

$\mathcal{Z}$ -STAR⁺: A zero-shot style transfer method adjusting style distribution

Yingying Deng¹, Xiangyu He², Fan Tang³ (✉), and Weiming Dong²

© The Author(s) 2026.

Abstract Style transfer presents a significant challenge, primarily centered on identifying an appropriate style representation. Conventional methods employ style loss, derived from second-order statistics or contrastive learning, to constrain style representation in the stylized result. However, these pre-defined style representations often limit stylistic expression, leading to artifacts. In contrast to existing approaches, we have discovered that latent features in basic diffusion models inherently contain natural style and content distributions. This allows for direct extraction of style information and seamless integration of generative priors into the content image without the need for retraining. Our method adopts dual denoising paths to represent content and style references in the latent space, subsequently guiding the content image denoising process with style latent codes. We introduce a cross-attention reweighting module that utilizes local content features to query style image information best suited to the input patch, thereby aligning the style distribution of the stylized results with that of the style image. Furthermore, we have designed a scaled adaptive instance normalization approach to mitigate global inconsistencies in color distribution between style and stylized images. Through theoretical analysis and extensive experimentation, we demonstrate the effectiveness and superiority of $\mathcal{Z}$ -STAR⁺, our diffusion-based zero-shot style transfer method using style distribution adjustment.

Keywords diffusion model; zero-shot; style transfer; attention mechanism

1 Introduction

Artists produce paintings characterized by distinctive strokes, colors, and compositions. Migrating the style of artistic paintings to a real image using image translation and stylization algorithms has received significant research attention, with numerous machine learning techniques utilized, such as convolutional neural networks (CNN), flow-based networks, visual transformers (ViT), and diffusion models.

The style transfer task is inseparable from unearthing the style distribution in input images. Prior works have proposed to use second-order statistics (such as Gram matrices, or mean and variance statistics) or contrastive learning for style representation. Such methods rely on minimizing a pre-defined style loss and optimizing network parameters, which restricts style expression to fixed patterns and leads to insufficient style rendering. In the context of artistic representation, *the contours and forms of a painting should be subject to the changing preferences of the artist's painting techniques, rather than being rigidly determined by content and style images*. This perspective necessitates a reevaluation of the role of training methodologies in style transfer processes, leading to the revelation that the generative model employed to describe the distribution of images has, in fact, already assimilated the nuanced art of transfer.

The emergence of diffusion models has catalyzed unprecedented interest in text-guided image manipulation and translation. These models demonstrate a remarkable capacity to generate artistic renditions when provided with style-related

1 School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China. E-mail: dyy15@outlook.com.

2 MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. E-mail: X. He, hexiangyu17@mails.ucas.ac.cn; W. Dong, weiming.dong@ia.ac.cn.

3 Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China. E-mail: tangfan@ict.ac.cn (✉).

Manuscript received: 2025-04-28; accepted: 2025-08-13

textual prompts. However, it is crucial to note that the efficacy of textual prompts is often limited by their inherent coarseness, which frequently proves inadequate in articulating the intricate details of desired stylistic elements. Previous methodologies, such as InST [1] and VCT [2], have attempted to utilize image-controlled diffusion models for the purpose of image style transfer and translation. However, these approaches are constrained by the requirement to optimize a unique style embedding for each input style image. This constraint not only complicates the extraction and representation of precise stylistic features, but also leads to undesirable outcomes, where the results significantly deviate from the intended input style while inadequately preserving the original content (as illustrated in Fig. 1(b)). Some researchers have trained a style-adapter network to extract style information from artistic paintings and use the image-based

style information for text prompts to ensure proper style. However, this cannot apply to specific content images. Based on IP-Adapter [3], InstantStyle-Plus [4] introduces an image extractor for content preservation, and can realize typical style transfer. However, it also faces the problem of content leakage and style deviation (as illustrated in Fig. 1(c)). In contrast to InST and VCT, where control information is encoded as text, we suggest that the basic diffusion model is capable of extracting style information directly from the desired style image and fusing it into the content image *without requiring re-training or tuning*.

In this paper, we leverage the style distribution from latent diffusion [5] and propose $\mathcal{Z}$ -STAR⁺, a zero-shot style transfer method using style distribution adjustment to generate satisfying stylized results: see Fig. 2. To obtain generative image priors, we employ dual diffusion paths to invert the style

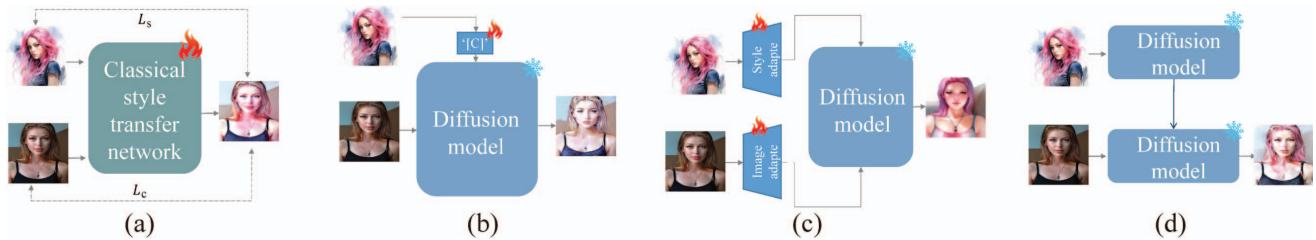


Fig. 1 Pipelines for style transfer, showing schematic stylized results generated by (a) StyTr², (b) InST, (c) InstantStyle-Plus, and (d) our method.



