

# Optimizing Digital Art Style Transfer Using a Diffusion Model

Da Liu<sup>1</sup>, Li-Chao Si<sup>2</sup>

<sup>1</sup>Fine Arts and Film College, Taiyuan Normal University, Taiyuan 030619, Shanxi, China

<sup>2</sup>Arts and Design College, Taiyuan University, Taiyuan 030032, Shanxi, China

Corresponding author: Da Liu (e-mail: ld254572315@126.com)

**ABSTRACT** Current generative models are prone to content distortion when guided by strong digital art styles, limiting the preservation of structural and semantic information during style transfer. Such controllable visual representation and semantic consistency are also important for intelligent imaging, computational perception, and electromagnetic information processing systems that require reliable feature preservation across heterogeneous visual domains. To address this issue, this study proposes a latent diffusion model (LDM)-based framework for optimized digital art style transfer. Strong style is quantified by the normalized norm of the CLIP-encoded style embedding in the latent space, and its distribution is employed as a continuous indicator to regulate style injection during diffusion. A Cross-Attention mechanism is introduced into the inverse denoising process to dynamically fuse content and style features while preventing structural distortion caused by excessive style guidance. In addition, a linear strength factor enables continuous adjustment of style intensity, and a pre-trained ControlNet generates regional guidance masks to spatially constrain style propagation in the latent space. Guided by style encoding, the predicted noise is iteratively refined under a fixed diffusion schedule to recover high-quality latent representations while preserving semantic consistency. Experiments on the WikiArt and COCO datasets demonstrate that the proposed method effectively alleviates content distortion and achieves a PSNR of  $27.6 \pm 1.2$  dB and an SSIM of  $0.85 \pm 0.02$ . The proposed framework provides an effective solution for controllable image generation and offers potential value for computational imaging, intelligent visual sensing, and electromagnetic information processing applications requiring high-fidelity semantic representation.

**INDEX TERMS** latent diffusion model, digital art style transfer, cross-attention, computational imaging, electromagnetic information processing

## I. INTRODUCTION

WITH the rapid development of artificial intelligence in art creation, digital art style transfer technology gradually becomes a hot research topic in image generation and visual content editing, a trend that holds profound implications for the industry by offering a powerful tool to rapidly prototype, visualize, and apply complex artistic styles to new fabric designs and garment patterns [1,2]. Most of the early methods complete the image content and style feature extraction through convolutional neural networks (CNNs), and then optimize the target image to keep the balance between content map and style map [3]. However, these methods often have high computation overhead and poor generalization ability, which makes it hard to apply them in real-time requirements or complex generation scenarios. In pattern transfer, the curved garment layout introduces local distortions during style injection, making controlled style modulation and region constraints essential for maintaining pattern direction. Subsequently, Generative Adversarial Networks (GAN) gradually became

mainstream. Fast Style Transfer [4], StyleGAN [5] and AdaIN (Adaptive Instance Normalization) [6] and other methods gradually developed based on GAN, which made some breakthroughs in speed and style. However, compared with traditional methods, they still generally have problems such as poor style expression ability, low fidelity of image details, and poor style control ability, making it hard to meet the constantly increasing requirements of artistic style transfer images.

This paper proposes an optimization scheme of digital artistic style transfer based on latent diffusion model, and constructs a multimodal input and style control generation framework, which enables more refined expression of art style and more flexible transfer control. In the overall design, it takes pre-trained model CLIP (Contrastive Language-Image Pre-Training) as style encoder, and uses cross-attention mechanism to dynamically fuse content graph and style graph. In the training process, a multi-objective optimization system is constructed by diffusion loss, style consistency loss and content preservation loss. In the generation

process, conditional control can guide the gradual denoising process in the latent space to finish the modeling of the whole process from style perception to image generation.

## II. RELATED WORKS

In recent years, with the continuous evolution of deep learning technology, image style transfer methods have continued to develop, gradually shifting from the early optimization-driven paradigm to a technical path dominated by generative models [7,8]. Zhou Y et al. first proposed an OCT (Optical Coherence Tomography) image speckle noise suppression method based on image style transfer and conditional generative adversarial networks, establishing a new framework that combines cross-domain image style matching with speckle noise removal [9]. Subsequently, Fernandez-Fernandez R proposed the NPST (Neural Policy Style Transfer) method, which achieved style transfer of control policies through deep reinforcement learning [10]. Zhu Q proposed the LPAdaIN (Light Progressive Attention Adaptive Instance Normalization) model, which further introduced the Convolutional Block Attention Module (CBAM) and reconstruction loss, improving the quality of style transfer while enhancing the ability to preserve image structure [11].

Based on this, the application of diffusion models in artistic style transfer tasks has gradually attracted the attention of researchers, and a series of innovative methods have emerged. He H proposed the Diff-Font framework, which combines content embedding, style encoding and stroke/component features to guide the diffusion process to achieve high-fidelity and structurally complete one-time font generation [12]. Yang L introduced a context prediction mechanism in CONPREDIFF, and significantly improved the semantic consistency and generation quality of the diffusion model in image generation through multi-step neighborhood distribution modeling and optimal transmission constraints [13]. Zhao S designed a dual adapter structure of local and global control in Uni-ControlNet to achieve unified control of multiple conditions and flexible combination of style transfer, and improved the fidelity and controllability of generated images through multi-scale condition injection and extended hint mechanism [14]. Mukherjee D proposed the AGGrGAN framework, which synthesizes high-quality brain tumor MRI images by fusing multiple GAN generation results and combining them with style transfer, effectively alleviating the problem of scarce medical image data [15]. The above research shows that the diffusion model has high expressiveness and controllability in style transfer tasks, especially showing unique advantages in cross-modal guidance, multi-style fusion, and fine-grained adjustment.

## III. IMPLEMENTATION OF DIGITAL ART STYLE TRANSFER METHODS BASED ON LDM

### OVERALL SYSTEM ARCHITECTURE

The digital art style transfer system constructed in this study is centered around LDM, and the overall process follows

an encoding-guidance-diffusion-decoding architecture. By mapping the image generation task into a latent space, the system significantly reduces computational overhead in the high-dimensional image domain and improves the stability and controllability of the generation process. The architecture of this system is shown in Figure 1:

The system is composed of three modules: First, the style guidance module encodes style features from images or texts and feeds them into the latent space. Specifically, the style feature is represented as a conditional vector in the latent space by either CLIP model or a built style encoder according to the style pattern or description of language. Second, the conditional control module guides and controls style in a local region with structured information, which includes edge maps, semantic maps and depth maps. Besides, the lightweight architecture is introduced to enhance the control accuracy of style, that is, LoRa and ControlNet. Finally, the diffusion generation module reverses the denoising process based on U-Net backbone network. It adopts cross-attention to combine the latent representation of content and style to assist dynamic style transfer in multi-steps generation. Finally, VAE (Variational Auto Encoder) decoder recovers the image and outputs a high-quality image.

### DATA PREPARATION AND PREPROCESSING

To train a latent diffusion model with style transfer capabilities, this paper constructs paired training samples between content and style maps during the data preparation phase. Furthermore, the raw data undergoes uniform augmentation and preprocessing to improve the model's generalization and stability across diverse styles and structures.

#### Construction of Content Graph and Style Graph Pairs

This study collects raw image data from two dimensions, general images and artworks, to construct training samples for the digital art style transfer task. The content images are primarily sourced from the COCO 2017 (Common Objects in Context 2017) and Flickr30k [16] datasets, covering a variety of semantic categories such as natural scenery, portraits, urban architecture, and everyday objects. After preliminary screening and deduplication, a total of 12,000 high-quality content images were selected to ensure broad representation in terms of texture complexity and compositional structure.

