An Efficient and Harmonized Framework for Balanced Cross-Domain Feature Integration

Shaoxu Li, Ye Pan*

John Hopcroft Center for Computer Science, Shanghai Jiao Tong University
lishaoxu, whitneypanye@sjtu.edu.cn

Abstract

Despite significant advancements in image generation using advanced generative frameworks, cross-image integration of content and style remains a key challenge. Current generative models, while powerful, frequently depend on vague textual prompts to define styles—creating difficulties in balancing content semantics and style preservation. We propose a novel framework that utilizes customized models to learn style representations. It enhances content preservation through cross-model feature and attention modulation, leveraging the inherent semantic consistency across models. Additionally, we introduce fixed feature and adaptive attention fusion to achieve the desired balance between content and style. We further develop spatial (mask-guided localized) and temporal (multi-style compositional) multi-model combinations, enabling flexible fusion of models and styles. Extensive experiments demonstrate that our method outperforms state-of-the-art approaches in balancing content preservation and stylistic coherence.

Code — https://github.com/lishaoxu1994/DiffStyler

Introduction

Paintings, powerful artistic expressions, convey unique perspectives through stylistic elements, textures, and motifs. But their subtle aesthetic qualities are hard to put into words, hindering example-based artistic style transfer. Image style transfer—integrating artistic styles with content—has attracted interdisciplinary attention(Wang et al. 2004; Gatys, Ecker, and Bethge 2015; Zhang et al. 2023b; Lv et al. 2024; Zhou et al. 2024). Text-driven methods, especially those using diffusion models(Song, Meng, and Ermon 2022; Ramesh et al. 2022), have advanced the field. But they face difficulty crafting detailed prompts to capture a style's essence. Some methods use texture inversion(Gal et al. 2022; Jeong et al. 2024) to depict styles. However, these methods are often time-consuming and inaccurately represent key attributes, limiting practical use.

To address this, we present **DiffStyler**, a diffusion-based framework that learns styles from a single image (e.g., cartoons, paintings) and applies them to natural images. It supports mask-guided localized style transfer and multi-style compositional transfer, offering exceptional control with computational efficiency. As shown in Figure 1, our method enables region-specific transfer via masks and simultaneous multi-style transfer for creative flexibility, generating cohesive outputs that blend content with target styles. Building upon text-to-image diffusion models—renowned for high-quality synthesis—our method leverages attention manipulation (Hertz et al. 2022; Tumanyan et al. 2023) and LoRA fine-tuning (Hu et al. 2021). We observe that diffusion models integrated with LoRA maintain the semantic consistency of features, enabling LoRA to learn style attributes from reference images and guide generation through carefully designed manipulation of feature and attention maps.

Based on this, we introduce mask-guided feature blending during denoising for region-specific transfer (conditioned on style references and masks) and a dynamic LoRA switching mechanism during denoising for hierarchical fusion with multiple references. These two approaches, enabling multi-LoRA combinations in spatial and temporal domains, are applicable to general diffusion methods. Our contributions can be summed up as follows: 1) We provide new empirical insights into internal spatial features across different LoRA-integrated text-to-image diffusion models. 2) We introduce a practical framework leveraging pre-trained and LoRA-integrated diffusion models for high-quality arbitrary image style transfer. 3) We propose mask-guided localized style transfer and multi-style compositional style transfer for flexible model and style fusion. 4) Quantitative and qualitative results show our method outperforms state-of-the-art approaches, achieving a significantly better balance between preserving content semantics and style attributes.

Related Work

Image Style Transfer. Image stylization converts images into diverse artistic styles. Originating from traditional handcrafted feature-matching between content and style images(Wang et al. 2004; Zhang et al. 2013), the field advanced substantially with neural style transfer(Gatys, Ecker, and Bethge 2015; Liu et al. 2021; An et al. 2021; Tan, Ji, and Pan 2024)—which leverages pre-trained CNNs for feature extraction. Recent works integrate pre-trained diffusion models(Cui et al. 2024; Yang, Wu, and He 2024; Wang et al. 2024a,b) (e.g., InST(Zhang et al. 2023b), VCT(Cheng et al. 2023)) to guide reference-based synthesis and style

*Corresponding author

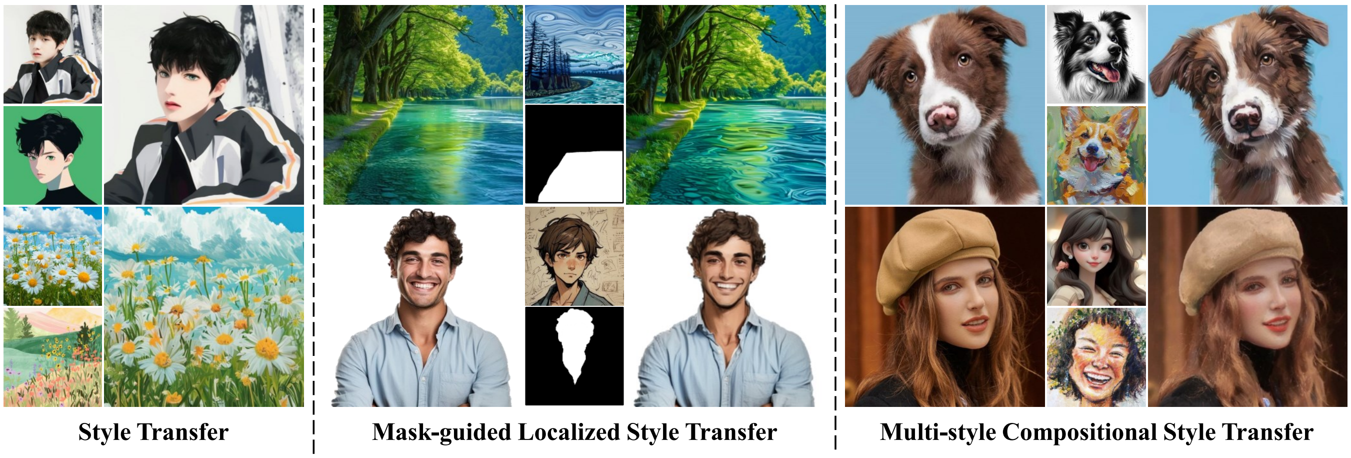

Figure 1: Style transfer results using our proposed architecture. By leveraging a content image and a style image as input, our method achieves precise attribute-level style propagation while preserving high-fidelity semantic content integrity. Beyond global style transfer, our approach enables two extensions: mask-guided localized and multi-style compositional style transfer.

