: Local style transfer results of our method. The transfer area is the foreground animal region by default.

Attention Swapping. As shown in Figure 8, when attention swapping is not used, the generated images only retain the target style but fail to maintain the image content. As the k value increases, the structure and other content-related information in the synthesized image become increasingly similar to the content image, while the style gradually diminishes. This demonstrates the effectiveness of the attention swapping technique in preserving content consistency.

4.4 Further Applications

In addition to style transfer tasks, our method supports other applications including local style transfer, texture transfer, style fusion, and style-guided text-to-image generation.

Local Style Transfer. Local style transfer applies style only to regions specified by a user-provided mask. Within the masked areas, we use SigStyle for style transfer, while denoising reconstruction is applied to non-masked areas to maintain consistency. Blending operations then integrate these regions seamlessly, producing a complete image and achieving local style transfer, as shown in Figure 9.

Texture Transfer. Texture, appearance, and style are interrelated concepts best learned by the same module, the UNet decoder. By replacing "style" with "appearance" in prompts while keeping inversion and transfer processes unchanged, a mask constrains the texture transfer region. Figure 10 demonstrates high-quality cross-domain texture transfer, preserving the original image's pose, structure, identity, and other content.

Style Fusion. We can fuse multiple styles to create a new style for transfer and customized generation, leading to more intriguing results. This is achieved by fine-tuning with multiple style images. As shown in Figure 11, our method effectively transfers the fused style onto the content image.

Style-guided Text-to-Image Generation. Our fine-tuning mechanism represents style as a special token *, enabling style-guided text-to-image generation. With a single style image, we can generate images guided by that style (see the first row of Figure 12). When using multiple style images, our method fuses them into a new style for more creative outputs (see the second row of Figure 12).

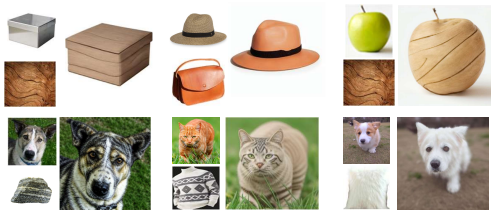

Figure 10: Cross-domain texture transfer of our method.

Figure 11: Style transfer based on multiple style references.

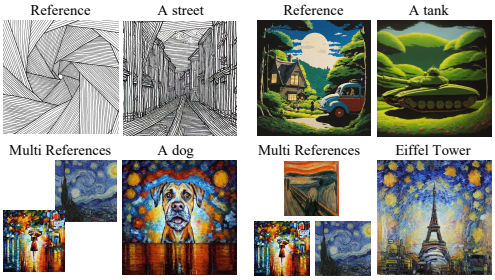

Figure 12: The style-guided text-to-image generation results.

5 Conclusion

In this paper, we presented SigStyle, a novel framework for high-quality signature style transfer using only a single style reference image. By introducing a hypernetwork-powered style-aware fine-tuning mechanism, our approach enhances the accuracy and efficiency of style inversion while addressing severe single-image overfitting issues. Additionally, our time-aware attention swapping technique ensures content consistency during the style transfer process. Extensive experiments show that SigStyle outperforms existing methods in preserving signature styles and supports various applications, including global and local style transfer, texture transfer, and style-guided text-to-image generation etc. Currently, SigStyle still needs to fine-tune the diffusion model for each given style image during inference, which limits its deployment on resource-constrained devices. How to further reduce the computation cost of the style learning and transferring process will be worthy to investigate. Moreover, it will also be interesting to explore how to use more specified text prompt to guide more refined and more controllable style transfer.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (No. 62202199, No. 62406134) and the Science and Technology Development Plan of Jilin Province (No. 20230101071JC).

References

Chen, W.; Hu, H.; Li, Y.; Ruiz, N.; Jia, X.; Chang, M.-W.; and Cohen, W. W. 2024. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36.

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.; Sun, W.; Huang, F.; and Xu, C. 2020. Arbitrary style transfer via multi-adaptation network. In *Proceedings of the 28th ACM international conference on multimedia*, 2719–2727.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gal, R.; Arar, M.; Atzmon, Y.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4): 1–13.

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.

Gatys, L. A.; Ecker, A. S.; Bethge, M.; Hertzmann, A.; and Shechtman, E. 2017. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3985–3993.

Gu, J.; Wang, Y.; Zhao, N.; Fu, T.-J.; Xiong, W.; Liu, Q.; Zhang, Z.; Zhang, H.; Zhang, J.; Jung, H.; et al. 2024a. Photoswap: Personalized subject swapping in images. *Advances in Neural Information Processing Systems*, 36.