### DATA AUGMENTATION AND FORMAT CONVERSION

During image data preprocessing, this paper first performs a uniform size normalization operation on all collected image samples. All content maps and style maps are scaled to a fixed resolution of  $512 \times 512$  pixels, and pixel values are normalized per channel to the interval  $[-1, 1]$  to meet the input image format requirements of the latent diffusion model during training. After normalization, to enhance the model's robustness to diverse scenes and style variations, this paper further introduces various online data augmentation strategies. In each round of training, image samples

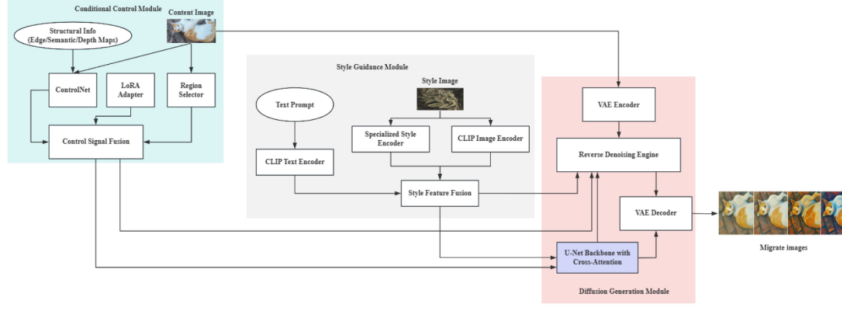


FIGURE 1. System Architecture

are randomly subjected to operations including horizontal flipping, brightness and contrast perturbations (the perturbation range is set to  $\pm 10$ ), random cropping (the cropping ratio is set within the interval  $[0.9, 1.0]$ ), and a mild Gaussian blur. These augmentations are implemented jointly by the Albumentations and TorchVision data augmentation libraries and embedded into the training data stream of loaded images, ensuring that each input sample seen by the model exhibits a certain degree of diversity and perturbation. The final processed data is split into a training set and a validation set in an 8:2 ratio.

In terms of conditional encoding, to meet the requirements of the style guidance mechanism for modeling external style features, this paper uniformly uses the pre-trained CLIP-ViT/32 model to encode the style map. Assuming the input style map is  $I_s$ , its corresponding style encoding vector  $c_s$  is obtained by the following formula:

$$c_s = \text{CLIP}_{\text{image}}(I_s) \quad (1)$$

Here,  $\text{CLIP}_{\text{image}}(\cdot)$  represents the image branch encoder. The resulting  $c_s \in \mathbb{R}^{512}$  vector is used as the conditional input in the Cross-Attention module for style control during the diffusion process. Similarly, when using text descriptions as style cues, the input text  $T_s$  is converted into a semantic vector  $c_t$  through the CLIP text encoder, represented as follows:

$$c_t = \text{CLIP}_{\text{text}}(T_s) \quad (2)$$

To achieve consistent style control across images and text, all conditional codes are mapped to a unified latent space dimension and injected into each step of the diffusion model's reverse denoising process during the generation phase.

### LATENT DIFFUSION MODEL (LDM) ARCHITECTURE

#### VAE Encoder/Decoder

In this study, the VAE encoder consists of a multi-layer CNN, specifically four convolutional blocks. Each convolutional block consists of two consecutive  $3 \times 3$  convolutional layers with a stride of 2. Combined with batch normalization and ReLU (Rectified Linear Unit) activation functions, this effectively achieves gradual spatial compression and feature

extraction. Through layer-by-layer downsampling, the input image with a resolution of  $512 \times 512$  is ultimately mapped to a latent space tensor  $z \in \mathbb{R}^{64 \times 64 \times 4}$ , significantly reducing the data dimension while preserving semantic information. A linear layer is connected to the end of the encoder to generate the mean  $\mu$  and logarithmic variance  $\log \sigma^2$  of the latent space. The latent variables are sampled using the reparameterization technique, as shown in the following formula:

$$z = \mu + \sigma \odot \epsilon, \quad \epsilon \sim N(0, I) \quad (3)$$

This sampling method ensures the differentiability of the latent space representation, enabling joint end-to-end training of the encoder and decoder. The resulting latent variable  $z$  serves as the initial input to the diffusion process.

### U-NET INVERSE DIFFUSION NETWORK

In this study, the inverse diffusion network employs a multi-scale U-Net architecture, specifically designed for noise removal in the latent space. The network takes as input the noisy latent variable  $z_t$ , the number of diffusion steps  $t$ , and conditional information  $c$ . The model predicts the noise residual at the current step, as follows:

$$\hat{\epsilon}_\theta = \text{U-Net}(z_t, t, c) \quad (4)$$

The network consists of a symmetrical encoder and decoder. The encoder extracts multi-scale features through layer-by-layer convolution and downsampling, while the decoder gradually restores spatial resolution

through deconvolution and upsampling. Skip connections are used between the encoder and decoder to transmit detailed information, ensuring that features at different levels are fully utilized.

### CONDITIONAL GUIDANCE MODULE

To achieve flexible control over the digital art style transfer process, the latent diffusion model designed in this paper supports multiple types of conditional inputs. First, the model receives the style map encoding extracted in the previous preprocessing stage as the primary style guidance vector, which is used to drive style injection during the generation process. Secondly, textual description information also serves as an important conditioning channel. Style

instructions such as “oil painting style” and “cyberpunk style” are converted into semantic embedding vectors via the CLIP text encoder, enabling natural language-driven image style control. All conditioning information is injected into the latent space features via the Cross-Attention module in the diffusion model to support the fusion of style and structural information. Furthermore, to improve parameter efficiency and adjustability, the model introduces a lightweight modulation mechanism, LoRA, within the multi-layer network structure. This further refines the path through which conditioning controls affect the generated image, thereby enhancing the flexibility and consistency of style expression.

### CROSS-ATTENTION GUIDANCE MECHANISM

#### Fusion Principle and Structural Design

In the latent diffusion model constructed in this study, the Cross-Attention module serves as the core conditional fusion mechanism, embedded within the intermediate layers and skip connections of the U-Net network. This module is responsible for enabling efficient interaction between latent space features and style conditions. In each diffusion step  $t$ , the model takes the current latent space representation as the query vector  $Q$  and receives the style vector encoded by the style map or text instructions as a key-value pair  $(K, V)$ . The model then computes the fusion result using the scaled dot-product attention mechanism, as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (5)$$

where  $d_k$  is the dimension of the key vector. To ensure feature alignment and comparability, all input vectors undergo linear mapping before entering the attention module, projecting them to the same embedding

dimension. The method introduces a temporally dependent weighting function during inverse denoising, where the noise level at each timestep is mapped to a continuous coefficient applied to the style key-value attention scores, producing a progressive modulation that maintains low style influence at early noisy stages and increases it as the latent becomes cleaner.

This mechanism supports multiple stylistic conditional injection methods. These guide vectors are assigned to different scale levels of the U-Net, enabling the model to achieve hierarchical injection of style features across multi-resolution feature spaces, thereby improving the semantic consistency and stylistic expressiveness of the transferred results. To enhance the controllability of the style fusion process, the model further introduces a style intensity factor ( $\alpha$ ) to linearly scale the Cross-Attention output to adjust the influence of style guidance on the generated image. The overall computational form of this mechanism can be expressed as:

$$z_{\text{out}} = z + \alpha \cdot \text{Attention}(Q, K, V) \quad (6)$$

where  $z$  represents the feature representation of the current latent space, and  $\alpha \in [0, 1]$  is the weight coefficient for style fusion. By adjusting the  $\alpha$  value, the model can effectively control the strength of style injection while maintaining the stability of the image structure, achieving fine-tuning and expanding the diversity of the generated results. In the style control path, the intensity factor  $\alpha$  participates in the scaling of the cross-attention output as a continuous real number, allowing style features to diffuse along a stable, gradual trajectory in the reverse step. Attention responses at different  $\alpha$  values produce coherent style evolution in the feature maps of each layer, gradually transitioning from weak texture shifts at low intensity to significant style manifestations at high intensity, maintaining the coherence of style direction and the consistency of structural semantics. All diffusion timesteps follow the original schedule, and the style encoding modulates only the predicted noise  $\epsilon$  within the cross-attention path. In this weighting path, the style and content coefficients are not static; they are jointly determined by the noise estimator and the normalized style-embedding norm. The norm contributes to the modulation term, creating a dual-input mapping whose output induces a nonlinear response in the attention matrix as the style intensity varies, resulting in differentiated fusion strength across feature levels.