translation, while AttnD(Zhou et al. 2025), StyleID(Chung, Hyun, and Heo 2024) and Z*(Deng et al. 2023) achieve zero-shot transfer via attention manipulation. StyleShot(Gao et al. 2025) enables generalized transfer without test-time tuning using a style-aware encoder (decoupling training) and content-fusion encoder, and Lin et al.(Lin et al. 2024) propose an unsupervised content-style learning method for cross-domain translation.

Semantic Localized Image Editing. Traditional style transfer often lacks semantic coherence, whereas text-to-image diffusion models offer flexible, controlled editing by distilling semantics from noise. DDIM(Song, Meng, and Ermon 2022) and Glide(Nichol et al. 2022) use sophisticated pixel manipulation during denoising for precise edits. Paint by Word(Andonian et al. 2023) and CLIP-inspired methods(Crowson 2023; Avrahami, Lischinski, and Fried 2022) use CLIP-guided gradients for text-image alignment in specific regions. RePaint(Lugmayr et al. 2022) refines diffusion iterations using unmasked image info. DiffusionCLIP(Kim, Kwon, and Ye 2022) and SDEdit(Meng et al. 2022) leverage inversion/SDE priors for text-driven, realistic edits. DiffEdit(Couairon et al. 2022) enables targeted editing via automatic mask extraction. ILVR(Choi et al. 2021) and Prompt-to-Prompt(Hertz et al. 2022) achieve localized edits without explicit masks via attention map correlations.

Composable Image Generation. Compositional image generation has drawn interest, with approaches categorized into: structured compositionality enhancement (enhancing control via scene graphs/spatial layouts(Johnson, Gupta, and Fei-Fei 2018; Yang et al. 2022; Singh, Gould, and Zheng 2023; Li, Ma, and Pan 2025)), diffusion model adaptation (adapting generative mechanisms for compositional specs(Feng et al. 2023; Huang et al. 2023)), multi-concept customization (integrating multiple semantics(Kumari et al. 2023; Han et al. 2023; Gu et al. 2023; Kwon et al. 2024; Kong et al. 2024; Wang et al. 2023)), and modular model composition (composing independent models under constraints(Du et al. 2024; Li et al. 2022)). Notably, combining multiple stylistic attributes is underexplored, which our work addresses by investigating style target integration.

Method

This work aims to generate a new image I^* that retains the semantic content of content image I_c while embodying the style of style image I_s. Using Stable Diffusion (SD)(Rombach et al. 2021) as the generative backbone, we introduce techniques including DDIM inversion, LoRA training, cross-LoRA feature/attention injection, mask-guided DDIM denoising, and multi-style DDIM denoising (Fig. 2). A key insight is that spatial features in different LoRA-integrated diffusion models exhibit strong semantic consistency. Based on this, our framework extracts features and attention maps from I_c's generation process in the original SD, then injects them (along with prompts c) into I_s's generation process in the LoRA-integrated model. This consistency enables multi-LoRA combined generation: we propose mask-guided localized style transfer (via feature fusion with masks) and multi-style compositional style transfer (via LoRA-switch-based sampling).

Preliminaries

Latent Diffusion Models. Diffusion models(Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015; Song et al. 2021) represent a class of probabilistic generative models trained to reverse a diffusion process. As the latest advancement in generative models, diffusion models have garnered significant attention due to their superior capability to generate high-quality images. In the image generation, the forward diffusion process incrementally introduces noise into an initially clean image x_0:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t} z \quad (1)$$

where $z \sim N(0, I)$ and $\{\alpha_t\}$ are the noise schedule over the time step t. The back process progressively removes noise from an initial Gaussian noise image. Typically, a neural network $\epsilon_\theta(x_t, t)$ is trained to predict the added noise.