Gu, J.; Wang, Y.; Zhao, N.; Xiong, W.; Liu, Q.; Zhang, Z.; Zhang, H.; Zhang, J.; Jung, H.; and Wang, X. E. 2024b. Swapanything: Enabling arbitrary object swapping in personalized visual editing. *arXiv preprint arXiv:2404.05717*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

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.

Jeong, J.; Kwon, M.; and Uh, Y. 2023. Training-free style transfer emerges from h-space in diffusion models. *arXiv preprint arXiv:2303.15403*, 3.

Kolkin, N.; Salavon, J.; and Shakhnarovich, G. 2019. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10051–10060.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

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

Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 624–632.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; and Yang, M.-H. 2017. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30.

Li, Y.; Liu, M.-Y.; Li, X.; Yang, M.-H.; and Kautz, J. 2018. A closed-form solution to photorealistic image stylization. In *Proceedings of the European conference on computer vision (ECCV)*, 453–468.

Liao, J.; Yao, Y.; Yuan, L.; Hua, G.; and Kang, S. B. 2017. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*.

Lu, M.; Zhao, H.; Yao, A.; Chen, Y.; Xu, F.; and Zhang, L. 2019. A closed-form solution to universal style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5952–5961.

Ma, Y.; Yang, H.; Wang, W.; Fu, J.; and Liu, J. 2023. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.

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.

Qi, T.; Fang, S.; Wu, Y.; Xie, H.; Liu, J.; Chen, L.; He, Q.; and Zhang, Y. 2024. DEADiff: An Efficient Stylization Diffusion Model with Disentangled Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8693–8702.

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.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023a. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Ruiz, N.; Li, Y.; Jampani, V.; Wei, W.; Hou, T.; Pritch, Y.; Wadhwa, N.; Rubinstein, M.; and Aberman, K. 2023b. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*.

Ryu, S. 2023. Low-rank adaptation for fast text-to-image diffusion fine-tuning.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1921–1930.

Voynov, A.; Chu, Q.; Cohen-Or, D.; and Aberman, K. 2023. $P+$: Extended Textual Conditioning in Text-to-Image Generation. *arXiv preprint arXiv:2303.09522*.

Wang, B.; Wang, W.; Yang, H.; and Sun, J. 2004. Efficient example-based painting and synthesis of 2d directional texture. *IEEE Transactions on Visualization and Computer Graphics*, 10(3): 266–277.

Wang, H.; Li, Y.; Wang, Y.; Hu, H.; and Yang, M.-H. 2020a. Collaborative distillation for ultra-resolution universal style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1860–1869.

Wang, H.; Wang, Q.; Bai, X.; Qin, Z.; and Chen, A. 2024a. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*.

Wang, H.; Xing, P.; Huang, R.; Ai, H.; Wang, Q.; and Bai, X. 2024b. InstantStyle-Plus: Style Transfer with Content-Preserving in Text-to-Image Generation. *arXiv preprint arXiv:2407.00788*.

Wang, Z.; Zhao, L.; Chen, H.; Qiu, L.; Mo, Q.; Lin, S.; Xing, W.; and Lu, D. 2020b. Diversified arbitrary style transfer via deep feature perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7789–7798.

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.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*.

Xu, Y.; Wang, Z.; Xiao, J.; Liu, W.; and Chen, L. 2024. FreeTuner: Any Subject in Any Style with Training-free Diffusion. *arXiv preprint arXiv:2405.14201*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

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, W.; Cao, C.; Chen, S.; Liu, J.; and Tang, X. 2013. Style transfer via image component analysis. *IEEE Transactions on multimedia*, 15(7): 1594–1601.

Zhang, Y.; Dong, W.; Tang, F.; Huang, N.; Huang, H.; Ma, C.; Lee, T.-Y.; Deussen, O.; and Xu, C. 2023a. ProSpect: Expanded Conditioning for the Personalization of Attribute-aware Image Generation. *arXiv preprint arXiv:2305.16225*.

Zhang, Y.; Fang, C.; Wang, Y.; Wang, Z.; Lin, Z.; Fu, Y.; and Yang, J. 2019. Multimodal style transfer via graph cuts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5943–5951.

Zhang, Y.; Huang, N.; Tang, F.; Huang, H.; Ma, C.; Dong, W.; and Xu, C. 2023b. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10146–10156.

Zhang, Y.; Li, M.; Li, R.; Jia, K.; and Zhang, L. 2022. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8035–8045.