Fig. 2 Image style transfer results generated by our proposed $\mathcal{Z}$ -STAR⁺. Our method demonstrates versatility in generating stylized outputs across a diverse range of style and content reference pairs. When presented with a style image and a content image, our approach successfully produces stylized results that exhibit vibrant and distinctive style characteristics while accurately preserving the essential content details.

and content images. The features extracted from diffusion models inherently encapsulate both content and style information. Subsequently, we perform local style adjustment through an attention mechanism and global style adjustment via normalization operations. For local style adjustment, we propose a cross-attention reweighting strategy that adeptly manipulates local content and style information from images and seamlessly fuses them within the diffusion latent space. By leveraging a tailored attention mechanism, our diffusion model naturally addresses the constraints of content and style without necessitating additional supervision. We conduct an in-depth analysis of the style distribution differences between latent features of the style image and generated results during the denoising process. Our findings demonstrate that the proposed attention mechanism effectively aligns the style distribution, thus enhancing the fidelity of style transfer while maintaining content integrity. However, even though the style distribution is closer, there is also a global style gap between the latent features of style and generated results, which affects color performance. To address this issue, we propose a scaled adaptive instance normalization (SAIN) method to adjust the global style distribution in the first denoising step without affecting the content representation. Moreover, by considering the attention and energy loss between frame features, our method can be flexibly applied to video style transfer tasks. In summary, our main contributions are as follows:

- a flexible image style transfer method considering local and global style distribution knowledge in the diffusion model to perform image stylization without re-training/tuning,
- a reweighted attention mechanism to perform local style rendering by disentangling and fusing content/style information in the diffusion latent space,
- a scaled adaptive instance normalization method to adjust the global style distribution in the initial step, which improves color alignment without damaging content information, and
- extensions of our method to video style transfer.

Various experiments demonstrate that our method can generate outstanding style transfer results, naturally fusing and balancing content and style information.

This work extends our conference paper $\mathcal{Z}^*$: *Zero-shot style transfer via attention reweighting* which was published in CVPR 2024 [6]. Firstly, we extend the original method by considering local and global style adjustment and reveal the style distribution in the latent space of the diffusion model. Secondly, we propose a scaled adaptive instance normalization technique to better align the color distribution in stylized results with style images. Furthermore, we demonstrate the applicability of our method to video-style transfer by leveraging inter-frame correlation. We report on extensive experiments and related analyses to explain the principles underlying the effectiveness of our method.

2 Related work

2.1 Style transfer

Style transfer aims to generate an image that retains the content layout of the input content image while adopting a style similar to that of the input style image. With the development of machine learning, researchers have designed various neural networks to realize style transfer. The networks are trained by ensuring that the generated image and the content/style image exhibit content/style similarity.

Gatys et al. [7] discovered that Gram matrices of image features extracted from pre-trained VGG models could be used to represent style and propose an optimization-based method to generate stylized results by minimizing Gram matrix differences. AdaIN [8] introduces an adaptive instance normalization technique to globally adjust the mean and variance of content images to match style images. Li et al. [9] employed a whitening and coloring transform (WCT) to align the covariance of style and content images, facilitating the transfer of multi-level style patterns and yielding enhanced stylized outcomes. SANet [10] and MAST [11] leverage attention matrices between content and style features to imbue content with appropriate stylistic patterns. AdaAttN [12] introduce a per-point attentive normalization approach to achieve feature distribution alignment. ArtFlow [13] proposes a flow-based network architecture designed to minimize image reconstruction errors and mitigate recovery bias. Several recent studies utilize the long-range feature representation ability of Transformer [14] to enhance

stylization effects. StyleFormer [15] incorporates the transformer mechanism into a traditional CNN-based encoder–decoder framework. StyTr² [16] addresses the biased representation issue inherent in CNN-based style transfer methodologies by employing a purely transformer-based architecture. However, the aforementioned approaches predominantly rely on second-order statistics of the entire image to quantify style information, which may prove insufficient for comprehensive style representation. Recent advances have introduced contrastive loss as an alternative to the traditional Gram matrix-based style loss, demonstrating enhanced efficacy in processing fine-detailed stylistic patterns. CAST [17] innovatively considers the interrelationships between positive and negative examples, thereby addressing a limitation of existing style transfer models by fully utilizing extensive style information. Similarly, IEST [18] leverages the synergy of contrastive learning and external memory mechanisms to enhance visual quality.

Notwithstanding the continuous advances in existing methods, the challenge of precise style representation persists. Inaccuracies in style expression can lead to suboptimal stylized outcomes. Acknowledging this ongoing challenge, our research aims to develop a zero-shot style transfer approach that circumvents reliance on explicit style constraints, potentially offering a more robust and versatile solution to the style transfer task.

2.2 Image generation based on diffusion models

Diffusion models have gained significant popularity in the generative research domain, and work by corrupting training data by continuously adding Gaussian noise and then learning to recover the data by reversing this noise addition process. Diffusion models have demonstrated impressive results in image generation tasks [19–25]. Certain methodologies, such as Imagic [26] necessitate fine-tuning the entire diffusion model for each instruction, resulting in a process that is both time-consuming and computationally intensive. To mitigate this limitation, Prompt-to-Prompt [27] modifies the attention map between text prompts and edited images, offering a more efficient alternative. Additionally, NTI [28] proposes null-text optimization based on Prompt-to-Prompt, enabling real image editing capabilities.

In order to reduce dependence on text prompts, StyleDiffusion [29] incorporates a mapping network to invert the input image into a context embedding, which is subsequently utilized as a key in the cross-attention layers. Plug-and-Play [30] and MasaCtrl [31] focus on spatial features, leveraging self-attention mechanisms within the U-Net architecture of the latent diffusion model.

While these methods demonstrate ability for text-guided style transfer through the use of prompts such as “a pencil drawing”, simple textual descriptors may not suffice to capture the intricacies of finely detailed style patterns. Textual inversion [32] learns a style embedding using a small group of style images. T2I-adapter [33] and prompt-free diffusion [34] train a style extraction module to convert a style image to a context embedding to guide the generated result exactly. The ControlNet team also proposed a “shuffle” model [35] to realize image-guided generation. However, the content image cannot be applied to control the semantics of the results. To address this limitation, besides employing a style transfer/translation scheme that can train a style image to give a style embedding to guide the generated results, InST and VCT use an inversion process to maintain the input content semantics.

In this paper, we demonstrate that latent diffusion models can effectively perform image-guided style transfer using only style images, without the need for pseudo-text guidance or additional training. Our findings indicate that style images alone provide sufficient information for these models to achieve successful style transfer.