#### Style Control and Region Guidance Mechanism

In the style transfer system constructed in this study, to achieve fine-grained control of image style and preserve the structure of key regions, the Cross-Attention module incorporates a region guidance mechanism

based on spatial structure when incorporating style guidance information. When the model input contains structured information such as edge maps, semantic segmentation maps, or depth maps, the system utilizes a pre-trained ControlNet network to encode these local structural conditions and convert them into a region guidance mask  $M$  aligned with the latent space representation. The value range is  $[0, 1]$ , which indicates the spatial strength of the style injection. The selection of the ControlNet architecture in this module is driven by its latent feature alignment with the diffusion backbone, yielding a region representation whose spatial

resolution, feature density and latent-domain consistency remain stable across diffusion steps, and this stability allows the regional mask to match the multi-scale latent features generated during inverse denoising without additional transformation layers.

In the Cross-Attention calculation, the regional guidance mask  $M$  acts on the attention weight matrix  $A = \text{softmax}(QK^\top / \sqrt{d_k})$  to achieve weighted adjustment:

$$\tilde{A}_{ij} = M_{ij} \cdot A_{ij} \quad (7)$$

After normalization, the adjusted attention weights ( $\tilde{A}$ ) are used for the final weighted style feature injection. This design ensures that style features are injected with high

intensity only within the structure-guided regions, while significantly suppressing style interference in unlabeled regions. The regional mask produced by the ControlNet encoder maintains coherence with the hierarchical latent structure within the U-Net path, avoiding discrepancies between semantic grouping and latent-space positioning, and this property reduces the need for an auxiliary alignment network and preserves the decoder's spatial mapping fidelity during reconstruction. The region-guided mechanism and the style strength factor ( $\alpha$ ) work together to dynamically apply style strength to different image regions, ensuring structural fidelity in detail areas and flexible expression of overall style. The dynamic fusion further adjusts attention distributions independently across multi-scale feature blocks. At high-resolution layers, the style weights remain stable under region masking, while at low-resolution layers they exhibit increased concentration driven by the style-embedding norm, aligning local textures, global structures, and stylistic elements throughout the fusion process. The region mask is generated in the same latent dimensional configuration as the style and content encodings, ensuring that the spatial modulation signal remains numerically aligned with the cross-attention fusion path, which prevents errors introduced by projecting externally segmented masks into the latent domain and maintains consistent spatial semantics throughout all reverse timesteps.

## MODEL TRAINING PHASE

### TRAINING OBJECTIVE AND LOSS FUNCTION DESIGN

The training objective of this model is composed of multiple loss functions to ensure a balance between style transfer effectiveness and content preservation. The core losses include diffusion loss, style consistency loss, content preservation loss, perceptual loss, and CLIP semantic preservation loss. The overall loss function is defined as:

$$L_{\text{total}} = \lambda_{\text{diff}} L_{\text{diff}} + \lambda_{\text{style}} L_{\text{style}} + \lambda_{\text{content}} L_{\text{content}} + \lambda_{\text{perc}} L_{\text{perc}} + \lambda_{\text{clip}} L_{\text{CLIP}} \quad (8)$$

The diffusion loss  $L_{\text{diff}}$  is the core goal of training the latent diffusion model, which aims to guide the U-Net network to accurately predict the noise component in the forward Gaussian perturbation process. Its loss form is:

$$L_{\text{diff}} = \mathbb{E}_{x, \epsilon, t} \left[ \|\epsilon - \epsilon_{\theta}(z_t, t, c)\|^2 \right] \quad (9)$$

Here,  $z_t$  represents the noisy latent space variable,  $c$  represents the conditional input (content map encoding and style encoding),  $\epsilon$  represents the actual noise, and  $\epsilon_{\theta}$  represents the model's predicted noise. To ensure that the transferred result retains the visual characteristics of the target style, the system introduces a style consistency loss ( $L_{\text{style}}$ ). In this implementation, a pre-trained VGG19 (Visual Geometry Group 19) network is used to extract feature maps from multiple convolutional layers of the generated image and the target style map. The Gram matrix is calculated as the style

statistic, and the difference between the two is measured using the mean squared error (MSE):

$$L_{\text{style}} = \sum_l \|G(F_l(x_{\text{gen}})) - G(F_l(x_{\text{style}}))\|^2 \quad (10)$$

Here,  $G(F_l)$  represents the Gram matrix of the feature map at layer  $l$ , capturing the texture and color distribution patterns in the style. Meanwhile, the content-preserving loss  $L_{\text{content}}$  focuses on the semantic structure and edge information of the content map. It uses the same VGG network intermediate layers to extract feature representations of the content map and the generated map, constraining the fidelity of the structural information during style transfer:

$$L_{\text{content}} = \|F_l(x_{\text{gen}}) - F_l(x_{\text{content}})\|^2 \quad (11)$$

In addition, to further improve the overall perceptual quality of the generated images, the system introduces perceptual loss  $L_{\text{perc}}$ . This loss uses a multi-level feature representation to optimize the perceptual similarity of images at multiple scales by accumulating errors from multiple perceptual layers:

$$L_{\text{perc}} = \sum_{l \in L} \|F_l(x_{\text{gen}}) - F_l(x_{\text{content}})\|^2 \quad (12)$$

To achieve semantic alignment of text-conditioned style descriptions, the system introduces the CLIP multimodal loss  $L_{\text{CLIP}}$ . This loss, based on a pre-trained CLIP model, feeds the generated image into an image encoder  $\phi_{\text{img}}(\cdot)$  and the style text into a text encoder  $\phi_{\text{text}}(\cdot)$ . By maximizing the cosine similarity between their embeddings, it achieves coherence between style semantics and visual results:

$$L_{\text{CLIP}} = 1 - \cos(\phi_{\text{img}}(x_{\text{gen}}), \phi_{\text{text}}(t_{\text{style}})) \quad (13)$$

During the training phase, all loss functions are weighted and summed using fixed weights to form the final joint loss function  $L_{\text{total}}$ , which guides the LDM in achieving accurate, semantically clear, and structurally faithful image generation under complex multimodal conditions.

### TRAINING PROCESS

The latent diffusion model constructed in this study uses a variational diffusion modeling framework based on a latent space. During training, the model learns the high-order distribution laws of image style transfer by simulating forward perturbations and inverse restoration processes. The training process consists of two stages: forward diffusion construction and inverse denoising prediction.

In the forward propagation phase, the system maps the content image to a latent space representation  $z_0 = E(x)$  through a pre-trained VAE encoder. Subsequently, Gaussian noise is gradually added to this latent variable to generate

a noisy representation  $z_t$  that satisfies the following perturbation formula:

$$z_t = \sqrt{\bar{\alpha}_t} \cdot z_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon, \quad \epsilon \sim N(0, I) \quad (14)$$

$\bar{\alpha}_t$  represents the cumulative noise retention coefficient at the  $t$ -th time step, and  $\epsilon$  represents standard Gaussian noise. During each training round, time step  $t$  is randomly sampled from the interval  $[1, T]$  to simulate the perturbation of the latent variable under any diffusion phase.

In the reverse phase, the model receives the noisy latent variable  $z_t$ , the corresponding temporal encoding  $t$ , and the conditional vector  $c$  through a conditionally guided U-Net network, and outputs a prediction for the noise term. The reverse process keeps the predefined  $\bar{\alpha}_t$  sequence unchanged, and the style encoding affects only the predicted noise  $\epsilon$ . This prediction is compared with the true noise term, and the deviation is calculated using a loss function for backward updates.