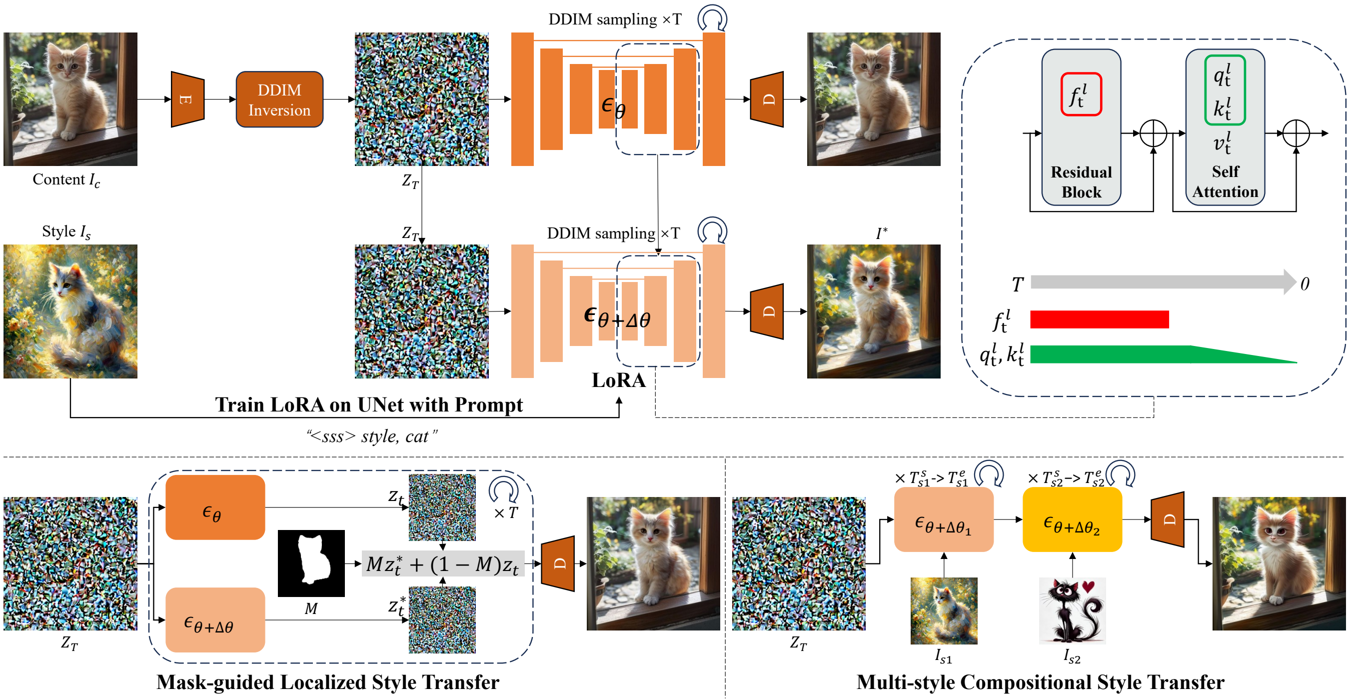

Figure 2: Method pipeline. We train LoRA on UNet with a simple prompt for the style attribute transfer. To maintain the content semantics, we execute the DDIM inversion and denoising. In the DDIM denoising process, the features and attention maps are injected into the LoRA-integrated SD model. In the denoising, we override the feature in certain steps. For attention injection, we override the q, k in certain steps and overlap them adaptively in the remaining steps. For mask-guided localized style transfer, we execute the mask-guided feature combination in the denoising process. For multi-style composition style transfer, we execute the LoRA switch in the denoising process.

Latent Diffusion Models (LDM)(Rombach et al. 2021) uses a variational auto-encoder (VAE)(Kingma and Welling 2022) to encode the images to the latent image embeddings. A text-conditioned denoising UNet(Ronneberger, Fischer, and Brox 2015) $\epsilon_\theta(z_t, t, c)$ is trained in the latent space. Our method is built upon Stable Diffusion (SD), a widely used pre-trained text-to-image model based on the LDM framework. A layer of the UNet in the SD involves a residual block, a self-attention block, and a cross-attention block. The attention module in UNet can be formulated as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_K}})V \quad (2)$$

where Q denotes the query features, K, V denotes the corresponding key and value feature, and d_K represents the dimension of the query and key vectors. Self-attention is computed purely on the spatial features. Cross-attention is computed between spatial features and the text embeddings.

Low-Rank Adaption (LoRA). Low-Rank Adaptation (LoRA)(Hu et al. 2021) is an efficient method initially employed for fine-tuning large language models by adapting a low-rank residual $\Delta\theta$, which can be decomposed into low-rank matrices, thus offering a streamlined approach to modifying the entire model parameters θ. Recently, LoRA has demonstrated its effectiveness and ease of integration in fine-tuning diffusion models, particularly in personalizing text-to-image generation. Our method utilizes LoRA to transfer the attributes of the style target in image style transfer.

Spatial Features across LoRA Models

We explore SD features for cross-LoRA image synthesis by fine-tuning two LoRA models in distinct stylistic domains. PCA comparisons with domain-specific images analyze features from each decoder layer at all timesteps t (original SD, in-domain/out-of-domain LoRA-modified SDs). Figure 3 shows initial layers capture broad semantics, later layers fine details, and LoRA-enhanced SDs have minimal feature discrepancy. Natural images exhibit indistinct model-specific feature variations, while styled images show domain-aligned patterns—specific layers (e.g., layer 5 for cartoon hair/-dog faces) align more with in-domain LoRA, with out-of-domain cartoon face features reduced at layer 5. Quantitative evaluation (20 cases, Tab. 1) confirms high feature similarity between in- and out-of-domain LoRA-modified SDs. Overall, LoRA integration subtly impacts spatial features, validating the feasibility of cross-LoRA spatial feature injection/fusion for mask-guided localized and multi-style compositional transfer.

Cross-LoRA Feature and Attention Injection

Building upon the preceding observations, we propose a novel approach to style transfer. Our method learns the tar-

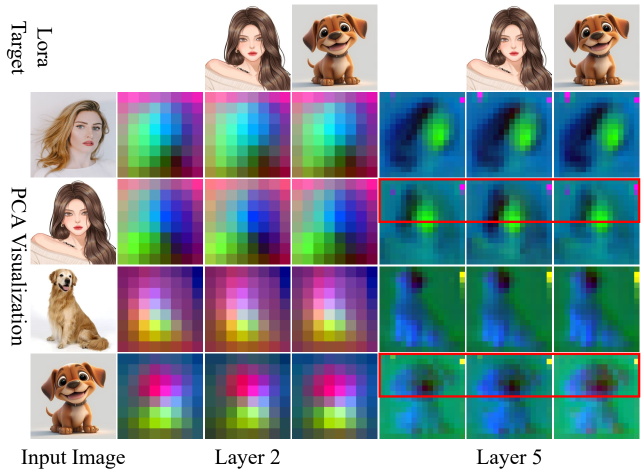

Figure 3: Analysis of features from different LoRA-integrated SD: Spatial features were extracted from decoder layers at 50% generation (t=540). PCA on features of original SD, in-domain/out-of-domain LoRA-integrated SD (visualizing top 3 components) shows early layers focus on semantics, later on textures. LoRA-integrated models have little feature difference, with in-domain ones showing better correspondences in some layers (e.g.,layer 5).

	Layer 2	Layer 5	Layer 7
In-domain↑	0.9970	0.9943	0.9922
Out-of-domain↑	0.9942	0.9883	0.9721

Table 1: Cosine similarity of features.

get's style attributes using a LoRA model, simultaneously preserving the content's semantic integrity through integrating features and attention mechanisms.