3 Preliminaries

3.1 Attention mechanism

The attention mechanism, initially introduced by Bahdanau et al. [36] as a powerful tool for information aggregation in neural network architectures, was subsequently adopted and refined by Vaswani et al. [37] as a fundamental building block for machine translation. The core operation of the attention mechanism is defined as

$$\text{Attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (1)$$

3.2 Diffusion models

Diffusion models are categorized as generative models

that use Gaussian noise for data sample generation. These models operate bidirectionally through forward and reverse processes. The reverse process, which is central to the model's generative capability, involves the gradual removal of this added noise. Conversely, the forward process systematically introduces noise to an initial data sample x_0 , yielding a noise sample x_t at time step t , in accordance with a pre-defined noise schedule α_t :

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}z, \quad z \sim \mathcal{N}(0, \mathbf{I}) \quad (2)$$

Furthermore, a corresponding denoising diffusion model (DDIM) reverse process can also be defined:

$$\begin{cases} x_{t-1} = \sqrt{\alpha_{t-1}}\hat{x}_0^t + \sqrt{1 - \alpha_{t-1} - \sigma_t^2}\epsilon_\theta(x_t, t) + \sigma_t z \\ \hat{x}_0^t = \frac{x_t - \sqrt{1 - \alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \end{cases} \quad (3)$$

where $\hat{x}_0^t$ is a prediction of the noise-free sample. The reverse process aims to gradually denoise $x_T \sim \mathcal{N}(0, \mathbf{I})$, where a cleaner image x_{t-1} is obtained at each step. This is accomplished by a neural network $\epsilon_\theta(x_t, t)$ that predicts the added noise z .

A prevalent approach in contemporary research involves the utilization of a U-Net architecture, integrated with an attention mechanism, to serve as the noise prediction function $\epsilon_\theta(x_t, t)$. This configuration facilitates the implementation of self-attention to capture long-range dependencies between image features, while simultaneously employing cross-attention to incorporate guiding signals from the provided text prompt. The attention mechanisms within this framework are formally expressed as

$$f_t^l = \text{Attn}(Q_t^l, K_t^l, V_t^l) \quad (4)$$

It is noteworthy that while the key and value for the l -th layer may be derived from diverse sources, such as image spatial features or textual embeddings, they consistently adhere to the standardized attention formulation. This versatility allows for the seamless integration of multimodal information within the unified attention framework.

4 Method

4.1 Overview

In this paper, we adopt local and global style distribution adjustment strategies in the pre-trained diffusion model. The attention module within Stable

Diffusion can effectively align key features K and value features V with the query Q . Previous studies, such as MasaCtrl, have used the self-attention layer to extract information from key and value features, which represent local spatial features in the DDIM inversion process, for image editing applications. Therefore, we have designed a dual-path network to generate Q , K , and V during the reverse process for content and style images respectively. Then we use a cross-attention reweighting technique to better align local content features with style features. To match the style distribution of content and style features globally, we analyze the style distribution trend in the latent space and use *scale adaptive instance normalization* to realize global feature adjustment. Finally, we also discuss video style transfer, considering the relationship between frames' features.

4.2 Dual-path networks

In our proposed method, we address a critical limitation inherent in conventional diffusion models. Specifically, we tackle the issue where text embeddings remain static throughout the reverse process, for time steps $t \in [0, T]$, despite the inherent necessity for style features to adapt dynamically to the evolving denoised stylized image. This adaptation is paramount, given that the initial stages of denoising primarily involve reconstructing the image's fundamental shape and color, while the later stages focus on refining intricate details such as contours and brushstrokes. To do so, we introduce an innovative dual-path framework that concurrently generates both the denoised style image and the content image at the same time step T . This is mathematically expressed as

$$I_s = \mathcal{G}_\theta(\epsilon_{I_s}, \{f_s\}, T), \quad I_c = \mathcal{G}_\theta(\epsilon_{I_c}, \{f_c\}, T) \quad (5)$$

This approach ensures that the features in both networks are intrinsically aligned along the temporal dimension. Specifically, given a content image I_c and a style image I_s , our objective is to derive a stylized output $\hat{I}_c$ that preserves the content of I_c while incorporating the stylistic elements of I_s . This is achieved through Eq. (6):

$$\hat{I}_c = \mathcal{G}_\theta(\epsilon_{I_c}, \{f_s, f_c\}, T) \quad (6)$$

where $\mathcal{G}_\theta(\cdot, \cdot, T)$ denotes the application of T denoising steps in the diffusion model, utilizing fixed pre-trained weights θ . The term ϵ_{I_*} represents the

noise x_T generated during the forward process by progressively introducing Gaussian noise to I_c or I_s , as delineated in Eq. (2). The notation $\{f_s, f_c\}$ refers to the spatial U-Net features derived from the style and content images.

As Fig. 3 shows, we employ DDIM inversion to invert the style and content images, thereby obtaining the diffusion trajectories $x_{[0:T]}^c$ and $x_{[0:T]}^s$. Initially, we perform global style distribution adjustment using our SAIN module. Subsequently, we use cross-attention reweighting to disentangle and fuse content and style information locally (i.e., f_c and f_s , denoted Q , K , and V) within the diffusion latent space using a U-Net at time step t . Through T denoising steps, we transform the stylized latent features $\hat{f}_c$, generated by the reweighting attention, into the final style transfer result $\hat{I}_c$.

4.3 Local style adjustment

In this section, we introduce the cross-attention reweighting used to adjust the style distribution of the input content in local fashion according to the target style. Our attention mechanism incorporates two types of attention calculation between f_s and f_c : style-cross attention for the integration of content and style features, and content self-attention for the preservation of structure. Since we use standard self-attention, we primarily discuss the proposed style-cross attention in this section.

4.3.1 Style-cross attention

The style-cross attention mechanism utilizes content features to extract the most relevant information from

style images for each input patch. It is natural to encode content information, such as the structure of an image, with the query (Q), while style information, encompassing attributes like color, texture, and object shape, is captured through the key (K) and value (V) features. Formally, the inputs of style-cross attention are features from the content latent space c and style latent space s where

$$\hat{f}_c = \text{Attn}(Q_c, K_s, V_s) = \text{Softmax}\left(\frac{Q_c K_s^T}{\sqrt{d}}\right) V_s \quad (7)$$

Despite its apparent simplicity, empirical observations reveal that the naive fusion approach delineated in Eq. (7) exhibits a propensity to overemphasize style patterns, often at the expense of preserving original content structures. Figure 4 illustrates this phenomenon through a heatmap depicting the cosine similarity of the outputs of cross-attention $\text{Attn}(Q_c, K_s, V_s)$ and self-attention $\text{Attn}(Q_c, K_c, V_c)$. The heatmap reveals a notable correlation between regions of low similarity scores and areas where content information is significantly attenuated or lost. This observation suggests that the naive fusion approach may inadvertently prioritize style transfer over content preservation, particularly in regions where the style and content features diverge substantially.