The training process uses a fixed VAE encoder and decoder architecture, whose parameters are frozen during training to ensure latent space stability and maintain consistent image decoding quality. All trainable parameters in the U-Net network and conditional injection module were iteratively updated using the AdamW optimizer, with an initial learning rate of  $1 \times 10^{-5}$  and a weight decay coefficient of 0.001. During the optimization process, a gradient clipping mechanism was incorporated to mitigate the risk of gradient explosion and ensure the numerical stability of the deep network during training. Regarding the model's hyperparameter settings, this study conducted multiple rounds of experiments and determined the following optimal hyperparameter configuration: The learning rate is set to  $1 \times 10^{-4}$ , the weight decay coefficient is 0.01, the batch size is 4, the diffusion steps are  $T = 1000$ , the style fusion factor is  $\alpha = 0.6$ , the style consistency loss weight is  $\lambda_{\text{style}} = 50$ , the content preservation loss weight is  $\lambda_{\text{content}} = 30$ , the perceptual loss weight is  $\lambda_{\text{percep}} = 2$ , the CLIP multimodal semantic loss weight is  $\lambda_{\text{clip}} = 5$ , and the default coefficients of the remaining loss terms are 1. Sinusoidal positional embedding is used for all time steps, and the conditional vectors are uniformly projected into a 512-dimensional latent space.

The model training cycle lasted for 400 epochs. During training, the training loss and SSIM (Structural Similarity) on the validation set were used as evaluation metrics to measure the model's artistic style transfer performance. This was used to monitor the model's learning and style transfer performance in real time. To improve training efficiency and model generalization, the system introduced an Early Stopping strategy. If the validation metrics showed no significant improvement after 20 consecutive epochs, training was terminated early and the currently optimized model parameters were retained. Furthermore, the trends of

key performance indicators were recorded during each training round. Under the optimal hyperparameter configu-

ration, the total loss and SSIM curves during model training are shown in Figure 2:

Figure 2 shows the total loss and SSIM curves during model training. The horizontal axis represents the number of training epochs, and the vertical axis corresponds to the total loss and SSIM values, respectively. As training progresses, the total loss shows a clear downward trend, reaching a minimum around the first 250 epochs and then stabilizing, indicating that the model has converged to a good state during this period. Meanwhile, the SSIM value shows an upward trend, reaching a maximum around the 250th epoch and then stabilizing, indicating that the structural similarity between the generated images and the target images has been significantly improved and maintained a high degree of stability during subsequent training.

## IV. EXPERIMENTS AND PERFORMANCE EVALUATION

### A. EXPERIMENTAL SETUP

All experiments were conducted in a hardware environment equipped with high-performance image generation computing capabilities. The specific configuration is as follows: Ubuntu 20.04 LTS operating system, dual NVIDIA A100 (80GB) GPUs, a 64-core Intel Xeon Gold 6326 CPU, and 512GB of RAM, capable of supporting parallel training and inference of large-scale deep models. In terms of the software environment, the paper used PyTorch 2.1.0 as the deep learning framework, with CUDA version 11.8. The paper also integrated libraries such as Transformers (v4.36), Diffusers (v0.18), OpenCV, and scikit-image to assist with preprocessing and image evaluation.

The experimental data was collected from COCO 2017, Flickr30k, WikiArt, and Behance Artistic Media. The content images covered a variety of semantic categories, including natural landscapes, portraits, urban architecture, and everyday objects. The style images encompassed over ten genres, including Impressionism, Cubism, Expressionism, Futurism, modern illustration, and digital painting.

### QUANTITATIVE EVALUATION RESULTS AND ANALYSIS

#### STYLE REPRESENTATION AND SEMANTIC PRESERVATION EVALUATION

To comprehensively evaluate the model's ability to accurately represent the target style and its ability to preserve the semantic information of the content in the digital art style transfer task, this study designed two independent experiments: style similarity analysis and style classification testing.

For style similarity analysis, a mixed test set was constructed based on the data collected previously. This set contained 500 content images, which were then combined with 10 representative artistic styles to generate stylized images, totaling 5,000 test samples. Using the pre-trained CLIP-ViT/B32 model, each generated image was encoded with its corresponding target style reference image. The image embedding vectors were extracted, and the cosine similarity between the two was calculated to measure the

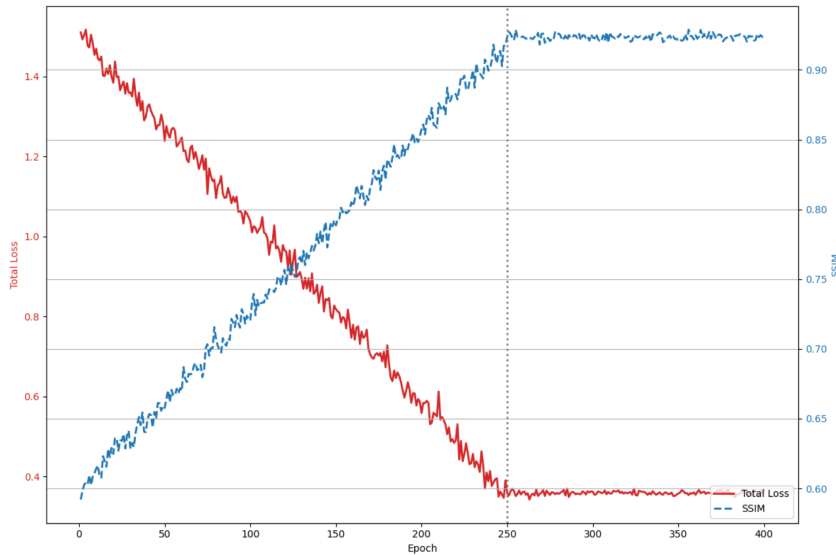


FIGURE 2. Model Training Loss and SSIM Curves

style similarity between the generated image and the target style in a high-level semantic space. The results are shown in Figure 3:

Figure 3 shows a comparison of cosine similarity between different methods, which measures the style similarity between the generated image and the target style in a high-level semantic space. Four groups are included: Control 1, Control 2, Control 3, and the Experimental group. As can be seen from the boxplot, the experimental group's style similarity is significantly higher than that of the other control groups, indicating that the proposed model performs best in the style transfer task and is able to more accurately capture and express the target style. Control group 2 performed second best, with similarity distribution concentrated between 0.75 and 0.85. Control groups 1 and 3 had relatively low similarities, particularly Control group 1, whose similarity was primarily concentrated between 0.55 and 0.7, indicating poor style consistency. In the style classification test, this study trained a ResNet-50 style classifier based on the WikiArt dataset, covering 10 major artistic styles. The stylized images generated in the style similarity analysis experiment were fed into the classifier, and the predicted labels were obtained. The top-1 and top-3 accuracy scores were recorded to evaluate the model's performance on the style recognition task. All experiments were run under identical conditions, with each test set repeated three times to reduce the impact of random error. The final results are summarized in Table 1:

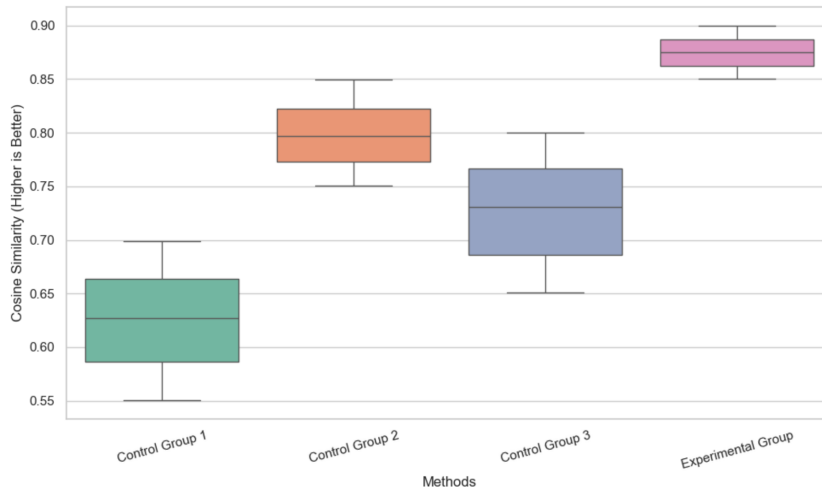
TABLE 1. Comparison of Accuracy of Different Methods on the Style Classification Task (Top-1/Top-3)

| Style                  | Control Group 1<br>(Top-1/Top-3) | Control Group 2<br>(Top-1/Top-3) | Control Group 3<br>(Top-1/Top-3) | Experimental Group<br>(Top-1/Top-3) |
|------------------------|----------------------------------|----------------------------------|----------------------------------|-------------------------------------|
| Impressionism(%)       | 62.4 / 78.1                      | 70.8 / 85.6                      | 67.3 / 82.5                      | 76.9 / 89.4                         |
| Cubism(%)              | 59.8 / 76.5                      | 69.4 / 84.2                      | 65.2 / 81.0                      | 75.3 / 88.7                         |
| Expressionism(%)       | 60.5 / 77.3                      | 71.2 / 86.3                      | 66.8 / 83.2                      | 77.1 / 89.2                         |
| Futurism(%)            | 61.2 / 78.9                      | 72.5 / 87.1                      | 68.1 / 84.5                      | 78.5 / 90.1                         |
| Modern illustration(%) | 63.1 / 79.4                      | 73.8 / 88.0                      | 69.5 / 85.7                      | 79.8 / 90.8                         |
| Digital Painting(%)    | 64.5 / 80.2                      | 74.6 / 88.5                      | 70.3 / 86.1                      | 80.6 / 91.3                         |
| Post-Impressionism(%)  | 61.8 / 78.3                      | 71.5 / 85.9                      | 67.9 / 82.8                      | 79.2 / 90.5                         |
| Abstract Art(%)        | 60.2 / 77.6                      | 70.3 / 84.8                      | 66.5 / 81.7                      | 78.7 / 89.9                         |
| Surrealism(%)          | 62.5 / 79.1                      | 72.8 / 86.7                      | 68.4 / 84.2                      | 80.1 / 91.0                         |
| Baroque(%)             | 63.7 / 79.8                      | 73.4 / 87.2                      | 69.2 / 85.5                      | 81.3 / 91.5                         |

Table 1 shows the style classification accuracy (Top-1 / Top-3) of different methods across ten artistic styles. As can be seen, the proposed method significantly outperforms the three control groups across all style categories, achieving an average Top-1 accuracy of 78.8% and a Top-3 accuracy of nearly 90.2%, demonstrating its clear advantages in both consistency and diversity of style expression. Control group 1, a traditional optimization method, exhibits the weakest overall performance, likely due to its reliance on direct optimization in the pixel space, which makes it difficult to effectively model high-level semantic information, resulting in content distortion or insufficient style expression during style transfer. Control Group 2, based on the StyleGAN-NADA framework, demonstrated strong capabilities in language-driven style transfer, outperforming both Control Groups 1 and 3.

CONTENT STRUCTURE FIDELITY ANALYSIS

To evaluate the model's ability to preserve the structural information of the original image in the digital art style transfer task, this study used three mainstream metrics: SSIM, PSNR (Peak Signal-to-Noise Ratio), and LPIPS (Learned Perceptual Image Patch Similarity) for quantitative evaluation, focusing on structural similarity and perceptual consistency.



**FIGURE 3.** Comparison of Style Similarity Distributions among Different Methods

During the experiment, a hybrid test set constructed from the WikiArt and COCO datasets was used. 500 images not used in training were selected as baseline input. These images were then combined with 10 typical artistic styles to generate stylized images, totaling 5,000 test samples. Each set of generated images was paired pixel-by-pixel with its corresponding content image. SSIM, PSNR, and LPIPS values were calculated at the global scale and at multiple local scales ( $256 \times 256$ ,  $128 \times 128$ , and  $64 \times 64$ ). The average and standard deviation of all

samples across three independent rounds of experiments were used as performance metrics. Table 2 shows the overall performance of proposed method and three control methods on these three metrics:

**TABLE 2.** Overall Performance of Each Method on SSIM, PSNR, and LPIPS

| Method             | SSIM $\uparrow$ | PSNR $\uparrow$ (dB) | LPIPS $\downarrow$ |
|--------------------|-----------------|----------------------|--------------------|
| Control Group 1    | $0.76 \pm 0.05$ | $24.1 \pm 1.8$       | $0.28 \pm 0.04$    |
| Control Group 2    | $0.81 \pm 0.03$ | $26.0 \pm 1.5$       | $0.22 \pm 0.03$    |
| Control Group 3    | $0.79 \pm 0.04$ | $25.3 \pm 1.6$       | $0.25 \pm 0.03$    |
| Experimental Group | $0.85 \pm 0.02$ | $27.6 \pm 1.2$       | $0.18 \pm 0.02$    |

Table 2 shows the comprehensive performance of proposed method and three control methods in terms of SSIM, PSNR, and LPIPS. As can be seen, proposed method significantly outperforms the three control methods in all metrics, demonstrating its clear advantages in image structural fidelity, perceptual consistency, and style control. In terms of SSIM, proposed method achieves  $0.85 \pm 0.02$ , significantly higher than  $0.76 \pm 0.05$  (Control 1),  $0.81 \pm 0.03$  (Control 2), and  $0.79 \pm 0.04$  (Control 3), demonstrating that proposed method excels in preserving the overall image structure. In the PSNR metric, proposed method achieved a score of  $27.6 \pm 1.2$  dB, also outperforming the other three methods, further demonstrating its superiority in detail restoration and noise suppression. In the LPIPS metric, proposed method achieved a value of  $0.18 \pm 0.02$ , significantly lower than

the  $0.28 \pm 0.04$  (Control 1),  $0.22 \pm 0.03$  (Control 2), and  $0.25 \pm 0.03$  (Control 3).

In addition, to further analyze the differential distribution of generated images at different perceptual levels, this study statistically analyzed the average variation trend of LPIPS in the feature space of layers 1 to 5 of the VGG network. The results are shown in Figure 4:

Figure 4 shows the distribution trend of LPIPS for content and generated images in the feature space of layers 1 to 5 of the VGG network. As the number of VGG layers increases, the LPIPS value shows a clear downward trend, indicating that the perceptual difference between the generated images and the content images gradually decreases.

### CONDITIONAL CONTROL ANALYSIS

To verify the controllability of this model over style intensity and impact area during style transfer, two types of experiments were designed: style intensity adjustment and area control.

In the style intensity adjustment experiment, the style fusion factor  $\alpha \in \{0.2, 0.4, 0.6, 0.8\}$  was set, and the style injection intensity was adjusted by linearly scaling the cross-attention output. Based on a hybrid test set constructed from WikiArt and COCO, the paper selected 200 content images and generated 1,000 sets of samples from five artistic styles: Impressionism, Cubism, Expressionism, Futurism, and Digital Painting. Each set of samples was generated using the same input, varying only the  $\alpha$  value, and the variation in style expression was recorded. The paper used CLIP similarity to measure style consistency and calculated the deviation between the actual variation rate and the theoretical expectation. The results are shown in Table 3:

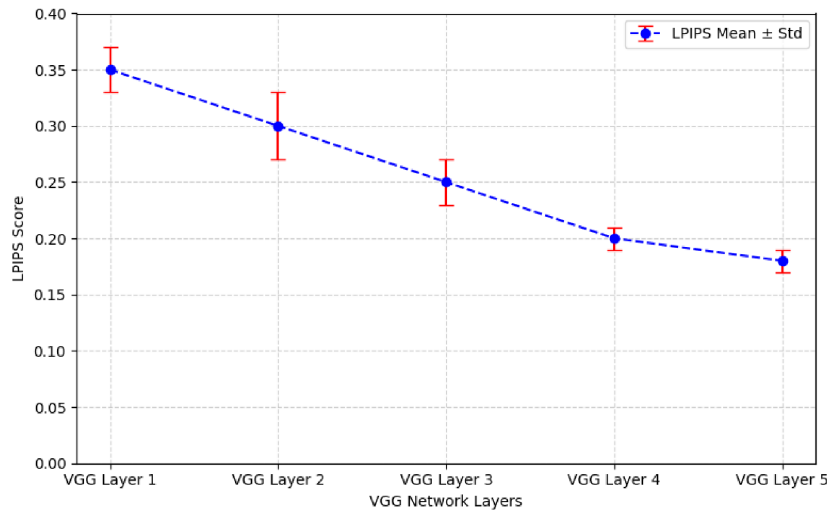


FIGURE 4. Distribution Trend of LPIPS for Content and Generated Images in Multiple Layers of Perceptual Space