LoRA Training. Given a style image, we first train a LoRA $\Delta\theta$ on the SD UNet ϵ_θ. The learning objective for training is:

$$L(\Delta\theta) = \mathbb{E}_{\epsilon,t}[||\epsilon - \epsilon_{\theta+\Delta\theta}(\sqrt{\alpha_t}z_0 + \sqrt{1-\alpha_t}\epsilon, t, c)||] \quad (3)$$

where z_0 denotes the encoded latent embedding of the input image, ϵ denotes the random sampled Gaussian noise, $\epsilon_{\theta+\Delta\theta}$ denotes the LoRA-integrated UNet, c denotes the text embedding of the text prompt. With the LoRA-integrated SD, attributes of the style target can be injected into generating images with corresponding prompts. To avoid overfitting, the LoRA training is only carried out on the projection matrices Q, K, V in the attention modules of UNet.

Feature injection. Given a content image, we invert the encoded image embedding into the initial noise using DDIM(Song, Meng, and Ermon 2022). The design of shared initial noise has been proved to lead to visual similarity in the denoising(Tumanyan et al. 2023; Zhang et al. 2023a).

At each denoising step, we extract the guidance features f_t^l from the original denoising step: $z_{t-1} = \epsilon_\theta(z_t, t, \varnothing)$. We then inject these features into the LoRA-integrated UNet to guide the semantics of style transfer. The injection is accomplished with feature override. The operation can be expressed by:

$$z_{t-1}^* = \hat{\epsilon}_{\theta+\Delta\theta}(z_t^*, t, c; \{f_t^l\}) \quad (4)$$

where $\hat{\epsilon}_{\theta+\Delta\theta}(\cdot; \{f_t^l\})$ denotes the denoising with injected features on the LoRA-integrated UNet. To balance the structure-preserving and the deviating, we choose layer 4 as the injection layer with a predefined injection steps threshold, following the design of Plug-and-Play(Tumanyan et al. 2023). By doing this, a balance between content structure preserving and appearance deviating can be achieved.

Attention Injection. Attention injection, as a supplement to feature injection, aids in preserving details or identity consistency in image generation. In Stable Diffusion, self-attention (SA) handles spatial features, while cross-attention (CA) processes text embeddings: SA maintains spatial geometry amid style/domain variations(Tumanyan et al. 2023; Chung, Hyun, and Heo 2023; Zhang et al. 2023a), and CA preserves localized attributes or identities(Hertz et al. 2022; Huang et al. 2024; Li, Ma, and Pan 2025). Layers 4-11 are chosen as injection layers to balance structure preservation and stylistic deviation. Fixed-threshold attention injection ensures rough structural similarity to the source, yet LoRA-induced attribute changes may cause unintended detail edits (e.g., hand deformations). To address this, we propose adaptive attention override: full attention injection is applied in early sampling steps; after threshold T_a, adaptive override commences.

$$\{A_t^l\}^{**} = \begin{cases} \{A_t^l\}, & t \geq T_a \\ (1-\kappa_t)\{A_t^l\} + \kappa_t\{A_t^l\}^*, & t < T_a \end{cases} \quad (5)$$

where $\{A_t^l\}$ denotes the original SD attention of the l_{th} layer, $\{A_t^l\}^*$ denotes the LoRA-integrated SD attention, T_a denotes the specified threshold, $\kappa_t = (T_a - t)/(T - T_a) \in [0, 1]$ dynamically regulates override strength.

During each denoising step, we inject attention matrices parallel to feature injection. The modified denoising operation is formulated as follows:

$$z_{t-1}^* = \hat{\epsilon}_{\theta+\Delta\theta}(z_t^*, t, c; \{f_t^l\}, \{A_t^l\}^{**}) \quad (6)$$

Our default configuration injects self-attention into all decoder layers of the UNet architecture.

Mask-guided Localized Image Style Transfer

Mask-guided localized image editing focuses on modifying specific regions of an image while preserving the attributes of non-target areas. Our method involves replacing the encoded latent features within mask boundaries, guiding the denoising process through mask-based feature manipulation. Masks are resized to match the dimensions of latent features. Mask-guided denoising is mathematically formulated as:

$$\hat{z}_t = Mz_t^* + (1-M)z_t. \quad (7)$$

Mask-guided cross-LoRA denoising enables mask-guided style transfer. Due to the pipeline's modular design, we can synthesize images with multiple style targets and corresponding masks:

$$\begin{aligned} \hat{z}_{(t-1)} = \sum M_i \hat{\epsilon}_{\theta+\Delta\theta_i}(\hat{z}_t, t, c; \{f_t^l\}, \{A_t^l\}^{**}) \\ +(1-\sum M_i)\epsilon_\theta(\hat{z}_t, t, c). \end{aligned} \quad (8)$$