4.3.2 Simple addition

To address the aforementioned challenge, we propose a straightforward solution that enhances the content information in $\hat{f}_c$ by reincorporating content self-attention. This approach is mathematically

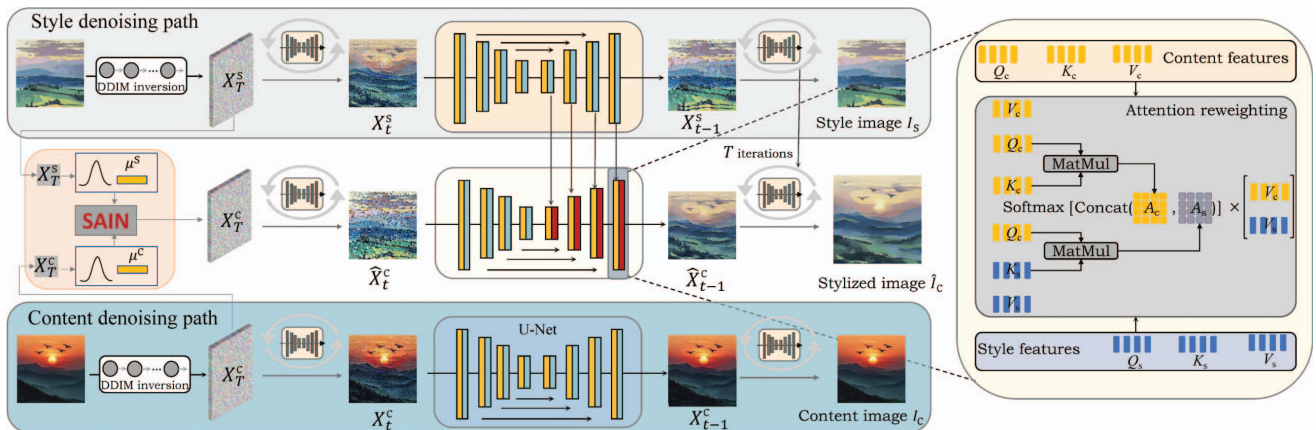


Fig. 3 Pipeline of our style transfer framework. Stylization operates in the latent space. The process begins with separate DDIM inversions for both the content and style images. We use the SAIN to adjust the global style distribution in the initial step. During the subsequent denoising process, we employ our novel cross-attention reweighting technique to integrate local style patterns into the content structure. Stylization is achieved through 30 iterative denoising steps, ultimately producing the final stylized output.

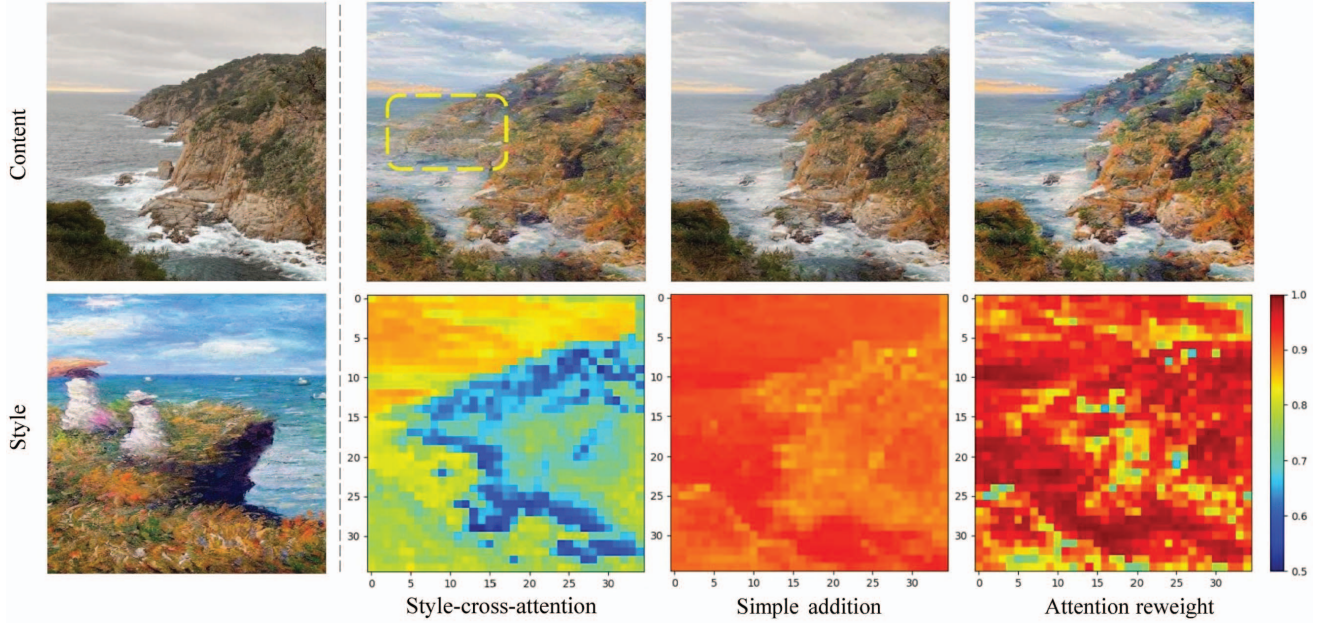


Fig. 4 Results obtained using style-cross attention demonstrate a tendency to overemphasize stylistic elements at the expense of preserving original content structures. Distortion is observed in areas where there is low correlation between self-attention and cross-attention outputs, indicating that V_s inadequately reconstructs these target regions. “Simple addition” retains an excessive amount of content features, whereas cross-attention reweighting achieves an optimal trade-off.

