AnyStyle: A Single LoRA is Sufficient for Image-Guided Style Transfer

Yongwen Lai and Chaoqun Wang*

School of Artificial Intelligence, South China Normal University, Guangzhou, China

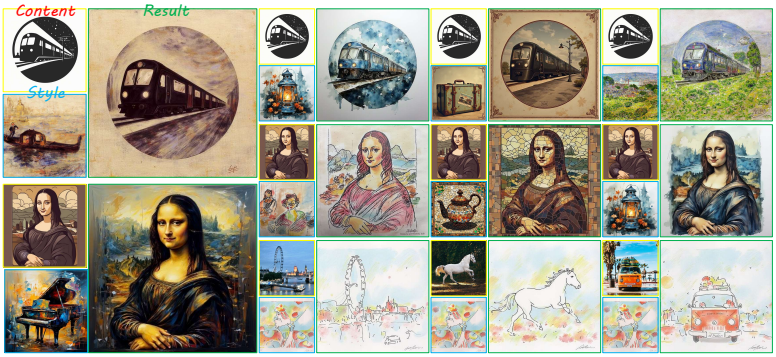

Fig. 1: Style transfer results of our method. In each set of images, the top-left image denotes the content reference, the bottom-left image denotes the style reference, and the image on the right presents the generated results. The results demonstrate that our method consistently achieves both content preservation and style alignment across diverse inputs, including real-world photographs and painted images.

Abstract. Image-guided style transfer aims to apply the artistic characteristics of a style image to a content image while preserving its semantic structure and layout. Despite advances in diffusion-based methods, existing approaches often face challenges in disentangling content and style, particularly when independently optimized adapters are naively combined, causing conflicts between adapters and limiting controllability over the content-style balance in inference. We further demonstrate that training-free structural guidance directly derived from the content image through the internal attention of pre-trained model outperforms a dedicated content LoRA adapter in terms of structural fidelity and computational efficiency. Building on these observations, we propose AnyStyle, a streamlined framework for image-guided style transfer. The framework adopts a unified single-adapter paradigm for coherent style capture from the style image and incorporates training-free structural guidance from the content image, thus avoiding complex entanglement between multiple adapters and improving controllability and stability. Extensive experiments show that our method delivers competitive quantitative performance and significantly improved perceptual quality. Code is available at `https://github.com/Yvan1001/AnyStyle`.

Keywords: Style Transfer · Low-Rank Adaptation · Rectified Flow

* Corresponding author.

1 Introduction

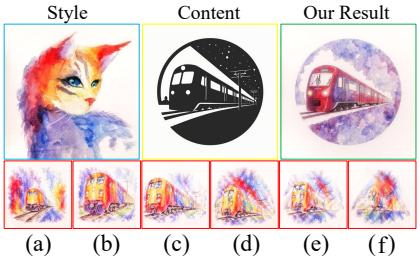

Fig. 2: The first row shows the input and the output of our method. Columns (a–f) present the results of the dual-LoRA approach under different fusion coefficients during inference(Adjust the coefficients to shift from style dominance to content dominance). As illustrated, the results remain unsatisfactory across different coefficient settings, achieving a balanced trade-off with the dual-LoRA method is difficult.

Image-guided style transfer [7, 10, 21, 26, 36] aims to apply the artistic characteristics of a style reference image to a content reference image while preserving as faithfully as possible the semantic structure and layout of the content reference image. This task holds significant value and has broad applications in digital art creation, image editing, and cross-modal content generation. With the evolution of deep learning, early methods [13,22,25,27] based on convolutional neural networks have achieved pioneering results by iteratively minimizing a joint content-style loss. However, these approaches exhibit strong dependency on data distributions specific to particular styles, and collecting high-quality datasets for style transfer remains challenging in practice. Subsequent advances in generative adversarial networks and diffusion models substantially improve generation quality and generalization. Modern approaches built upon large-scale pre-trained text-to-image diffusion models [9, 24] now represent the dominant paradigm for style transfer.

In recent years, the fine-tuning technique Low-Rank Adaptation (LoRA) has been widely adopted for personalized style generation. Taking advantage of this convenience, a prominent line of work [2, 10, 32, 56] employs a dual-LoRA strategy for the style transfer task. In this approach, separate LoRA weights are independently fine-tuned for content and style and then combined via simple weight fusion to denoise in inference. Although such methods advance the field, our empirical analysis reveals several inherent limitations. 1) Independently optimized adapters often cause complex entanglement between content and style adapters. Their fusion requires delicate tuning of mixing coefficients, rendering the content-style trade-off difficult, as illustrated in Fig. 2. Meanwhile, although modern diffusion models are often described as exhibiting a layered processing pattern, in which earlier layers encode low-level structural information and later layers emphasize higher-level semantics and style, such a division is only partially valid in practice. Feature representations of content and style in diffusion models remain inherently entangled across layers, which challenges the assumption that content and style can be reliably disentangled through a dual-LoRA design. 2) A dedicated content LoRA typically underperforms direct utilization of training-free structural guidance derived from the content reference image through the internal attention of the pre-trained model in structural fidelity and perceptual quality, while adding fine-tuning overhead. Under these circum-

stances, training-free structural guidance thus provides a more direct, efficient, and effective means of content preservation in the style transfer task.

Building upon these findings, we propose AnyStyle, an image-guided style transfer framework that combines a single-style LoRA with an attention modulation mechanism. The complete framework consists of two stages: During fine-tuning, we adopt a single-adapter design in place of the dual-LoRA paradigm, fine-tuning only one LoRA targeted at layers more suitable for encoding style, thereby effectively enhancing the ability to generate consistent style. During inference, training-free structural guidance is extracted from the content reference image and dynamically incorporated to reliably preserve content layout and structure. This design effectively avoids the complex entanglement and uncertainty introduced by multi-adapter fusion and achieves improved coordination among textual-prompt adherence, content fidelity, and style consistency. The main contributions of our work are as follows:

- Through experimental analysis, we uncover key limitations of the dual-LoRA strategy in style transfer and validate the superiority of training-free structural guidance from content reference image for content preservation.
- We introduce a simple yet effective pipeline that integrates a single LoRA with attention modulation and enhances the controllability and stability of style transfer.
- We demonstrate competitive performance against existing methods in both quantitative metrics, visual results, and user study evaluations.