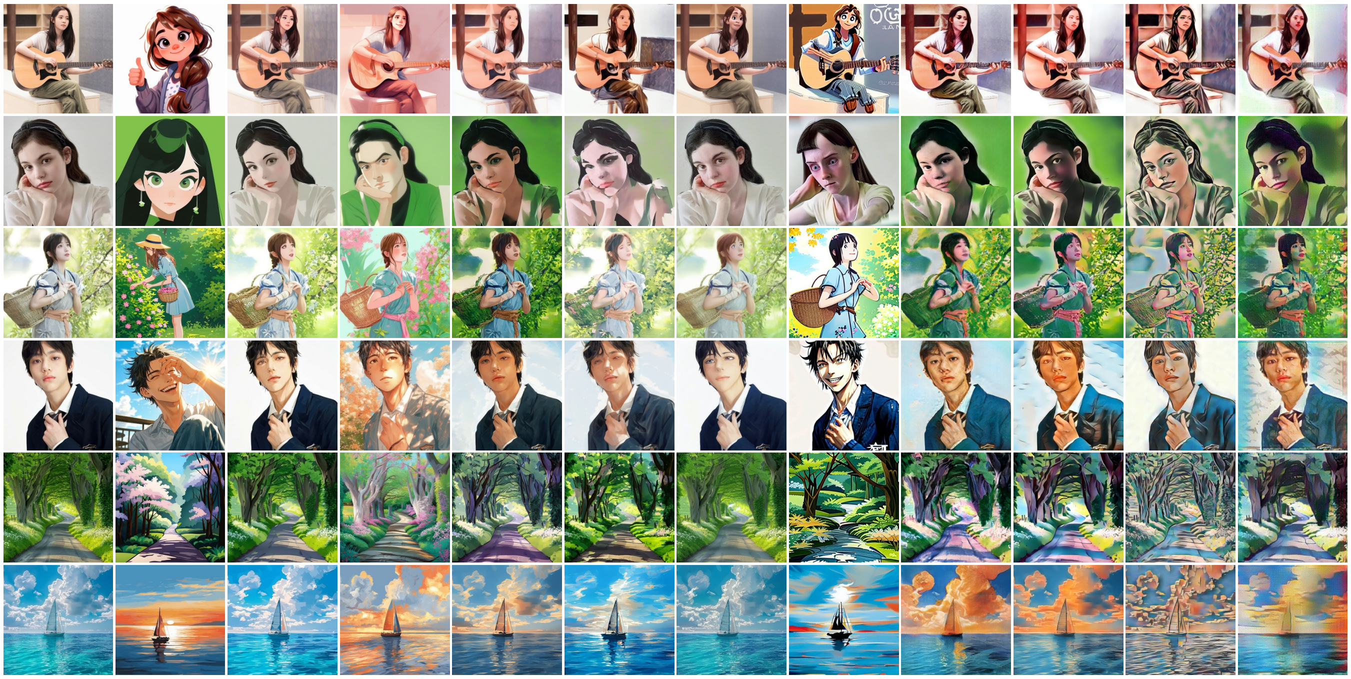

Figure 4: Qualitative comparison with some state-of-the-art image style transfer methods.

where $\hat{\epsilon}_{\theta+\Delta\theta_i}$ denotes the i_{th} LoRA-integrated UNet with injection, and M_i denotes the corresponding region mask.

Multi-style Composition Image Style Transfer

We note that the UNet is capable of encoding attributes at various scales during different sampling steps. The incorporation of LoRA has a nuanced impact on spatial feature representations. Building on these findings, we introduce adaptive LoRA-switch sampling for multi-style compositional image style transfer. When we have two LoRA-integrated UNets, we can use one for generating images at a coarse-grained level and the other for a fine-grained level ($T - T_1$ steps), as shown in Fig. 2. As a fundamental pipeline, multi-style LoRA-switch sampling can be extended to work with multiple (more than two) LoRA-integrated UNets, and it can be mathematically expressed as:

$$\hat{\epsilon}_{\theta+\Delta\theta} = \hat{\epsilon}_{\theta+\Delta\theta_i}, \qquad T^s_{si} > t \geq T^e_{si}. \tag{9}$$

where $\hat{\epsilon}_{\theta+\Delta\theta_i}$ denotes the i_{th} LoRA-integrated UNet with injection, and $[T^s_{si}, T^e_{si}]$ indicates the specific sampling time range for each LoRA model.

Experiments

We utilize Stable Diffusion v2.1-base as our generative backbone. Inversion and denoising are implemented using the diffusers(von Platen et al. 2022). All experiments are performed on a single NVIDIA GeForce RTX 3090 GPU. Input and output images are processed at 512×512 resolution. Images We train Low-Rank Adaptation (LoRA) to fine-tune the projection matrices Q, K, V in the attention modules of the UNet architecture within the diffusion model. Training parameters include a LoRA rank of 16, 200 training iterations, and a learning rate of 2×10^{-4} using the AdamW optimizer with weight decay. Deterministic DDIM inversion and denoising are performed with 50 sampling steps derived from the full diffusion process. Our default feature and attention injection thresholds are set to 30 and 25 denoising steps out of 50 total.

Comparison with Style Transfer Methods

We compare our method against leading image style transfer techniques, including five diffusion-based approaches: AttnD(Zhou et al. 2025), StyleID(Chung, Hyun, and Heo 2024), Z*(Deng et al. 2023), VCT(Cheng et al. 2023), and InST(Zhang et al. 2023b); alongside four non-diffusion-based methods: CAST(Zhang et al. 2022), StyTr^2(Deng et al. 2022), AdaAttN(Liu et al. 2021), and ArtFlow(An et al. 2021). Comparative results in Fig. 4 demonstrate that diffusion-based methods excel at transferring complex stylistic attributes, including the spatial structure and visual motifs, while non-diffusion-based methods primarily alter color palettes. StyleID, Z* , and VCT preserve content properties comparable to our method, whereas AttnD and InST induce significant content alterations. Notably, diffusion-based approaches generate high-quality, realistic outputs. While AttnD and InST produce aesthetically appealing results in certain cases, their style-content mismatches often result in artifacts like abnormal facial textures. StyleID, Z* and VCT focus more on color transformation rather than artistic style emulation. Our model achieves superior visual quality and effectively preserves content semantics amidst stylistic transformation.

	Ours	AttnD	StyleID	Z*	VCT	InST	CAST	StyTr^2	AdaAttN	ArtFlow
ArtFID ↓	**22.06**	23.62	26.36	34.29	36.79	29.41	34.37	30.65	29.85	34.79
FID ↓	**17.85**	19.02	19.76	25.58	28.20	23.51	22.23	20.59	20.33	23.16
LPIPS ↓	**0.17**	0.18	0.27	0.29	0.26	0.20	0.48	0.42	0.40	0.44
CLIP ↑	**0.36**	0.34	0.19	0.21	0.14	0.32	0.10	0.14	0.14	0.18

Table 2: Quantitative Comparison.