formulated as Eq. (8):

$$\hat{f}_c = \lambda \cdot \text{Attn}(Q_c, K_s, V_s) + (1 - \lambda) \cdot \text{Attn}(Q_c, K_c, V_c) \quad (8)$$

with $\lambda \in [0, 1]$. We find that selection of λ requires careful consideration. This is particularly evident when examining the correlation between content pixels and style elements, as quantified by the values of $\vec{q}_c K_s^T$ ($\vec{q}_c$ is the vector from Q_c). Figure 5(a) illustrates this relationship, where the blue bars represent instances of weak correlation (i.e., $\vec{q}_c K_s^T < 0$). To improve the model’s performance, it is advantageous to assign reduced attention weights

to these pixels, thereby mitigating their potential negative influence. However, the inherent properties of the softmax function, which disregards absolute magnitudes and solely amplifies differences between $\vec{q}_c K_s^T$ values, lead to counter-intuitive outcomes. As Fig. 5(b) shows, smaller $\vec{q}_c K_s^T$ values paradoxically result in larger attention weights after softmax normalization. To mitigate this issue, we introduce an additional variable, λ , as a compensatory measure. However, it is crucial to note that a pre-defined λ value cannot adequately address the diverse requirements of all content/style image pairs.

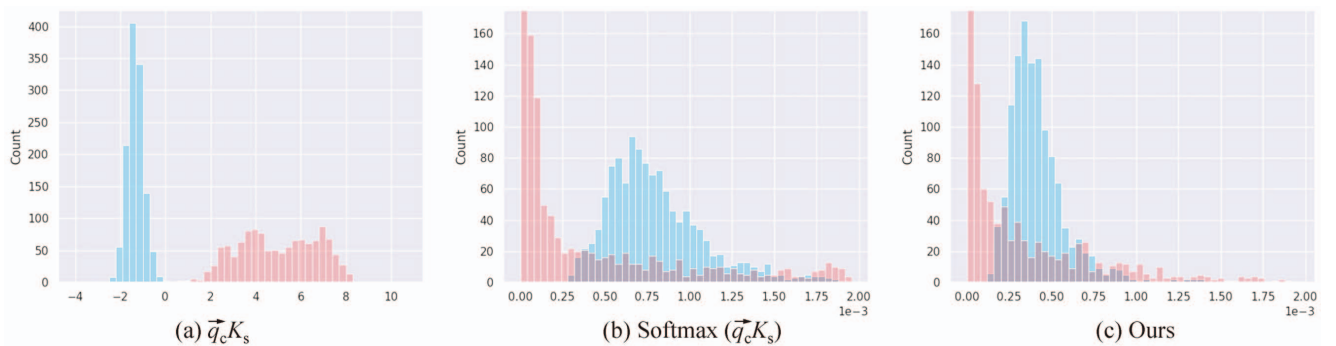


Fig. 5 Distribution of $\vec{q}_c K_s$ for two content feature points (represented by blue and red bars) before and after softmax in (a) and (b). In (c), we represent the distribution of $\vec{q}_c K_s$ normalized by Eq. (10). Note that softmax tends to overly amplify small values of $\vec{q}_c K_s$ (blue bars are shifted more significantly towards the right compared to the red bars in (b), despite the original $\vec{q}_c K_s$ values being negative). In contrast, the negatively correlated values are limited to small magnitudes before and after normalization in (c).

4.3.3 Cross-attention reweighting

While a handcrafted λ external to $\text{Attn}(\cdot, \cdot, \cdot)$ can not adapt to input images, incorporating λ within $\text{softmax}(\cdot)$ achieves the desired adaptability. By reformulating Eq. (8) in matrix form, we obtain

$$\begin{aligned}\hat{f}_c &= \left[\lambda \sigma \left(\frac{Q_c K_s^T}{\sqrt{d}} \right), (1 - \lambda) \sigma \left(\frac{Q_c K_c^T}{\sqrt{d}} \right) \right] * \begin{bmatrix} V_s \\ V_c \end{bmatrix} \\ &= A * V'^T\end{aligned}\quad (9)$$

Here, $\sigma(\cdot)$ denotes the softmax function, and each row in $A \in \mathbb{R}^{N \times 2N}$, denoted $\vec{a} \in \mathbb{R}^{2N}$, is normalized, i.e., $\vec{a} \cdot \vec{1}^T = 1$. This normalization inspires us to re-construct the matrix A in the form of applying softmax row-wise, i.e.,

$$A' = \sigma \left(\left[\lambda \cdot \frac{Q_c K_s^T}{\sqrt{d}}, \frac{Q_c K_c^T}{\sqrt{d}} \right] \right) \quad (10)$$

$$\hat{f}_c = A' * V'^T = \sigma \left(\left[\lambda \frac{Q_c K_s^T}{\sqrt{d}}, \frac{Q_c K_c^T}{\sqrt{d}} \right] \right) * \begin{bmatrix} V_s \\ V_c \end{bmatrix} \quad (11)$$

Specifically, Eq. (10) integrates λ within softmax, which inherently normalizes the output to the $[0, 1]$ range, eliminating the requirement for a $1 - \lambda$ term. In contrast to the previous attention formulation presented in Eq. (9), the newly proposed reweighted attention matrix $A' \in \mathbb{R}^{N \times 2N}$ simultaneously considers both intra-content feature differences and inter-content-style feature differences during the application of the softmax function for output normalization. The reweighted attention matrix effectively enhances significant values of both $\vec{q}_c K_c^T$ and $\vec{q}_c K_s^T$ at each pixel, while automatically suppressing small values of $\vec{q}_c K_s^T$ when the content pixel corresponding to $\vec{q}_c$ exhibits low relevance to all style pixels.