2 Related Work

2.1 Style Transfer

Image-guided style transfer [8, 16] is a classic task, which aims to apply the artistic style of a style reference image to a content reference image while preserving the semantic structure of the object and layout in the content reference image as faithfully as possible. The early approaches are primarily based on convolutional neural networks [13, 22, 23, 25, 27, 28, 34], achieving a content-style balance through iterative optimization of a combined loss function. These methods mark a significant breakthrough in the style transfer task, but suffer from strong dependency on datasets and limit output diversity. With the advent of deep generative models, many methods [5, 12, 37] based on generative adversarial networks have emerged. These approaches substantially improve generation quality and generalization. However, they typically require large domain-specific datasets, incur high retraining costs, and are prone to mode collapse or structural distortion in semantically complex scenes.

In recent years, diffusion models [17, 43, 48] have fundamentally transformed the paradigm of style transfer by leveraging the powerful generative priors of pre-trained text-to-image models such as Stable Diffusion 1.5 [43], SDXL [40]. In more recent advances, diffusion model backbones have transitioned from traditional U-Net [44] to more powerful Transformer [50]. Models including DiT [39],

SD3 [9], and FLUX [24] have enabled diffusion models to scale toward high-resolution and semantically complex image generation. By exploiting the rich prior knowledge embedded in pre-trained diffusion models, recent methods investigate image stylization and editing by modifying the generation process through fine-tuning and training-free methods. For example, InST [60] proposes a textual inversion-based method that encodes a target style into learned textual embeddings. StyleDiffusion [54] seeks to separate style from content by incorporating a CLIP-based [41] style disentanglement loss during the fine-tuning of diffusion models for style transfer. Style-Friendly [4] aggressively shifts the signal-to-noise ratio distribution toward higher noise levels during fine-tuning to focus on noise levels where stylistic features emerge. During the same period, many training-free methods [1, 15, 51] emerge. Specifically, Only-Style [1] and StyleAligned [15] achieve style consistency across a sequence of generated images by combining attention feature sharing with the AdaIN [19] mechanism. However, these methods do not explicitly incorporate content control, which may lead to the loss of structural information from the content reference image.

2.2 Attention and Structural Guidance

In image generation tasks, guiding the sampling process with attention and structural information is a widely adopted strategy. Methods such as Ctrl-X [31] achieve structural control via a dedicated module but require clean structural input, imposing restrictive assumptions in practice. FreeControl [30] performs one-step attention extraction from a single, optimally chosen key timestep and reuses it throughout denoising to enable efficient structural guidance. P2P [14] injects cross-attention maps from the reconstruction branch into the editing branch. RF-Solver-Edit [53] and PnP [49] reuse values from self-attention layers of the source inversion. Training-based methods such as ControlNet [58] and T2I-Adapter [35] leverage condition maps to guide spatial structure, achieving strong low-level alignment but requiring retraining per condition and per model, which incurs substantial computational cost and limits robustness to noisy inputs. Higher-level approaches, including GLIGEN [29] and IP-Adapter [57], enable layout-aware or visual conditioning, yet still depend on specific training and provide limited fine-grained structural control. In addition, a growing number of studies [3, 6, 52, 55, 61] have investigated methods to improve content preservation by enforcing structural and semantic consistency during generation. Previous methods tend to emphasize either content preservation or style consistency, making it challenging to achieve a balanced coordination of both.

2.3 LoRA-based Image Style Transfer

Early work such as DreamBooth [45] initiate this line of research by fine-tuning diffusion models on a small set of subject-specific images to encode identity information, but at the cost of substantial computational overhead and increased

risk of overfitting. To address these limitations, LoRA [18] enables parameter-efficient fine-tuning by applying low-rank updates to selected model layers, significantly reducing memory and computation while maintaining generation quality. Consequently, LoRA-based methods [2, 10, 32, 46, 56] have emerged as the predominant approach for style transfer, supporting a wide range of applications. Specifically, ZipLoRA [46] enables flexible subject–style composition by fusing independently trained LoRA weights (one for content, one for style) through column-wise merging coefficients, but each fusion still requires additional optimization, limiting reuse. B-LoRA [10] leverages block specialization in the SDXL to achieve implicit content–style disentanglement via joint optimization on selected blocks, although coarse block partitioning often leads to subject leakage. UnZipLoRA [32] further improves disentanglement by introducing prompt, column, and fine-grained block separation. Despite these advances, challenges remain in generating stylized images that preserve content structure and align with the desired style.

In our work, we adopt only one LoRA targeted at layers more suitable for encoding style as the foundation for achieving style consistency. Building upon this, we guide the sampling process using attention modulation that incorporates training-free structural guidance from a content reference image, thereby enabling collaborative stylization that jointly preserves content and style.

3 Preliminaries

3.1 Rectified Flow Model

Rectified flow [33] is a continuous generative framework that facilitates smooth and efficient transport between the image data distribution X and a Gaussian distribution $N \sim \mathcal{N}(0, I)$. Unlike traditional diffusion models that rely on stochastic multi-step transitions between noise and data, rectified flow establishes a deterministic and nearly linear trajectory for mapping noise to data, which is governed by an ordinary differential equation (ODE) parameterized by a neural network θ. The intermediate latent state at continuous timestep $t \in [0, 1]$ is formulated as a linear interpolation:

$$X_t = (1 - t) \cdot X + t \cdot N, \tag{1}$$

where $t = 1$ corresponds to pure noise and $t = 0$ denotes the clean data domain. During inference, starting from a Gaussian noise sample, the pre-trained model iteratively predicts the velocity v_θ at each discrete timestep and updates the latent state accordingly. Specifically, the update of the latent state Z_t follows the recurrence relation:

$$Z_{t_{i-1}} = Z_{t_i} + (t_{i-1} - t_i) v_\theta(Z_{t_i}, t_i). \tag{2}$$

After completing the iterative denoising process, the final latent state is decoded into the pixel space through a decoder to obtain the generated image.