Quantitative Comparison. Quantitative evaluation of style transfer, though challenging due to limited ground truth and large source-target style gaps, remains critical. Following prior work, we use ArtFID [(1 + LPIPS)·(1 + FID), where LPIPS quantifies content fidelity and FID assesses style fidelity] and CLIP (measuring text-image embedding similarity between outputs and style images). Evaluating 800 representative generated images (Table 2), our method outperforms existing approaches with the lowest ArtFID, FID, LPIPS and highest CLIP score, achieving an optimal balance between content integrity and style fidelity and setting a new benchmark.

User Study. To compare with leading SOTA algorithms, we conducted a user study with 40 participants (18 males, 22 females, 18–40 years). They assessed 30 content-style image pairs, comparing our method with a random SOTA approach across content preservation, style capture, and overall preference. Collecting 1200 votes (Table 3), our method received significantly more votes in all metrics, demonstrating strong balance and alignment with human perception and aesthetics, validating its efficacy and appeal.

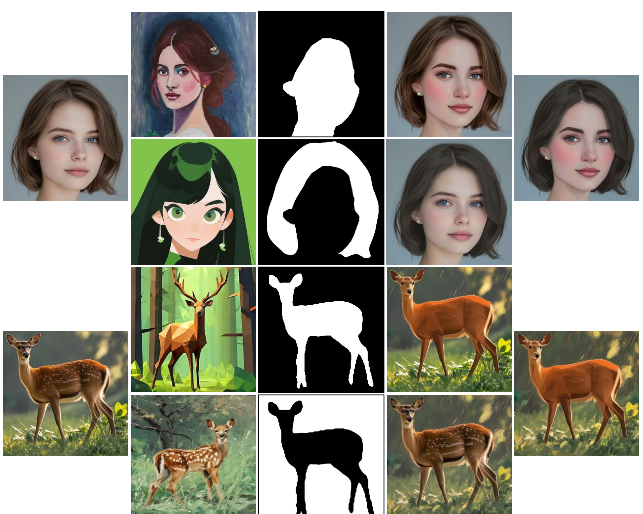

Figure 5: Mask-guided localized style transfer results.

Mask-guided Localized Style Transfer Results

Beyond global style transfer, our method enables mask-guided localized image style transfer, as shown in Figure 5 showcasing two case studies. For each content image, two style targets and corresponding masks are provided, with results including mask-guided and fused outputs. Our approach effectively synthesizes mask-guided results aligned with the mask and style target, leaving other regions unaltered. These localized results differ slightly from full style transfer outcomes. Comparisons reveal our method ensures harmonious generation of content and stylized regions and efficiently enables combinatorial generation of multiple styles across spatial domains.

Multi-style Composition Style Transfer Results

We propose the multi-style composition style transfer using LoRA-switch sampling. Fig. 6 shows two cases. For each case with two style targets A and B, we show style transfer results of A, B, $A + B$ and $B + A$. For single-style stylization, outputs align with stylistic expectations. For style combinations, earlier sampling steps (0-30) influence global stylistic attributes while later steps (31-50) refine local details, enabling nuanced artistic expression. Our method achieves the desired stylization for any combination of two different style targets.

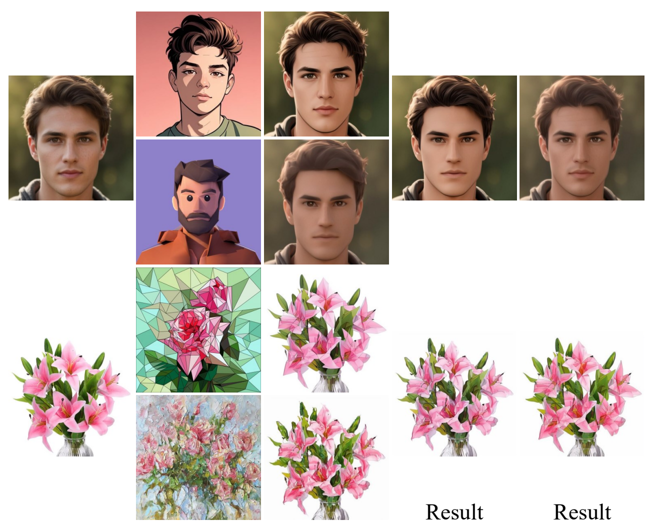

Figure 6: Multi-style composition style transfer results.

Ablation Study

Attention injection layers have been widely studied in prior work(Tumanyan et al. 2023; Couairon et al. 2022; Hertz

	AttnD	StyleID	Z*	VCT	InST	CAST	StyTr^2	AdaAttN	ArtFlow
Content	37%	21%	23%	45%	26%	37%	32%	42%	37%
Style	44%	29%	37%	28%	41%	17%	16%	15%	14%
Overall	42%	25%	34%	25%	30%	11%	15%	13%	16%

Table 3: User Study. The results show the percentage of votes where SOTA method was preferred over ours.

et al. 2022; Zhang et al. 2023a). Our analysis focuses on two key aspects: cross-LoRA feature and attention injection and adaptive attention injection. These analyses validate the effectiveness of our design components.

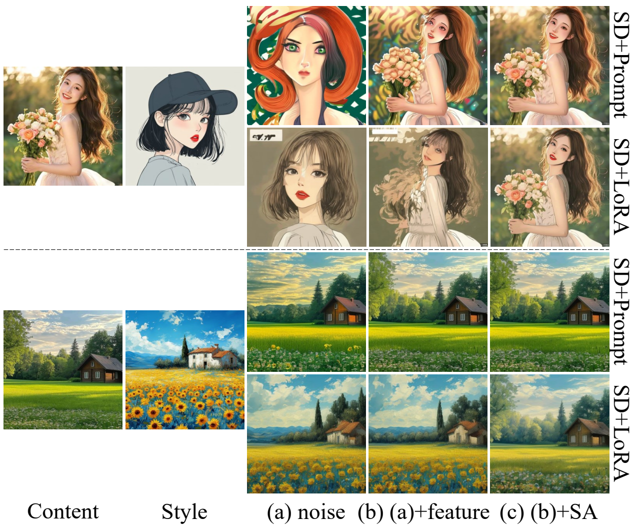

Figure 7: Cross-LoRA feature and attention injection.