The cross-attention reweighting mechanism demonstrates enhanced flexibility and adaptability, characterized by several significant properties: (i) in scenarios where the correlation between the style and content images is weak, i.e., when each element $\vec{q}_c \vec{k}_s^T$ in $Q_c K_s^T$ approaches $-\infty$, the modified attention $\hat{f}_c = A' * V'^T$ converges to standard self-attention of content images, expressed as $\text{Attn}(Q_c, K_c, V_c)$, (ii) in instances with strong correlation between style and content images, where the maximum value of $\vec{q}_c \vec{k}_s^T$ approximates the maximum value of $\vec{q}_c \vec{k}_c^T$, and the softmax operation generates an approximate one-hot probability distribution, then $\hat{f}_c = A' * V'^T$ is equivalent to Eq. (8), and (iii) notably, Eq. (8) can be reformulated using A' as Eqs. (12) and (13):

$$\begin{aligned}\hat{f}_c &= \frac{1}{2} \text{Attn}(Q_c, K_s, V_s) + \frac{1}{2} \text{Attn}(Q_c, K_c, V_c) \\ &= \sigma \left(\left[\frac{Q_c K_s^T}{\sqrt{d}} + C, \frac{Q_c K_c^T}{\sqrt{d}} \right] \right) * \begin{bmatrix} V_s \\ V_c \end{bmatrix}\end{aligned}\quad (12)$$

where

$$C = \ln \frac{\sum_j \exp([Q_c K_c^T]_{\cdot, j})}{\sum_j \exp([Q_c K_s^T]_{\cdot, j})} \quad (13)$$

In Eq. (12), the simple addition of self-attention to cross-attention introduces an additional term C . This variable serves to magnify all elements within $Q_c K_s^T$, including small values that represent weak correlations between style and content features, which are typically considered inconsequential and should ideally be disregarded. Thus, the incorporation of C may introduce an increased level of noise into the softmax function. This noise can potentially lead to sub-optimal outcomes in the stylization process, highlighting a limitation of the simple addition approach in effectively balancing self-attention and cross-attention for style transfer tasks.

4.4 Global style adjustment

4.4.1 Approach

The cross-attention reweighting module enables effective disentanglement and fusion of content and style information. This module utilizes local content features to query the most suitable information from style images for each input patch by computing attention between content and style features. This operation significantly narrows the gap between the style distribution of generated results and that of style images. We calculate the style differences between style features and content/stylized features at different steps in diffusion models using Eq. (14).

$$\begin{cases} \mathcal{L}_{\text{style}}^t = \|G_t(\hat{I}_c) - G_t(I_s)\|_2 \\ G_t(x) = \frac{1}{C_t H_t W_t} \sum_{h=1}^{H_t} \sum_{w=1}^{W_t} \phi_t(x)_{h,w,c} \phi_t(x)_{h,w,c'} \end{cases} \quad (14)$$

where $\phi_t(\cdot)$ denotes the features extracted from the t -th step in the diffusion model, and $G_t(\cdot)$ is proportional to the uncentered covariance of the extracted features.

We visualize features at each step during the generation process and trends in style differences in Fig. 6. As the number of denoising steps increases, the style differences also increase due to the emergence of more visual information. Our cross-attention reweighting module effectively reduces the style

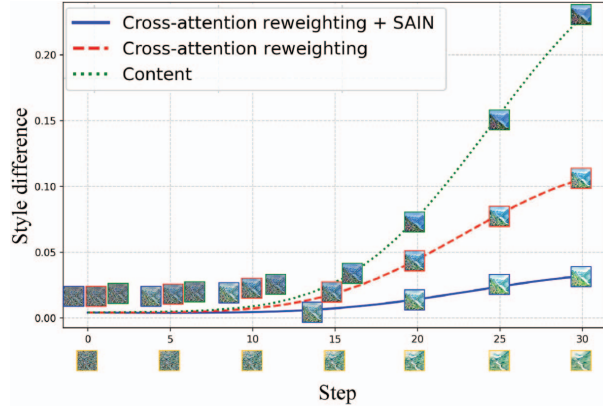


Fig. 6 Style differences between stylized/content features and style features in the denoising process. Images enclosed in yellow boxes represent visualized style features, while those in green boxes depict visualized content features. Images within red boxes show stylized features obtained through local style adjustment. Images in blue boxes demonstrate stylized features resulting from both local and global adjustments.

differences between style features and stylized features compared to the differences between style features and content features. This observation demonstrates the module’s capability to transfer style patterns from the style image to the content image. Although the style distribution is narrowed by the cross-attention reweighting, we neglect that the global content and style features have deviations in style distribution which affect the attention between them. This deviation causes the content image’s color to influence the stylization result, potentially weakening the stylization effect. This observation raises a crucial question: how can we reduce global differences and further eliminate inconsistencies in color distribution?

In Fig. 6, in the early steps, the features are almost pure noise. The style loss is small but exhibits slight differences. We assume that the implicit style is determined by the noise feature obtained by the DDIM inversion process. Adjusting the style distribution initially can gradually eliminate inconsistencies in color distribution between style and stylized images.

As established by AdaIN [8], the mean and variance of feature representations serve as effective descriptors of style characteristics. Therefore, we select a group of style images across 3 styles and use the t-SNE to visualize the variance and mean of initial noise features in Fig. 7. The mean of initial noise features is gathered according to different styles, while the variance shows less style-specific agglomeration. This empirical observation supports our theoretical

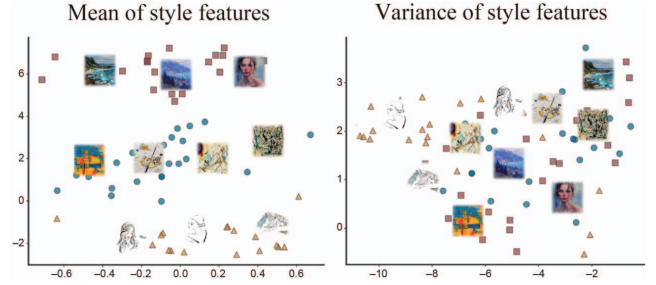


Fig. 7 t-SNE visualization using mean and variance latent features of a set of style images.

understanding that in the noise feature domain, global color distribution is primarily encoded in the mean statistics, whereas textural style elements which demonstrate weaker representation in the noise domain are predominantly associated with variance features. Therefore, we can adjust the mean of stylized features in the initial step to enhance alignment between the output and the reference style.