TABLE 3. CLIP Similarity Deviation for Each Style Under Different Style Fusion Factors

| Style Type       | $\alpha = 0.2$ | $\alpha = 0.4$ | $\alpha = 0.6$ | $\alpha = 0.8$ |
|------------------|----------------|----------------|----------------|----------------|
| Impressionism    | 2.1%           | 1.8%           | 1.5%           | 1.3%           |
| Cubism           | 2.3%           | 2%             | 1.7%           | 1.4%           |
| Expressionism    | 2.5%           | 2.1%           | 1.9%           | 1.6%           |
| Futurism         | 2.2%           | 1.9%           | 1.6%           | 1.4%           |
| Digital Painting | 2%             | 1.7%           | 1.4%           | 1.2%           |

Table 3 shows the deviation between the CLIP similarity of generated images and theoretical expectations for different style fusion factors. The data shows that as the value of  $\alpha$  increases, the deviation of each artistic style decreases, demonstrating that the model has good linear response characteristics in controlling the intensity of style injection. At  $\alpha = 0.2$ , the deviation of Expressionism reaches a maximum of 2.5%, which decreases to 1.6% at  $\alpha = 0.8$ . For digital painting, the deviation at the same  $\alpha$  value ranges from 1.2% to 2%, showing smaller overall fluctuations and greater style consistency. The trends in styles like Impressionism, Cubism, and Futurism also exhibit similar stability, with the deviations gradually converging as  $\alpha$  increases, demonstrating that the model can accurately represent the target style even with strong style injection. Images generated in different artistic styles at varying strengths of the style fusion factor are shown in Figure 5:

In Figure 5, each column corresponds to a complete and continuous interval from low  $\alpha$  to high  $\alpha$ . The generated results exhibit a stable hierarchical progression in color distribution, texture density, and region morphology. The visual changes between each  $\alpha$  level maintain consistency in style direction, creating a coherent semantic flow from weak to strong style expression, rather than the rigid transition caused by linear interpolation of non-latent variables.

In the region control experiment, ControlNet was used to perform semantic segmentation on the input image, generating a region mask aligned with the latent space. This

mask was then injected into the Cross-Attention module as an additional condition, limiting the style transfer to the specified region. This experiment used a person image as input, masking the background area. This allowed the style transfer to focus on the background while preserving the main structure of the person. The generated results are shown in Figure 6:

In Figure 6, this study applied various styles to the background region of the original image. The results demonstrate that proposed model successfully replaces the background style while preserving the original character image.

#### DIVERSITY AND SUBJECTIVE APPEAL EVALUATION

To comprehensively evaluate the diversity and artistic appeal of the images generated by proposed method in the digital art style transfer task, this study conducted experiments using both objective metrics and subjective evaluation.

To evaluate the diversity performance of each model in style transfer, this study introduced three objective metrics: FID (Fréchet Inception Distance), KL Divergence (Kullback-Leibler Divergence, KLD), and diversity score. On a mixed test set constructed from WikiArt and COCO, 200 content images were selected and combined with 10 typical styles to generate 2,000 sets of images for a unified performance evaluation of various methods. The FID measures the difference in perceptual feature distribution between generated and real images, reflecting overall generation quality and style discrimination. The KL divergence assesses the consistency of the distribution of high-level feature space between the generated image and the target style image; a lower KL divergence indicates more accurate style modeling. The diversity score measures the perceptual differences in the generated results of the same content image under different style guidance, reflecting the style sensitivity and expressiveness of the model response. Comparison of the three evaluation metrics is shown in Table 4:

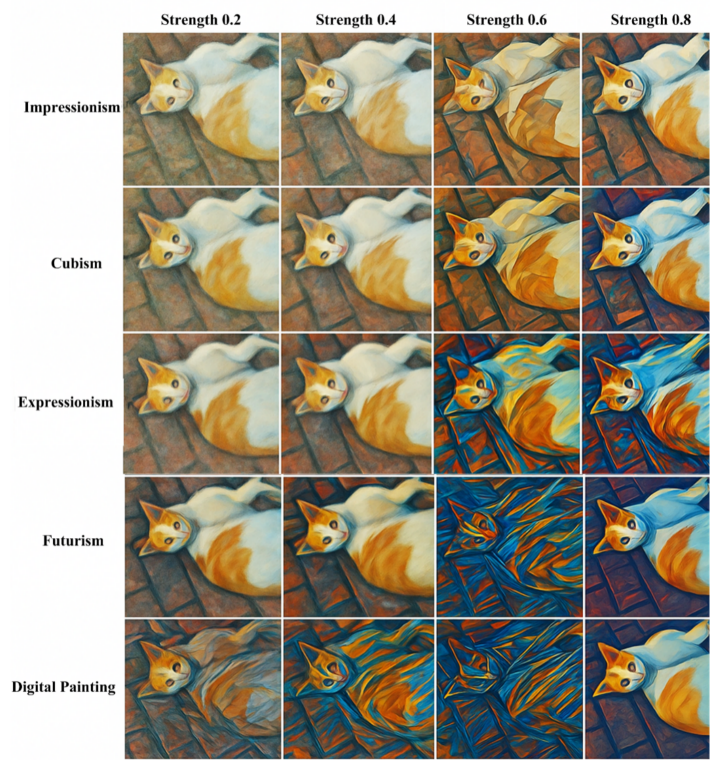


FIGURE 5. Comparison of visual effects under different style intensity controls

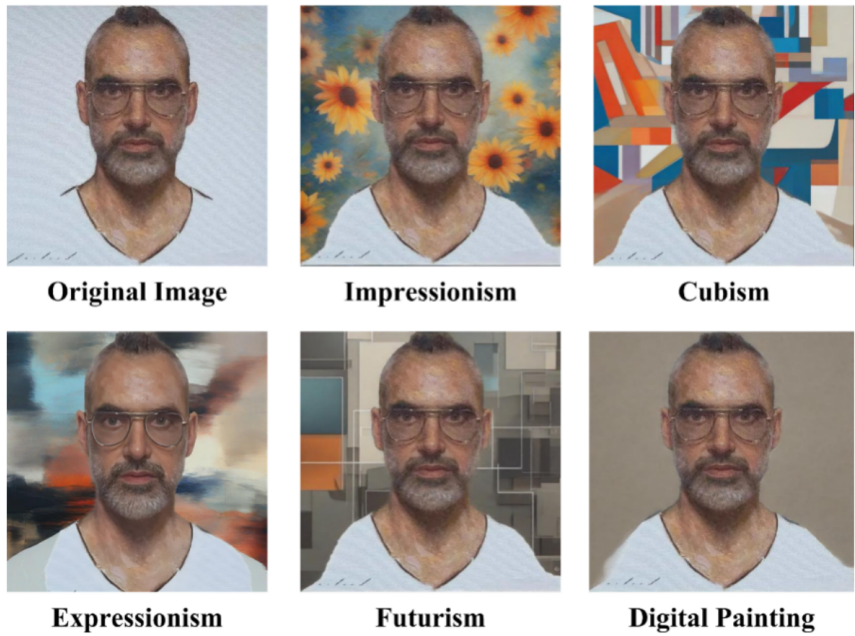


FIGURE 6. Example of Region Control Generation Results

TABLE 4. Objective Diversity Evaluation Results

| Method             | FID ↓      | KL Divergence ↓ | Diversity score ↑ |
|--------------------|------------|-----------------|-------------------|
| Control Group 1    | 38.6 ± 2.5 | 0.78 ± 0.09     | 0.21 ± 0.03       |
| Control Group 2    | 34.1 ± 2.1 | 0.65 ± 0.07     | 0.28 ± 0.04       |
| Control Group 3    | 36.4 ± 2.3 | 0.71 ± 0.08     | 0.24 ± 0.03       |
| Experimental Group | 29.8 ± 1.7 | 0.53 ± 0.05     | 0.36 ± 0.02       |