3.2 FlUX Architecture and LoRA Fine-tuning

FLUX Architecture. FLUX [24] is a state-of-the-art text-conditional generative framework built upon the rectified flow paradigm and the DiT [39] backbone. Its core architecture comprises 19 double-stream transformer blocks and 38 single-stream transformer blocks, which collaboratively process multimodal inputs. The input text prompt is encoded via CLIP [41] and T5 [42], ensuring that semantic guidance is effectively embedded into the generation process.

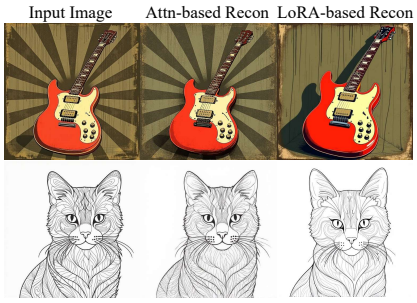

Fig. 3: Reconstruction. Attention-based reconstruction preserves structure and perceptual quality with a single inference, while LoRA-based requires extra fine-tuning and yields lower fidelity.

LoRA Fine-tuning. To efficiently adapt such large-scale pre-trained models to task-specific scenarios like style transfer, LoRA [18] is a parameter-efficient fine-tuning strategy that avoids full model retraining. For a pre-trained linear transformation layer, its weight matrix is denoted as $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ where d_{in} and d_{out} denote the dimensions of the input and output of the layer. This matrix maps an input vector $x \in \mathbb{R}^{d_{\text{in}}}$ to an output vector $y \in \mathbb{R}^{d_{\text{out}}}$ following the linear transformation $y = W_0 x$. LoRA introduces a low-rank residual update to capture task-specific adjustments, with the effective fine-tuned weight matrix and its low-rank decomposition jointly expressed as:

$$W = W_0 + \Delta W, \quad \Delta W = BA, \tag{3}$$

where W_0 is the original frozen pre-trained weight matrix, ΔW is the low-rank update introduced by LoRA fine-tuning. The rank of the low-rank decomposition r is defined as a small value typically satisfying $r \ll \min(d_{\text{in}}, d_{\text{out}})$ to minimize the number of trainable parameters. The matrices $A \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B \in \mathbb{R}^{d_{\text{out}} \times r}$ form the low-rank factorization of ΔW, and W denotes the effective weight matrix used during inference. During training, only the low-rank matrices A and B are updated, while the original pre-trained weights W_0 remain frozen. This drastically reduces the number of trainable parameters and mitigates the risk of overfitting. During inference, the low-rank updates can be merged into the original weight matrices without incurring additional computational overhead, ensuring efficient deployment.

4 Method

Existing style transfer methods typically adopt a dual-LoRA paradigm, separating content and style adaptation across different layer ranges. However, such designs rely on imperfect layer-wise disentanglement and often require delicate weight balancing due to the complex coupling relationship between content and style. To address these limitations, we propose a streamlined single-LoRA framework with attention modulation termed AnyStyle for image-guided style transfer.

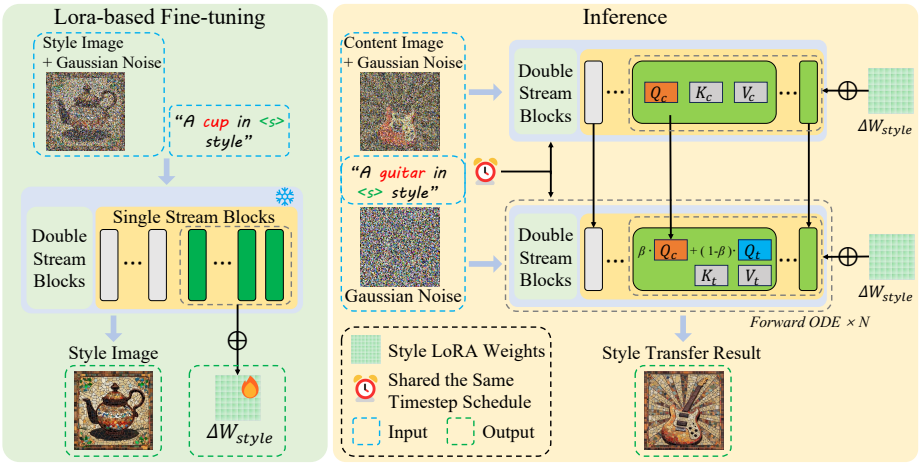

Fig. 4: Overview of the AnyStyle framework. Given a content reference image and a style reference image, together with a prompt such as "A <c> in <s> style", AnyStyle produces stylized results that preserve the structural integrity of the original object while ensuring stylistic consistency. The left panel illustrates the fine-tuning stage, whereas the right panel presents the inference stage. During inference, the style LoRA weights are integrated into the corresponding DiT blocks to enforce consistent style generation. In addition, an attention modulation module is employed to guide structural features, enabling effective style transfer while maintaining content fidelity.

4.1 Analysis of Dual-LoRA for Style Transfer

FLUX [24], based on a DiT [39] architecture, features double-stream and single-stream transformer blocks in its core denoising component. Through extensive experimentation and analysis, we observe that early single-stream transformer blocks primarily capture low-level structural details, whereas later blocks tend to encode higher-level semantic and stylistic attributes. Consistent with the above findings, many prior works [2,10,32,56] adopt dual-LoRA architectures for style transfer by fine-tuning separate LoRA adapters [18] for content and style in different layer ranges, and subsequently fuse the learned weights.