4.4.2 Scaled adaptive instance normalization

We apply Eq. (15) directly in the initial step to replace the mean of the content features with the mean of the style features.

$$\hat{f}_c = f_c - \mu_c + \mu_s \quad (15)$$

However, given that initial noise is highly correlated with the final content results, this mean adjustment operation may risk disrupting the input content. To address this challenge, we propose a scaled adaptive instance normalization method, which dynamically determines the degree of manipulation by measuring the distributional differences between the style and content images:

$$\hat{f}_c = f_c - w\mu_c + w\mu_s \quad (16)$$

where w is determined by the Kullback–Leibler (K–L) divergence between the initial content feature and style feature:

$$\begin{cases} w = e^{-\text{KL}(f_s \| f_c)} \\ \text{KL}(f_s \| f_c) = \sum p_s \log \frac{p_s}{p_c} \end{cases} \quad (17)$$

The notation p_s denotes the probability density function of f_s , estimated via histogram-based methods, and analogously, p_c represents the probability density function of f_c derived through the same approach. When the probability distribution difference between initial content and style is larger, the scale weight is smaller, ensuring that content is not destroyed by limiting the extent of the distribution shift. As Fig. 6 shows, with SAIN,

the style difference further decreases. The color distribution of stylized results is better aligned with style images.

From the perspective of optimal transport (OT), the transformation of content features f_c to better match the distribution of style features f_s can be formulated as finding a transport map T that minimizes the expected cost of shifting the content distribution P_c to the style distribution P_s . The Monge formulation of OT seeks to solve

$$\min_T \int c(x, T(x)) dP_c(x) \quad (18)$$

where $c(x, y)$ denotes the cost function, often taken as the squared Euclidean distance. In the special case where P_c and P_s are both Gaussian distributions with means μ_c , μ_s and identical covariances, the optimal transport map reduces to an affine transformation, specifically $T_{OT}(x) = x - \mu_c + \mu_s$. This underpins the classical mean-shifting operation in style transfer. However, when P_c and P_s differ substantially in higher-order statistics or modality, such a direct alignment can result in a high transport cost and loss of semantic content. To address this, our SAIN introduces a scaling factor w , adaptively determined by the divergence between the style and content feature distributions: $w = e^{-KL(f_s \| f_c)}$, where $KL(f_s \| f_c) = \int p_s(x) \log \frac{p_s(x)}{p_c(x)} dx$. The resulting transformation can be expressed as a convex combination of the identity map and the full mean-shifting OT map:

$$T_{SAIN}(x) = (1-w)x + w(x - \mu_c + \mu_s) = x - w\mu_c + w\mu_s.$$

This formulation ensures that when $KL(f_s \| f_c)$ is small, indicating high similarity between feature distributions, the scaling factor w converges to 1, and T_{SAIN} approximates the optimal mean alignment. Conversely, when the distributions are dissimilar and $KL(f_s \| f_c)$ is large, w converges to 0, and the transformation defaults to the identity, thereby preserving the original content features. Thus, SAIN can be interpreted as an adaptive, regularized OT map, where the degree of transformation is controlled by the statistical compatibility of the distributions, effectively balancing style transfer strength and content preservation in a theoretically principled manner.

4.5 Video style transfer

Compared to image style transfer, video style transfer presents greater challenges due to the dual

requirements of achieving satisfactory stylization effects while maintaining temporal consistency. While our proposed style distribution adjustment mechanism can effectively produce desirable stylized appearances, the persistent issue of inter-frame flickering remains unaddressed. To overcome this problem, we introduce an inter-frame attention mechanism coupled with a temporal consistency constraint to ensure coherent stylization across consecutive frames. Firstly, we extend our method to video style transfer by incorporating inter-frame attention. At each diffusion step, we obtain the latent features of the current frame f_i guided by the latent information of the reference frame. We choose the first frame and frame f_{i-1} to increase the consistency of the stylized video. See Fig. 8: we modify the attention mechanism for frame f_i by replacing its key and value with the concatenation of the corresponding key and value from frames f_{i-1} and f_0 . This approach enables the current frame to use information from both its immediate predecessor and the initial frame, promoting stylistic coherence throughout the video sequence.

$$K_i = \text{Concat}(K_0, K_{i-1}), \quad V_i = \text{Concat}(V_0, V_{i-1}) \quad (19)$$

While this inter-frame attention mechanism promotes video consistency, its effect is relatively subtle when based on a text-to-image diffusion model. Inspired by Ref. [38], we constrain consistency in the noise-free domain to further eliminate flickering in the generated video. We define an energy loss as $E_{t-1} = \|\hat{x}_{i,t-1}^0 - \hat{x}_{i-1,t-1}^0\|_2$ to minimize the difference between consecutive frames i and $i-1$ at each denoising step in the noise-free space, where $\hat{x}_{i,t-1}^0$ is predicted at each step using Eq. (3). Subsequently, we update the predicted sample f_{t-1}^i by minimizing E :

$$f_{t-1}^i = f_{t-1}^i - \nabla_{f_{t-1}^i} \cdot E_{t-1} \quad (20)$$

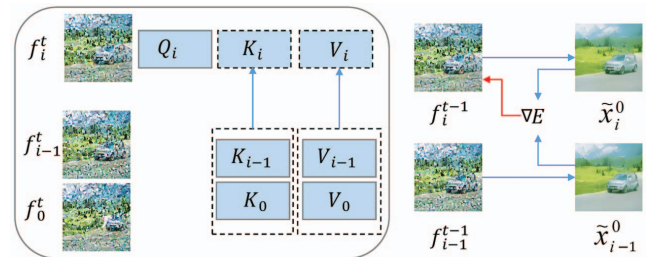


Fig. 8 Module injecting attention for various denoising steps and grad backward in the noise-free domain.

5 Experiments

5.1 Implementation details

Our study builds upon the concept of Stable Diffusion [39] and employs the v1.5 checkpoint of the model. In our experimental protocol, we configure the text prompts as null character strings to isolate their visual influence on the style transfer process. The denoising procedure is structured to encompass a total of 30 steps. We apply our novel cross-attention reweighting module between layers 10–15 during the 5-th–30-th denoising steps.

5.2 Evaluation

In this section, we compare our method to various state-of-the-art style transfer approaches: AdaAttN, IEST, StyTr², CAST, QuantArt [40], InST, VCT, InstantStyle-Plus, and StyleID [41].