Cross-LoRA feature and attention injection. We evaluate our design via two configurations (SD+Prompt, SD+LoRA) with three synthesis combinations: (a) noise initialization, (b) (a)+feature injection, (c) (b)+self-attention (SA) injection. For SD+Prompt, generation is guided by CLIP Interrogator-optimized prompts(Wen et al. 2023); SD+LoRA uses simplified prompts (e.g., "<sss> style, girl.") for training and synthesis. Analyses show noise and feature injection mainly shape global spatial distribution, while SA refines local details. The two configurations yield comparable global structures but differ in local stylistic details. Notably, in portraits, noise-only inputs cause inconsistent content structures in both, yet LoRA ensures target stylistic consistency; progressive feature and attention injection enhances content consistency for more coherent stylizations. In landscapes, optimized prompts fail to fully capture target styles, whereas LoRA enables precise transfer. Our method, integrating noise initialization, feature injection, SA modulation, and LoRA adaptation, excels at preserving semantic content while transferring target stylistic attributes.

Adaptive attention injection. To balance content and style, we propose adaptive attention injection with an empirical threshold. Before the threshold, full attention injection enforces content guidance; beyond it, injection strength diminishes gradually. Fig. 8 demonstrates its effect: partial attention causes severe content loss (style leakage), full attention leads to style absence, while adaptive injection optimally balances content integrity and style fidelity.

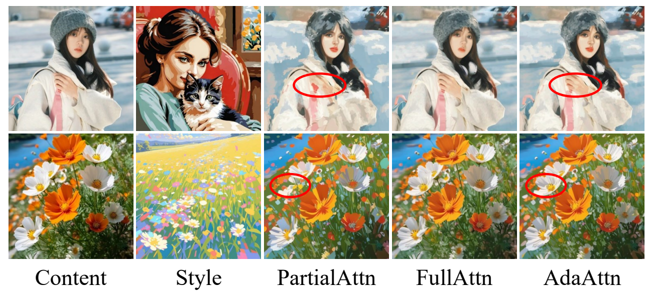

Figure 8: Attn-injection with threshold 30/50 (PartialAttn), 50/50 (FullAttn) and adaptive attention injection (AdaAttn).

Discussion and Limitations

LoRA excels at capturing nuanced target stylistic features. Confining fine-tuning to attention module projection matrices alleviates overfitting, yet challenges persist in preserving content semantic integrity: intricate style-structure entanglement hinders complete decoupling, and LoRA's absorption of stylistic properties may occasionally distort content semantics. Future work will focus on improving content-style decoupling for more reliable style transfer. Building on our approach, we propose two methods: mask-guided localized style transfer and multi-style compositional transfer. Both operate independently, leveraging Stable Diffusion's (SD) stepwise sampling independence to apply to general text-to-image models. Their use for such localized or multi-style text-to-image generation yields desired results.

Conclusion

In this paper, we introduce a novel style transfer framework that captures stylistic attributes from a single reference image via LoRA fine-tuning and guides image synthesis through feature and attention injection. The framework is rooted in a core insight: the inherent cross-LoRA semantic consistency within text-to-image diffusion models. Our approach outperforms state-of-the-art methods by achieving a superior balance between preserving content semantics and transferring stylistic attributes. Building on this insight, we further introduce two innovative techniques: mask-guided localized style transfer and multi-style compositional style transfer. Comprehensive experimental evaluations demonstrate that our approach surpasses existing SOTA techniques.

Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC, No. 62472285 and No. 62102255), Sponsored by CCF-NetEase ThunderFire Innovation Research Funding (NO. CCF-Netease 202508 and NO. CCF-Netease 202509).

References

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

Andonian, A.; Osmany, S.; Cui, A.; Park, Y.; Jahanian, A.; Torralba, A.; and Bau, D. 2023. Paint by Word. arXiv:2103.10951.

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended Diffusion for Text-driven Editing of Natural Images. In *CVPR*. IEEE.

Cheng, B.; Liu, Z.; Peng, Y.; and Lin, Y. 2023. General Image-to-Image Translation with One-Shot Image Guidance. *arXiv preprint arXiv:2307.14352*.

Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; and Yoon, S. 2021. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 14347–14356.

Chung, J.; Hyun, S.; and 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*.

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 *CVPR*, 8795–8805.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2022. DiffEdit: Diffusion-based semantic image editing with mask guidance. arXiv:2210.11427.

Crowson, K. 2023. CLIP-Guided-Diffusion. https://github.com/afiaka87/clip-guided-diffusion. Accessed: 2025-03-15.

Cui, X.; Li, Z.; Li, P. P.; Huang, H.; Liu, X.; and He, Z. 2024. InstaStyle: Inversion Noise of a Stylized Image is Secretly a Style Adviser. In *ECCV*.

Deng, Y.; He, X.; Tang, F.; and Dong, W. 2023. Z^*: Zero-shot Style Transfer via Attention Rearrangement. arXiv:2311.16491.

Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Pan, X.; Wang, L.; and Xu, C. 2022. StyTr ^{2}: Image Style Transfer with Transformers. In *CVPR*.

Du, Y.; Durkan, C.; Strudel, R.; Tenenbaum, J. B.; Dieleman, S.; Fergus, R.; Sohl-Dickstein, J.; Doucet, A.; and Grathwohl, W. 2024. Reduce, Reuse, Recycle: Compositional Generation with Energy-Based Diffusion Models and MCMC. arXiv:2302.11552.

Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A. R.; Narayana, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2023. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *The Eleventh International Conference on Learning Representations*.

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.