To evaluate the effectiveness of the content LoRA in the dual-LoRA strategy for the style transfer task, we analyze its assumptions and limitations by comparing two content reconstruction strategies. The first strategy fine-tunes content LoRA weights on early single-stream blocks (1–10) using a single reference image and applies them for reconstruction from noise. The second adopts an attention-based approach that adds noise to the content reference image, records query, key, and value tensors during denoising, and uses them to guide reconstruction from random noise without any LoRA weights. The attention-based method achieves superior structural fidelity and perceptual quality while eliminating additional fine-tuning overhead. The corresponding quantitative results are reported in Tab. 1. Detailed visual comparisons of the reconstruction

Table 1: Quantitative comparison of image reconstruction under two strategies. In this context, "Attn-based" refers to the approach that guides image reconstruction using the recorded query, key, or value tensors, whereas "LoRA-based" refers to the approach that guides image reconstruction using the fine-tuned LoRA weights.

Reconstruction	PSNR(↑)	DINO-Content(↑)	DS-Content(↓)	MSE(↓)	LPIPS(↓)
Attn-based	18.0029 ↑	0.9743 ↑	0.0617 ↓	0.0754 ↓	0.2438 ↓
LoRA-based	11.7899	0.9567	0.1273	0.3137	0.4553

can be found in Fig. 3. From the visualization, the attention-based approach demonstrates significantly superior reconstruction quality while requiring only a single inference process. In contrast, the LoRA-based approach not only produces lower reconstruction fidelity but also incurs additional time for LoRA fine-tuning. These results suggest that content LoRA weights provide limited value for content preservation. Recording query, key, and value tensors during denoising captures more precise structural information than fine-tuning a content LoRA module, making content LoRA largely redundant. Furthermore, dual-LoRA architectures suffer from complex entanglement: independently optimized content and style adapters remain coupled in single-stream blocks, and their fusion requires careful coefficient tuning, often leading to suboptimal trade-offs.

4.2 AnyStyle for Style Transfer

Motivated by these observations, we propose AnyStyle, a streamlined and effective framework for style transfer that employs only a single layer-specific LoRA for style alignment combined with a complementary attention modulation for content preservation. This design can avoid the complexities and instability of dual-LoRA fusion while achieving robust and predictable control over style and content. The overall pipeline is illustrated in Fig. 4.

For style alignment, we apply LoRA exclusively to the later single-stream transformer blocks (11–38) in FLUX, targeting the projection matrices of self-attention layers (query, key, value, and output projections), where corresponding low-rank updates are optimized to improve the ability to generate specific styles. Specifically, by differentiating the interpolation formula in Eq. (1) with respect to timestep t, the underlying flow ODE can be derived as:

$$\frac{dX_t}{dt} = X - N, \tag{4}$$

which defines the target velocity field that the neural network with style LoRA weights is designed to approximate, and the objective of fine-tuning is to minimize the mean squared error between the predicted velocity $v_{\theta,\Delta W_{style}}(X_t, t)$ and the target velocity across the entire time interval, expressed as:

$$\min_{\theta} \int_0^1 \mathbb{E}\left[\left\|(X - N) - v_{\theta,\Delta W_{style}}(X_t, t)\right\|_2^2\right] dt. \tag{5}$$

Algorithm 1 Full AnyStyle Algorithm

Input: $Z^c, Z^s, \{t_i\}_{i=0}^{T}, n, \beta_{max}, \beta_{min}$
Output: Style Transfer Result Z_0
$\Delta W \leftarrow \text{DreamBooth}(Z^s)$; $\Theta \leftarrow \theta + \Delta W$
Initialize $Z_T \sim \mathcal{N}(0, I)$; $\epsilon \sim \mathcal{N}(0, I)$
for $i = n$ to 1:
  $\beta \leftarrow \beta_{max} \cdot (t_i/T) + \beta_{min} \cdot (1 - t_i/T)$
  $Z^c_{t_i} \leftarrow (1 - t_i) \cdot Z^c + t_i \cdot \epsilon$
  $V^c_{t_i} \leftarrow V_\Theta(Z^c_{t_i}, t_i) \implies$ extract Q_c; $\quad V_{t_i} \leftarrow V_\Theta(Z_{t_i}, t_i, Q_c)$
  $Z_{t_{i-1}} \leftarrow Z_{t_i} - \Delta t \cdot V_{t_i}$
return Z_0

Each LoRA module is fine-tuned independently following the DreamBooth [45] objective, which reconstructs the reference image from Gaussian noise conditioned on a textual prompt. This single, targeted LoRA efficiently infuses stylistic priors, including colors, textures, and artistic patterns, into the layers where style representations naturally dominate. In inference, the fine-tuned style LoRA weights are loaded into the AnyStyle pipeline and integrated with standard sampling. By replacing the dual-LoRA strategy with this single-LoRA plus attention modulation framework, we mitigate complex entanglement and unstable trade-offs between content and style arising from the simple fusion of independently optimized adapters, supporting flexible stylization of arbitrary content-style pairs and allowing the target result to consistently maintain the intended structure while effectively applying the desired stylistic characteristics.

4.3 Structural Guidance with Attention Modulation

For content preservation, methods that rely solely on fine-tuning an additional content adapter under the dual-LoRA paradigm frequently result in distorted layouts or loss of object structure when applying aggressive style changes.