Table 4 shows the comprehensive performance of each method group on three metrics: FID, KL divergence, and diversity score. The data shows that the experimental group proposed in this paper significantly outperforms the three control groups in all metrics, demonstrating its advantages in style modeling accuracy, distribution matching, and gen-

erated diversity. The experimental group achieved the lowest FID score ( $29.8 \pm 1.7$ ) and KL divergence value ( $0.53 \pm 0.05$ ), indicating that its generated images not only have an overall distribution closer to real artistic style images, but also have stronger style expression consistency in the high-level feature space. The experimental group also achieved the highest diversity score, at  $0.36 \pm 0.02$ , demonstrating that this method can generate outputs with greater style variation given the same input content, demonstrating good control flexibility.

In the subjective attractiveness evaluation, 25 participants with backgrounds in art or design conducted a blind review of the generated images. Without knowing the source of the method, each participant rated the style transfer results of each method on a 1–5 scale based on three criteria: visual aesthetics, expressiveness of style, and overall harmony. Finally, the average preference score for each method across different style categories was calculated, as shown in Figure 7:

As shown in Figure 7, the experimental group achieved the highest user preference scores across most art styles, with particular strengths in styles like “Cubism,” “Modern Illustration,” “Baroque,” and “Post-Impressionism.” This demonstrates that this method has significant advantages in terms of visual appeal and overall harmony of stylistic expression. In contrast, Control Group 1 scored generally low across all styles, mostly between 2.9 and 3.5, with the lowest scores in “Surrealism” and “Modern Illustration,” indicating that traditional optimization methods still have significant limitations in terms of visual quality and artistic expression in style transfer. Control Groups 2 and 3 performed between the experimental group and Control Group 1, with Control Group 2 scoring relatively high in “Abstract” and “Expressionism,” demonstrating its expressive power in certain styles. Control Group 3 achieved the best performance in “Impressionism,” but its style transfer stability and consistency still need improvement. Overall, the experimental group method significantly outperformed the other control methods in subjective visual evaluation, further validating the effectiveness and user satisfaction of the proposed method in digital art style transfer tasks.

### ALGORITHMIC AND EXTREME GEOMETRIC STYLE ANALYSIS

Digital art encompasses rigorous algorithmic structures and high-frequency geometric features distinct from traditional painting aesthetics. To verify the robustness of the model under extreme geometric constraints, this study introduces input samples with strict spatial logic, selecting mathematical fractals, architectural wire frames, and trompe-l’œil geometry as test objects. These content images with precise structural blueprints are paired with radical style reference images possessing dense textures and disordered patterns, aiming to challenge the model’s limits in maintaining edge fidelity during drastic style transfer. In this comparison, the StyleGAN-NADA baseline is used as the reference

system, and the latent diffusion framework introduced in this study serves as the target method for evaluating structural preservation under extreme geometric constraints. The resulting visual generation effects and method comparisons are presented in Figure 8.

Baseline methods exhibit a clear tendency towards structural collapse when processing such high-frequency details. The rigid boundaries of architectural wireframes and the recursive features of fractals are assimilated by the chaotic textures of style maps, leading to severe edge blurring and geometric distortion. In contrast, the architecture proposed in this study, while injecting style colors and textures, fully preserves the physical properties of fine structural lines and complex tessellation patterns. This advantage stems from the effective anchoring of high-frequency information by the spatial guidance module, ensuring that the stylization process only acts on unstructured areas and tone mapping, without eroding the original geometric logic. The comparison results confirm that this framework has excellent structural integrity when handling precision-oriented professional digital design tasks.

### ABLATION ANALYSIS

When conducting an in-depth examination of the model’s style control mechanism, it is necessary to break down its role from three key dimensions: spatial adjustment, temporal adjustment, and style intensity adjustment, in order to determine the contribution of each sub-module to the overall generation quality. This section sequentially constructs variants with the region-guided mask removed, dynamic cross-attention fusion fixed, style norm constraints removed, style intensity factors removed, diffusion step adjustment removed, and a variant replacing the region mask generator with a lightweight network. A degenerate structure that simultaneously removes spatial masking and dynamic fusion is also added to observe performance changes along the three paths of structure preservation, content consistency, and style expression. All models are run under the same evaluation system, and the results are shown in Table 5.

**TABLE 5.** Ablation comparison data of style diffusion models under different structural configurations.

| Model Variant               | LPIPS ↓         | SSIM ↑          | CLIP Similarity ↑ |
|-----------------------------|-----------------|-----------------|-------------------|
| Full Model (Ours)           | $0.18 \pm 0.01$ | $0.84 \pm 0.01$ | $0.31 \pm 0.02$   |
| No Regional Mask            | $0.24 \pm 0.02$ | $0.78 \pm 0.02$ | $0.28 \pm 0.02$   |
| No Dynamic                  | $0.22 \pm 0.02$ | $0.79 \pm 0.01$ | $0.29 \pm 0.02$   |
| Cross-Attention Fusion      |                 |                 |                   |
| No Style Norm               | $0.21 \pm 0.02$ | $0.81 \pm 0.01$ | $0.26 \pm 0.02$   |
| No Strength Factor          | $0.21 \pm 0.02$ | $0.80 \pm 0.01$ | $0.27 \pm 0.02$   |
| No Step Modulation          | $0.22 \pm 0.02$ | $0.80 \pm 0.01$ | $0.28 \pm 0.02$   |
| Lightweight Mask Network    | $0.20 \pm 0.01$ | $0.82 \pm 0.01$ | $0.30 \pm 0.02$   |
| No Mask + No Dynamic Fusion | $0.27 \pm 0.02$ | $0.74 \pm 0.02$ | $0.25 \pm 0.02$   |

Judging from the data trends, each submodule played a role in the stable generation of the model to varying degrees. After removing the region masking, SSIM dropped

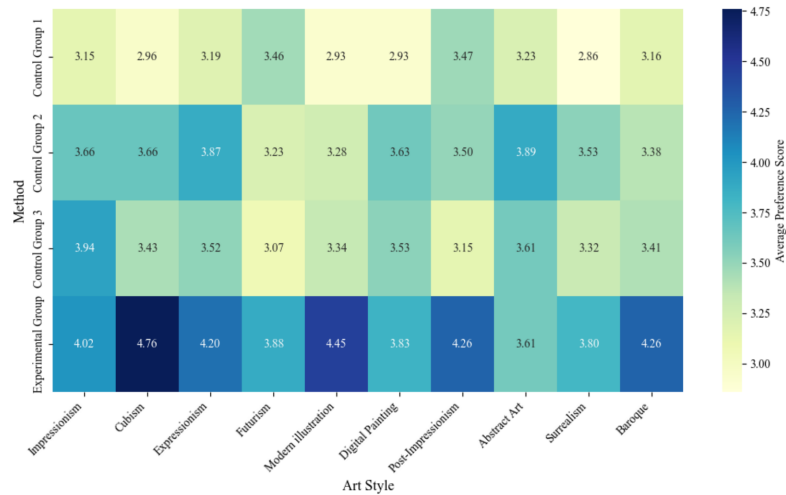


FIGURE 7. User Preference Score Distribution

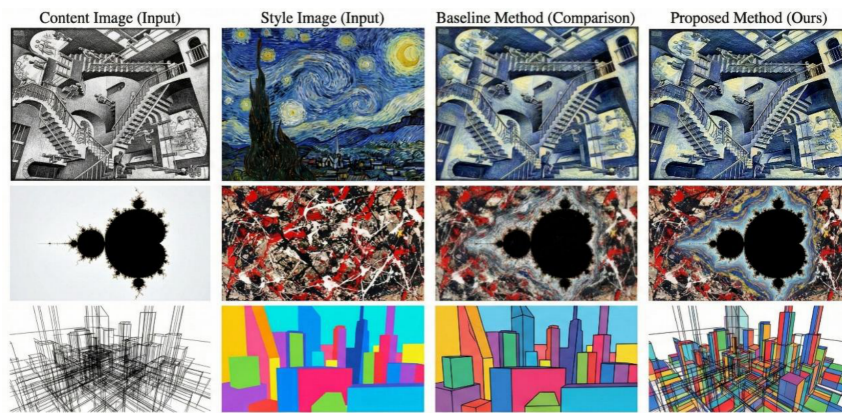


FIGURE 8. Qualitative comparison of extreme geometric and algorithmic style scenarios