Gao, J.; Sun, Y.; Liu, Y.; Tang, Y.; Zeng, Y.; Qi, D.; Chen, K.; and Zhao, C. 2025. Styleshot: A snapshot on any style. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2015. A Neural Algorithm of Artistic Style. *arXiv preprint arXiv:1508.06576*.

Gu, Y.; Wang, X.; Wu, J. Z.; Shi, Y.; Yunpeng, C.; Fan, Z.; Xiao, W.; Zhao, R.; Chang, S.; Wu, W.; Ge, Y.; Ying, S.; and Shou, M. Z. 2023. Mix-of-Show: Decentralized Low-Rank Adaptation for Multi-Concept Customization of Diffusion Models. *arXiv preprint arXiv:2305.18292*.

Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D.; and Yang, F. 2023. SVDiff: Compact Parameter Space for Diffusion Fine-Tuning. arXiv:2303.11305.

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

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Huang, N.; Dong, W.; Zhang, Y.; Tang, F.; Li, R.; Ma, C.; Li, X.; and Xu, C. 2024. CreativeSynth: Creative Blending and Synthesis of Visual Arts based on Multimodal Diffusion. arXiv:2401.14066.

Huang, Z.; Chan, K. C.; Jiang, Y.; and Liu, Z. 2023. Collaborative Diffusion for Multi-Modal Face Generation and Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jeong, J.; Kim, J.; Choi, Y.; Lee, G.; and Uh, Y. 2024. Visual Style Prompting with Swapping Self-Attention. arXiv:2402.12974.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image Generation from Scene Graphs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1219–1228.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *CVPR*, 2426–2435.

Kingma, D. P.; and Welling, M. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114.

Kong, Z.; Zhang, Y.; Yang, T.; Wang, T.; Zhang, K.; Wu, B.; Chen, G.; Liu, W.; and Luo, W. 2024. OMG: Occlusion-friendly Personalized Multi-concept Generation in Diffusion Models. arXiv:2403.10983.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-Concept Customization of Text-to-Image Diffusion.

Kwon, G.; Jenni, S.; Li, D.; Lee, J.-Y.; Ye, J. C.; and Heilbron, F. C. 2024. Concept Weaver: Enabling Multi-Concept Fusion in Text-to-Image Models. arXiv:2404.03913.

Li, S.; Du, Y.; Tenenbaum, J. B.; Torralba, A.; and Mordatch, I. 2022. Composing Ensembles of Pre-trained Models via Iterative Consensus. arXiv:2210.11522.

Li, S.; Ma, C.; and Pan, Y. 2025. DiffPortraitVideo: Diffusion-based Expression-Consistent Zero-Shot Portrait Video Translation. *IEEE Transactions on Visualization and Computer Graphics*, 1–12.

Lin, Z.; Chen, J.; Ma, X.; Li, C.; Zhang, H.; and Zhao, L. 2024. Unsupervised content and style learning for multimodal cross-domain image translation. *Scientific Reports*, 14(1): 29469.

Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Sun, Z.; Li, Q.; and Ding, E. 2021. AdaAttN: Revisit Attention Mechanism in Arbitrary Neural Style Transfer. In *Proceedings of the IEEE International Conference on Computer Vision*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Gool, L. V. 2022. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. arXiv:2201.09865.

Lv, C.; Zhang, D.; Geng, S.; Wu, Z.; and Huang, H. 2024. Color transfer for images: A survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(8): 1–29.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *ICLR*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. arXiv:2112.10741.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv*, abs/1505.04597.

Singh, J.; Gould, S.; and Zheng, L. 2023. High-Fidelity Guided Image Synthesis with Latent Diffusion Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5997–6006.

Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. arXiv:1503.03585.

Song, J.; Meng, C.; and Ermon, S. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*.

Tan, S.; Ji, B.; and Pan, Y. 2024. Style2Talker: High-Resolution Talking Head Generation with Emotion Style and Art Style. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5079–5087.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In *CVPR*, 1921–1930.

von Platen, P.; Patil, S.; Lozhkov, A.; Cuenca, P.; Lambert, N.; Rasul, K.; Davaadorj, M.; Nair, D.; Paul, S.; Berman, W.; Xu, Y.; Liu, S.; and Wolf, T. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

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.; 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.; Wang, X.; Xie, L.; Qi, Z.; Shan, Y.; Wang, W.; and Luo, P. 2023. StyleAdapter: A Unified Stylized Image Generation Model. *arXiv preprint arXiv:2309.01770*.

Wen, Y.; Jain, N.; Kirchenbauer, J.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2023. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36: 51008–51025.

Yang, R.; Wu, X.; and He, S. 2024. MixSA: Training-free Reference-based Sketch Extraction via Mixture-of-Self-Attention. *IEEE Transactions on Visualization and Computer Graphics*, 1–16.

Yang, Z.; Liu, D.; Wang, C.; Yang, J.; and Tao, D. 2022. Modeling image composition for complex scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7764–7773.

Zhang, K.; Zhou, Y.; Xu, X.; Pan, X.; and Dai, B. 2023a. DiffMorpher: Unleashing the Capability of Diffusion Models for Image Morphing. *arXiv preprint arXiv:2312.07409*.

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.; Huang, N.; Tang, F.; Huang, H.; Ma, C.; Dong, W.; and Xu, C. 2023b. Inversion-based Style Transfer with Diffusion Models. In *CVPR*, 10146–10156.

Zhang, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Lee, T.-Y.; and Xu, C. 2022. Domain Enhanced Arbitrary Image Style Transfer via Contrastive Learning. In *ACM SIGGRAPH*.

Zhou, Y.; Gao, X.; Chen, Z.; and Huang, H. 2025. Attention Distillation: A Unified Approach to Visual Characteristics Transfer. arXiv:2502.20235.

Zhou, Z.; Tang, F.; Zhang, Y.; Deussen, O.; Cao, J.; Dong, W.; Li, X.; and Lee, T.-Y. 2024. A Comprehensive Evaluation of Arbitrary Image Style Transfer Methods. *IEEE Transactions on Visualization and Computer Graphics*.