To overcome this limitation, we design an attention modulation that directly extracts training-free structural guidance from the content image itself, avoiding separate training or external adapters. The key is to leverage the inherent attention dynamics of the content image under controlled noise. Specifically, we introduce an auxiliary denoising path: at each timestep t, the input latent of the DiT model Z^c_t is $(1-t) \cdot Z^c + t \cdot N$, where Z^c represents the latent of the content image. During the denoising process, attention tensors from self-attention layers are recorded and indexed by timestep and block, and then selectively mixed into the main generation path. For each attention call at the corresponding timestep and block, the content query Q_c is retrieved and blended with the current query Q_t according to:

$$Q_t = \beta \cdot Q_c + (1 - \beta) \cdot Q_t, \tag{6}$$

where β controls the strength of the structural guidance. Key and value tensors remain unchanged to prevent any stylistic leakage from the content reference

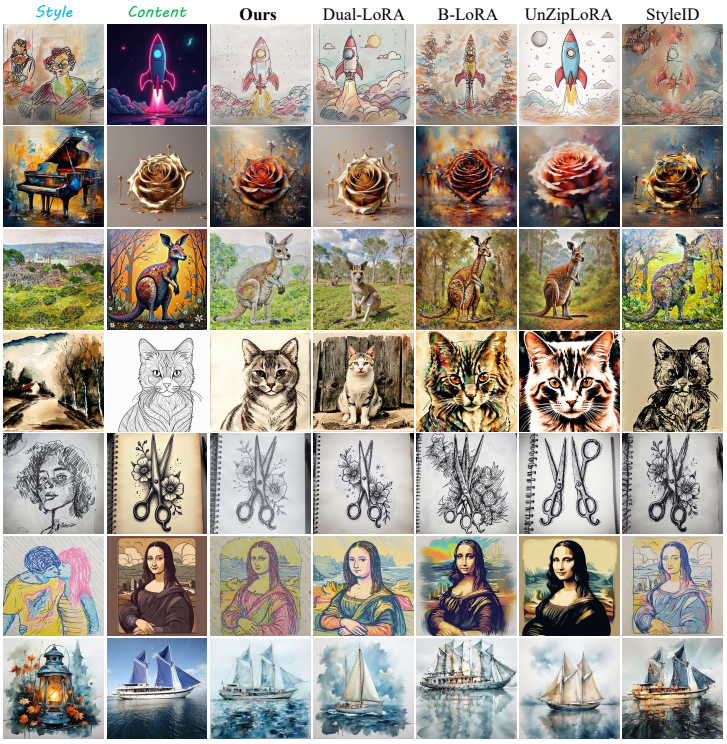

Fig. 5: Qualitative comparisons. The first and second columns correspond to the style reference image and the content reference image, respectively. The subsequent columns present the visual results obtained using different style transfer methods.

image. Ablation studies confirm that modulating only the query tensors is sufficient for effective content preservation, as queries primarily encode "what to attend to", such as layouts and objects, whereas modifying key or value tensors tends to introduce unwanted stylistic elements from the content reference image. Through thorough experimentation, this selective approach provides precise structural guidance without compromising style alignment. To ensure adaptive guidance across the denoising process, we apply time-dependent weighting:

$$\beta = \beta_{\max} \cdot (t/T) + \beta_{\min} \cdot (1 - t/T), \tag{7}$$

where T denotes the first number in denoising steps, $\beta_{\max}$ and $\beta_{\min}$ denote the maximum and minimum modulation strengths applied at the beginning and end of the denoising process, respectively. Actually, systematically decreasing β values from 1 to 0 results in a smooth and continuous transition from heavily content-preserving visual generations to style-dominant outputs. For details, please refer to the supplementary materials. This design delivers stronger modulation in early timesteps, where noise is dominant and the content structure is most vulnerable, and weaker modulation in later timesteps, which allows stylistic refinement once the core content structure has been established. The modula-

Table 2: Quantitative comparison across different style transfer methods. The style transfer results are assessed separately against the content reference image and the style reference image, yielding corresponding content and style scores. In addition, a text consistency score is computed to evaluate whether the generated images are consistent with the corresponding prompt.

Method	Prompt Align	Content Align		Style Align	
	CLIP-T(↑)	DS(↓)	DINO(↑)	CLIP(↑)	DINO(↑)
B-LoRA [10]	31.6738	0.4642	0.8367	<u>65.7344</u>	<u>0.6777</u>
UnZipLoRA [32]	<u>31.8346</u>	0.4230	0.8370	61.9083	0.6567
StyleID [6]	30.5093	**0.3179**	**0.8849**	64.9011	0.6634
Dual-LoRA	30.5151	0.5038	0.7947	63.3469	0.6874
AnyStyle (Ours)	**32.5573**	<u>0.4210</u>	<u>0.8382</u>	**67.9291**	**0.7147**

tion is applied across all 38 single-stream transformer blocks to achieve holistic and consistent content preservation. These timestep-dependent attention signals then dynamically guide the main generation path, preserving content structure while fully respecting the style imposed by the style LoRA adapter. Algorithm 1 summarizes the proposed Anystyle. Detailed algorithmic procedures and variable definitions are provided in the supplementary materials.

5 Experiments

5.1 Implementation Details

Datasets. A total of 40 images are collected for the experiments from B-LoRA [10], StyleDrop [47], and EditEval [20]. These images are evenly divided into two subsets, namely content reference images and style reference images, with 20 images in each subset. Each content reference image is paired with each style reference image, resulting in 400 distinct style transfer combinations.

Experimental Setup. Our implementation is based on FLUX [24] model, with the model weights, text encoders, and VAE frozen. During the fine-tuning process, the rank of style LoRA weights is set to 64. We use a learning rate of $5e-5$ and train for 200 steps on a single style reference image. The entire training process takes approximately 3 minutes on a single GPU. For inference, we set the total number of timesteps to 28, with target prompt guidance values of 5. The hyperparameters are set as follows: the maximum and minimum modulation strengths $\beta_{\max}$ and $\beta_{\min}$ are set to 0.7 and 0.3, respectively. All experiments are conducted on an NVIDIA RTX 5880 Ada GPU with 48GB of memory.

Evaluation Metrics. Following previous studies [2,32,56], we adopt CLIP [41], PSNR, LPIPS [59], MSE, DINO [38] and DS [11] as our evaluation metrics. Specifically, the evaluation scores are decomposed into content and style components, and the generated style transfer results are assessed separately against the content reference image and the style reference image, yielding the corresponding content and style scores. In addition, a score is set to evaluate the text consistency, reflecting the semantic matching between prompt and image result.