to 0.78, while LPIPS rose to 0.24, indicating that the content boundaries were more easily eroded by style signals. When dynamic fusion was missing, the temporal adjustment of structural information no longer took effect, with SSIM dropping to 0.79 and LPIPS rising to 0.22, showing a different structural degradation pattern than that of region masking. When the style norm and intensity factor were missing, respectively, CLIP similarity dropped to 0.26 and 0.27, indicating that scale control of style expression was impaired. The lightweight masking network was weak in region localization, causing SSIM to drop slightly to 0.82 and LPIPS to rise to 0.20, but no significant degradation of the unmasked structure was observed. The most significant performance decline occurred in the variant where both spatial and temporal adjustment were removed, with SSIM dropping to 0.74, LPIPS rising to 0.27, and CLIP similarity dropping to 0.25. Both structural and style consistency were compromised, indicating that the synergistic relationship of the core modules in the complete model is irreplaceable.

## V. CONCLUSION

This study addresses the problem of content distortion caused by generative models in digital art style transfer, proposing a style transfer optimization method based on the latent diffusion model that effectively improves style expression while preserving content structure and semantic information.

This structural preservation improves pattern continuity across garment pieces, aligning the generated results with real panel boundaries. The style control module performs a linear mapping based on the normalized norm of the style code, and uses the mapping result to directly apply to the cross-attention output, so that the style injection intensity changes monotonically in a continuous interval, enhancing the flexibility and controllability of style transfer. The intensity factor forms a continuously evolving style adjustment sequence in this mapping mechanism, so that there is a clear semantic progression relationship between different style intensities. The texture direction, color tendency and local structure in the visual result maintain a stable growth trend during the reverse diffusion process. This mapping introduces a temporal change in the modulation applied to

the predicted noise  $\epsilon$  while the diffusion schedule remains unchanged. Furthermore, by integrating the region-guided mask generated by ControlNet, the paper limits the style to specific regions, further ensuring the fidelity of key content.

Experiments demonstrate that this method exhibits excellent style consistency and content structure preservation across a wide range of artistic style and content image combinations, significantly outperforming existing methods based on optimization, GANs, and traditional diffusion models. This study provides a new technical path for high-quality, controllable digital art style transfer, which helps alleviate the contradiction between style expression and content fidelity in traditional methods and promotes the practical application value of image generation technology in the field of artistic creation.

### AUTHOR CONTRIBUTIONS

Conceptualization – Da Liu and Li-Chao Si; methodology – Da Liu and Li-Chao Si; investigation – Da Liu and Li-Chao Si; writing-original draft preparation – Da Liu and Li-Chao Si. All authors have read and agreed to the published version of the manuscript.

### CONFLICTS OF INTEREST

The authors declare no conflict of interest.

### FUNDING

This research received no external funding.

### ACKNOWLEDGEMENTS

Not applicable.

### REFERENCES

- [1] L. Yu and Q. Zheng, "AI-Enhanced Digital Creativity Design: Content-Style Alignment for Image Stylization," *IEEE Access*, vol. 11, pp. 143964-143979, 2023, doi: 10.1109/ACCESS.2023.3342444.
- [2] X. Han, Y. Wu, and R. Wan, "A method for style transfer from artistic images based on depth extraction generative adversarial network," *Applied Sciences*, vol. 13, no. 2, pp. 867-881, 2023, doi: 10.3390/app13020867.
- [3] H. Li, L. Wang, and J. Liu, "A review of deep learning-based image style transfer research," *The Imaging Science Journal*, vol. 73, no. 4, pp. 504-526, 2025, doi: 10.1080/13682199.2024.2418216.
- [4] Y. Xu, M. Xia, K. Hu, et al., "Style Transfer Review: Traditional Machine Learning to Deep Learning," *Information*, vol. 16, no. 2, pp. 157-220, 2025, doi: 10.3390/info16020157.
- [5] S. Gayatri, G. Reddy S, T. Sirisha, et al., "SC-GAN: A Style-Conditioned Generative Adversarial Network for High-Quality Artistic Image Generation," *Macaw International Journal of Advanced Research in Computer Science and Engineering*, vol. 11, no. 1, pp. 1-10, 2025.
- [6] Y. Zhang, B. Hu, Y. Huang, et al., "Adaptive style modulation for artistic style transfer," *Neural Processing Letters*, vol. 55, no. 5, pp. 6213-6230, 2023, doi: 10.1007/s11063-022-11135-7.
- [7] M. Xu, S. Yoon, A. Fuentes, et al., "Style-consistent image translation: A novel data augmentation paradigm to improve plant disease recognition," *Frontiers in plant science*, vol. 12, pp. 773142-773158, 2022, doi: 10.3389/fpls.2021.773142.
- [8] G. Song, L. Luo, J. Liu, et al., "Agilegan: stylizing portraits by inversion-consistent transfer learning," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1-13, 2021, doi: 10.1145/3450626.3459771.
- [9] Y. Zhou, K. Yu, M. Wang, et al., "Speckle noise reduction for OCT images based on image style transfer and conditional GAN," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 139-150, 2021, doi: 10.1109/JBHI.2021.3074852.
- [10] R. Fernandez-Fernandez, G. Victores J, J. Gago J, et al., "Neural policy style transfer," *Cognitive Systems Research*, vol. 72, pp. 23-32, 2022, doi: 10.1016/j.cogsys.2021.11.003.
- [11] H. He, X. Chen, C. Wang, et al., "Diff-font: Diffusion model for robust one-shot font generation," *International Journal of Computer Vision*, vol. 132, no. 11, pp. 5372-5386, 2024, doi: 10.1007/s11263-024-02137-0.
- [12] L. Yang, J. Liu, S. Hong, et al., "Improving diffusion-based image synthesis with context prediction," *Advances in Neural Information Processing Systems*, vol. 36, pp. 37636-37656, 2023, doi: 10.52202/075280-1636.
- [13] S. Zhao, D. Chen, C. Chen Y, et al., "Uni-controlnet: All-in-one control to text-to-image diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 11127-11150, 2023, doi: 10.52202/075280-0491.
- [14] D. Mukherjee, P. Saha, D. Kaplun, et al., "Brain tumor image generation using an aggregation of GAN models with style transfer," *Scientific reports*, vol. 12, no. 1, pp. 9141-9157, 2022, doi: 10.1038/s41598-022-12646-y.
- [15] W. Hu and Y. Zhang, "Research on Artistic Pattern Generation for Clothing Design Based on Style Transfer," *J. COMBIN. MATH. COMBIN. COMPUT.*, vol. 127, pp. 4539-4550, 2025, doi: 10.61091/jcmcc127b-248.
- [16] G. Castellano and G. Vessio, "Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview," *Neural Computing and Applications*, vol. 33, no. 19, pp. 12263-12282, 2021, doi: 10.1007/s00521-021-05893-z.