Table 3: Quantitative comparison of image generation results with and without loading the style LoRA weights. All other experimental settings are kept identical. The results demonstrate the impact of style-specific adaptation on style alignment.

Method	CLIP-T(↑)	CLIP-Style(↑)	DINO-Style(↑)
w/ Style LoRA	32.0717↑	68.2336↑	0.7234↑
w/o Style LoRA	29.9431	57.3314	0.6170

5.2 Results

Comparison with Existing Methods. We compare AnyStyle with recent state-of-the-art methods for image-guided style transfer. The quantitative results presented in Tab. 2 demonstrate that our approach achieves strong performance in most of the evaluation metrics. Although StyleID [6] attains the highest score in content alignment, it performs worst in prompt alignment and style alignment, which is consistent with the observations reported in ConsisLoRA [2]. Excluding this particular case, our method surpasses all baseline approaches in prompt, style, and content alignment. To validate the effectiveness of the proposed single-LoRA strategy, we conduct a comparative study against a dual-LoRA approach implemented on FLUX. The quantitative results are presented in the fourth method of Tab. 2, together with the qualitative comparisons shown in Fig. 5, which indicate that our method outperforms the "Dual-LoRA" baseline in terms of both content consistency and stylistic coherence. These results validate the effectiveness of the proposed single-LoRA strategy and attention modulation mechanism. The qualitative comparisons shown in Fig. 5 further support these conclusions, indicating that our approach achieves higher semantic precision in both content and style, resulting in more natural and coherent visual outcomes. More visualizations provided in Fig. 6.

Fig. 6: Additional visual examples of style transfer of our method.

User Study. To complement quantitative evaluation, we conduct a user study assessing the perceptual quality of AnyStyle against competing methods. A total of 70 participants are presented with the content reference image, style reference image, together with editing results from different methods displayed in randomized order. For each case, participants are asked to select the most visually satisfactory result, with a "Not Sure" option

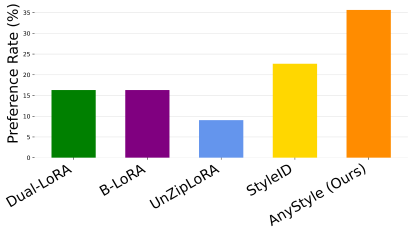

Fig. 7: User study. Our AnyStyle receives the highest number of user selections.

Fig. 8: Standard image generation vs. AnyStyle generation with style LoRA weights. The first column of each group is the style reference image. Under the same prompt "A <c> in <s> style", standard sampling without style LoRA (red boxes) yields diverse but stylistically inconsistent results, whereas sampling with the proposed style LoRA (green boxes) produces images with consistent style across different contents.

provided to avoid forced choices. Each participant evaluated approximately 10 sets. As shown in Fig. 7, AnyStyle receives the highest preference rate.

5.3 Ablation Study

Generation of Style Consistency. To evaluate the effectiveness of the fine-tuned style LoRA weights, we design a controlled comparative experiment. Under an identical textual prompt, images are generated using two configurations: one with the style LoRA weights loaded and one without loading the style LoRA weights. All other model parameters and inference settings are kept fixed to ensure a fair comparison. The generated results are subsequently compared with the corresponding style reference image. As shown in Tab. 3, the configuration with the style LoRA weights achieves substantially higher stylistic consistency with the style reference image, and a higher text consistency score. This conclusion is further supported by the visual comparisons in Fig. 8, where the configuration with style LoRA weights loaded exhibits more consistent and stable style.

Analysis of Structural Guidance Strength. The modulation parameters β_{min} and β_{max} jointly determine the structural control capacity. We conduct an exhaustive parametric sweep to analyze their impact on synthesis characteristics. As visualized in Fig. 9, systematically decreasing β values results in a smooth and continuous transition from heavily content-preserving visual generations to style-dominant outputs. This indicates that AnyStyle offers a highly controllable

Table 4: Quantitative comparison under different attention modulation.

Method	Prompt Align	Content Align		Style Align	
	CLIP-T(↑)	DS(↓)	DINO(↑)	CLIP(↑)	DINO(↑)
Query	**32.5573**	**0.4210**	**0.8382**	67.9291	0.7147
Key	32.2734	0.5198	0.7827	**68.4408**	**0.7225**
Query + Key	32.4606	0.4384	0.8333	67.1565	0.7034

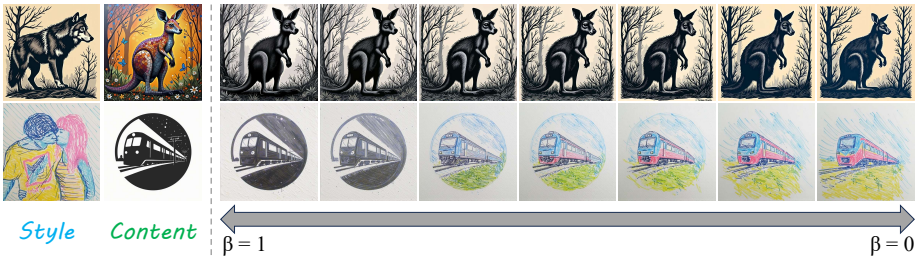

Fig. 9: Parametric ablation study of the structural guidance strength $\beta_{\max}$ and $\beta_{\min}$.

framework capable of modulating structural preservation boundaries according to diverse user preferences.

Effect of Attention Modulation. To investigate the influence of different attention modulation strategies on the final performance, we conduct a series of controlled experiments. Based on the single-LoRA framework, various attention modulation schemes are introduced during the denoising process. As reported in Tab. 4, applying modulation solely to the key tensors yields relatively high style alignment scores but leads to substantially degraded content preservation. When both query and key tensors are modulated, content-related metrics improve compared to key-only modulation, yet the gains in style alignment remain limited. In contrast, modulating only the query tensors achieves the best overall performance across evaluation metrics. This outcome is consistent with our expectations. Query-only modulation maximally exploits structural guidance, effectively preserving content structure. At the same time, it restricts the propagation of extraneous information from the content reference image into the main branch, thereby preventing unintended stylistic leakage and resulting in superior style transfer performance. Similarly, as illustrated in Fig. 10, query-only modulation demonstrates the most favorable trade-off between content preservation and style alignment.

Failure Cases. Typical failure cases tend to manifest when the pre-trained model fails to reliably recognize highly complex and non-standard semantic structures. For instance, when encountering uncommon semantics rarely seen in the training data, such as "a cartoon picture depicting a person submerged in a mobile phone screen" and "a colored pencil drawing depicting a cute little bear", AnyStyle exhibits an inability to comprehend such intricate textual

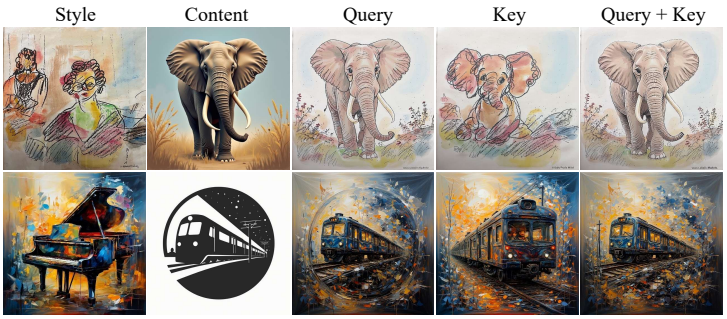

Fig. 10: Visualization of style transfer under different attention modulations. Key-only modulation yields strong stylistic consistency but weakens content preservation. Modulating both query and key improves content preservation over the key-only variant, yet reduces stylistic consistency. In contrast, query-only modulation better balances content and style, achieving the most effective overall style transfer performance.

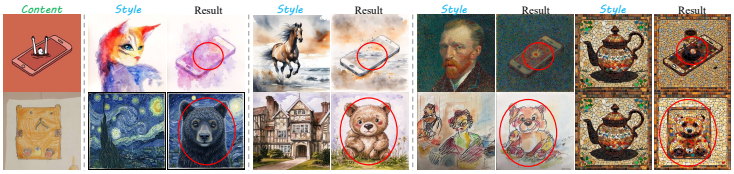

Fig. 11: Typical failure cases visualization of our proposed method.

descriptions. An illustrative example of this behavior is explicitly detailed in Fig. 11.

6 Conclusion

In this paper, we propose AnyStyle, a streamlined and effective framework for image-guided style transfer with improved controllability and stability. By employing a single-adapter strategy for consistent style generation, we avoid the complexities of multi-adapter fusion and the resulting entanglement between content and style. Furthermore, AnyStyle integrates attention modulation to provide precise structural guidance from the content reference image, ensuring faithful preservation of semantic layout and structure. Experiments demonstrate that our method achieves competitive performance compared to existing approaches, with enhanced coordination among prompt adherence, style consistency, and content fidelity in both quantitative metrics and perceptual evaluations.

7 Acknowledgments

This work was supported by Guangdong Basic and Applied Basic Research Foundation (Nos. 2024A1515140109 and 2023A1515110695).

References

1. Aravanis, T., Filntisis, P., Maragos, P., Retsinas, G.: Only-Style: Stylistic Consistency in Image Generation without Content Leakage. In: Greeks in AI Symposium 2025 (2025)
2. Chen, B., Zhao, B., Xie, H., Cai, Y., Li, Q., Mao, X.: ConsisLoRA: Enhancing content and style consistency for lora-based style transfer. arXiv preprint arXiv:2503.10614 (2025)
3. Cho, H., Lee, J., Chang, S., Jeong, Y.: One-Shot Structure-Aware Stylized Image Synthesis. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 8302–8311 (2024)
4. Choi, J., Shin, C., Oh, Y., Kim, H., Lee, J., Yoon, S.: Style-friendly snr sampler for style-driven generation. arXiv preprint arXiv:2411.14793 (2024)
5. Chong, M.J., Forsyth, D.: Jojogan: One shot face stylization. In: Proceedings of the European Conference on Computer Vision. pp. 128–152 (2022)
6. Chung, J., Hyun, S., Heo, J.P.: Style Injection in Diffusion: A Training-free Approach for Adapting Large-scale Diffusion Models for Style Transfer. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 8795–8805 (2024)
7. Cui, X., Li, Z., Li, P., Huang, H., Liu, X., He, Z.: InstaStyle: Inversion noise of a stylized image is secretly a style adviser. In: Proceedings of the European Conference on Computer Vision. pp. 455–472 (2024)
8. Efros, A.A., Freeman, W.T.: Image quilting for texture synthesis and transfer. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. p. 341–346 (2001)
9. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Proceedings of the International Conference on Machine Learning (2024)
10. Frenkel, Y., Vinker, Y., Shamir, A., Cohen-Or, D.: Implicit style-content separation using b-lora. In: Proceedings of the European Conference on Computer Vision. pp. 181–198 (2024)
11. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In: Advances in Neural Information Processing Systems. pp. 50742–50768 (2023)
12. Gal, R., Patashnik, O., Maron, H., Bermano, A.H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators. ACM Transactions on Graphics pp. 1–13 (2022)
13. Gatys, L.A., Ecker, A.S., Bethge, M.: Image Style Transfer Using Convolutional Neural Networks. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (2016)
14. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: Proceedings of the International Conference on Learning Representations (2023)
15. Hertz, A., Voynov, A., Fruchter, S., Cohen-Or, D.: Style aligned image generation via shared attention. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 4775–4785 (2024)
16. Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.H.: Image analogies. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. p. 327–340 (2001)

17. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Advances in Neural Information Processing Systems. pp. 6840–6851 (2020)
18. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: Proceedings of the International Conference on Learning Representations (2022)
19. Huang, X., Belongie, S.: Arbitrary Style Transfer in Real-Time With Adaptive Instance Normalization. In: Proceedings of the International Conference on Computer Vision (2017)
20. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Cao, L., Chen, S.: Diffusion Model-Based Image Editing: A Survey. Transactions on Pattern Analysis and Machine Intelligence pp. 4409–4437 (2025)
21. Jia, Y., Hoyer, L., Huang, S., Wang, T., Van Gool, L., Schindler, K., Obukhov, A.: Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In: Proceedings of the European Conference on Computer Vision. pp. 91–109 (2024)
22. Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., Song, M.: Neural Style Transfer: A Review. Transactions on Visualization and Computer Graphics pp. 3365–3385 (2020)
23. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European Conference on Computer Vision. pp. 694–711 (2016)
24. Labs, B.F.: FLUX. `https://github.com/black-forest-labs/flux` (2024)
25. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (2017)
26. Li, W., Fang, M., Zou, C., Gong, B., Zheng, R., Wang, M., Chen, J., Yang, M.: StyleTokenizer: Defining image style by a single instance for controlling diffusion models. In: Proceedings of the European Conference on Computer Vision. pp. 110–126 (2024)
27. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal Style Transfer via Feature Transforms. In: Advances in Neural Information Processing Systems (2017)
28. Li, Y., Liu, M.Y., Li, X., Yang, M.H., Kautz, J.: A Closed-form Solution to Photorealistic Image Stylization. In: Proceedings of the European Conference on Computer Vision (2018)
29. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)
30. Lin, J., Chen, X., Wu, S., Zhang, Z., Zhang, J., Wang, Y., Tang, Q., qian Wang, Yang, J., Yi, Z.: FreeControl: Efficient, Training-Free Structural Control via One-Step Attention Extraction. In: Advances in Neural Information Processing Systems (2025)
31. Lin, K.H., Mo, S., Klingher, B., Mu, F., Zhou, B.: Ctrl-X: Controlling Structure and Appearance for Text-To-Image Generation Without Guidance. In: Advances in Neural Information Processing Systems. pp. 128911–128939 (2024)
32. Liu, C., Shah, V., Cui, A., Lazebnik, S.: Unziplora: Separating content and style from a single image. In: Proceedings of the International Conference on Computer Vision. pp. 16776–16785 (2025)
33. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: Proceedings of the International Conference on Learning Representations. pp. 1–15 (2023)

34. Lu, M., Zhao, H., Yao, A., Chen, Y., Xu, F., Zhang, L.: A Closed-Form Solution to Universal Style Transfer. In: Proceedings of the International Conference on Computer Vision (October 2019)
35. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 4296–4304 (2024)
36. Nikolaidou, K., Retsinas, G., Sfikas, G., Liwicki, M.: DiffusionPen: towards controlling the style of handwritten text generation. In: Proceedings of the European Conference on Computer Vision. pp. 417–434 (2024)
37. Ojha, U., Li, Y., Lu, J., Efros, A.A., Lee, Y.J., Shechtman, E., Zhang, R.: Few-Shot Image Generation via Cross-Domain Correspondence. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 10743–10752 (2021)
38. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research Journal (2024)
39. Peebles, W., Xie, S.: Scalable Diffusion Models with Transformers. In: Proceedings of the International Conference on Computer Vision. pp. 4195–4205 (2023)
40. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: ICLR (2024)
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning. pp. 8748–8763 (2021)
42. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research pp. 1–67 (2020)
43. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis With Latent Diffusion Models. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
44. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241 (2015)
45. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
46. Shah, V., Ruiz, N., Cole, F., Lu, E., Lazebnik, S., Li, Y., Jampani, V.: Ziplora: Any subject in any style by effectively merging loras. In: Proceedings of the European Conference on Computer Vision. pp. 422–438 (2024)
47. Sohn, K., Jiang, L., Barber, J., Lee, K., Ruiz, N., Krishnan, D., Chang, H., Li, Y., Essa, I., Rubinstein, M., et al.: Styledrop: Text-to-image synthesis of any style. Advances in Neural Information Processing Systems pp. 66860–66889 (2023)
48. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

49. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023)
50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is All you Need. In: Advances in Neural Information Processing Systems (2017)
51. Wang, H., Spinelli, M., Wang, Q., Bai, X., Qin, Z., Chen, A.: Instantstyle: Free lunch towards style-preserving in text-to-image generation. arXiv preprint arXiv:2404.02733 (2024)
52. Wang, H., Xing, P., Huang, R., Ai, H., Wang, Q., Bai, X.: Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. arXiv preprint arXiv:2407.00788 (2024)
53. Wang, J., Pu, J., Qi, Z., Guo, J., Ma, Y., Huang, N., Chen, Y., Li, X., Shan, Y.: Taming rectified flow for inversion and editing. In: Proceedings of the International Conference on Machine Learning. pp. 1–11 (2025)
54. Wang, Z., Zhao, L., Xing, W.: StyleDiffusion: Controllable Disentangled Style Transfer via Diffusion Models. In: Proceedings of the International Conference on Computer Vision. pp. 7677–7689 (2023)
55. Xu, Y., Wang, Z., Xiao, J., Liu, W., Chen, L.: Freetuner: Any subject in any style with training-free diffusion. arXiv preprint arXiv:2405.14201 (2024)
56. Yang, Y., Wang, Y., Wang, C., Zhang, Y., Chen, Z., He, S.: SplitFlux: Learning to Decouple Content and Style from a Single Image. arXiv preprint arXiv:2511.15258 (2025)
57. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
58. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the International Conference on Computer Vision. pp. 3836–3847 (2023)
59. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the conference on computer vision and pattern recognition. pp. 586–595 (2018)
60. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-Based Style Transfer With Diffusion Models. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. pp. 10146–10156 (2023)
61. Zhang, Z., Zhang, Q., Xing, W., Li, G., Zhao, L., Sun, J., Lan, Z., Luan, J., Huang, Y., Lin, H.: Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7396–7404 (2024)