5.2.1 Qualitative evaluation

Visual outcomes of various style transfer methods are presented in Fig. 9. AdaAttN struggles to capture the intended style patterns particularly evident in rows 5 and 7. IEST demonstrates improved style rendering, but it still exhibits discrepancies between



Fig. 9 Comparison to other style transfer methods. First column: content images. Second column: style images. Other columns: stylized results generated by different methods.

the generated output and the input style reference (in rows 2 and 5). CAST and StyTr² fail to accurately reproduce artistic qualities, appearing “artificial in comparison to genuine artworks” due to inadequate style loss constraints (see rows 2, 3, and 5). The results of QuantArt display a diminished style performance, attributed to an overemphasis on content preservation. InST attempts to optimize style embedding but fails to generate outputs that faithfully represent both the input style and content (rows 6 and 7). VCT has enhanced content performance but continues to struggle with style bias issues (rows 2 and 4). InstantStyle-Plus tries to learn the intrinsic content and style presentation, but introduces the content bias and style deviation in row 1, and alters facial identity in row 6. StyleID also employs a training-free approach for style transfer, however, it struggles to effectively capture and reproduce fine-grained stylistic details (in rows 1 and 2).

In comparison, our method achieves visually appealing stylized results by effectively transferring style elements, such as brush strokes and line patterns, onto the input content images. These style elements are adeptly adapted to the semantic content, as exemplified in rows 2, 3, and 6 of Fig. 9.

5.2.2 User study

To assess the effectiveness of various stylization methods, we conducted a comprehensive user study to gauge public opinion. The study involved a diverse group of 50 randomly selected participants (52% male, 48% female) aged 18–45, with both technical (68%) and non-technical (32%) backgrounds. Participants were randomly selected and screened for normal color vision. For evaluation, we used 10 content images and 10 style images, generating a total of 100 stylized outputs using our proposed method and several comparator methods. Each participant was presented with 36 sets of questions. In each set, they were shown a randomly selected content/style image pair along with corresponding stylized results from both our method and a randomly chosen comparator.

For each image pair, we provided a 30 s viewing duration to ensure detailed observation, featuring side-by-side display with zoom functionality and randomized question ordering to mitigate potential bias. Participants were asked to evaluate the stylized outputs based on three criteria: (1) which stylization result better preserved the content, (2) which stylization result exhibited better style patterns, and (3) which stylization result had the better overall effect. The study yielded a total of 5400 individual votes.

The final results are summarized in Table 1. Analysis of the data reveals that our method consistently outperformed most comparator methods in terms of content preservation and style representation. Notably, our method demonstrated superior performance in overall effect, surpassing all comparator methods, indicating that our method successfully achieves a good balance between style transfer and content preservation, resulting in visually pleasing outcomes.

To validate the authenticity and reliability of our experiment, we quantitatively analyzed the consistency between different demographic groups: male versus female, age 18–30 versus age 30–45, and professional versus non-professional participants. We counted the votes for all groups separately, with the differences between pairwise groups calculated as Eq. (21).

$$d = \frac{1}{K} \sum_{k=1}^K |M_1 - M_2| \quad (21)$$

where K denotes the total number of compared methods, and M_1 and M_2 represent the percentage of votes counted for groups compared pairwise. The difference metrics are presented in Table 2. Most notably, gender differences correspond to the largest, suggesting distinct aesthetic preferences for male and female participants. Age-related variations demonstrate slightly smaller deviation than gender-based differences. Professional background has the smallest impact on evaluation outcomes, indicating

Table 1 User study results. Numbers represent the percentage of votes preferring the corresponding method to ours, using the criteria of preservation of content and style and overall quality, respectively

	StyleID	InstantStyle-Plus	VCT	InST	QuantArt	CAST	StyTr ²	IEST	AdaAttN
Content	54.0%	14.5%	31.5%	29.5%	45.5%	35.5%	45.0%	46.5%	32.0%
Style	40.0%	35.5%	33.0%	25.0%	21.5%	49.0%	47.5%	43.5%	40.5%
Overall	43.5%	30.0%	27.5%	19.0%	30.5%	31.5%	36.5%	35.5%	34.0%

Table 2 Analysis of evaluation consistency across three demographic dichotomies: gender, age, and professional specialization

	Content	Style	Overall
Male vs. female	17.6%	14.2%	15.5%
Age18–30 vs. age30–45	16.5%	13.4%	13.8%
Prof. vs. non-prof.	15.7%	12.1%	11.7%

that our user studies are simple but easy to understand, further validating the robustness of our experimental design and results.

5.2.3 Quantitative comparisons

To conduct a quantitative assessment, our evaluation framework incorporated two key metrics: content loss and style loss. These metrics quantify the effectiveness of content preservation and style rendering, respectively, allowing for a comprehensive comparison of performance between different methodologies. Content loss and style loss are commonly adopted in style transfer tasks, and are defined as

$$\mathcal{L}_c = \frac{1}{N_l} \sum_{i=0}^{N_l} \|\phi_i(\hat{I}_c) - \phi_i(I_c)\|_2 \quad (22)$$

where $\phi_i(\cdot)$ denotes features extracted from the i -th layer in a pre-trained VGG19 and N_l is the number of layers. The style perceptual loss $\mathcal{L}_s$ is defined as

$$\mathcal{L}_s = \frac{1}{N_l} \sum_{i=0}^{N_l} \|\mu(\phi_i(\hat{I}_c)) - \mu(\phi_i(I_s))\|_2 + \|v(\phi_i(\hat{I}_c)) - v(\phi_i(I_s))\|_2 \quad (23)$$

where $\mu(\cdot)$ and $v(\cdot)$ denote the mean and variance of extracted features, respectively. The content loss is computed using the features extracted from the 5-th layer, while the style loss is calculated using the features from the 1-st to 5-th layers.

Table 3 presents the results of our comparative analysis. Our method exhibits the lowest style loss, indicating its exceptional capability for style rendering. Conversely, StyTr² achieves the lowest content loss, highlighting its effectiveness in preserving content. A notable aspect of our method is its ability to achieve comparatively low values for

both content and style loss metrics, suggesting a favorable balance between content preservation and style rendering.

We also present an inferencing speed comparison in Table 3. While our method is slower than CAST, StyTr², IEST, and AdaAttN, it achieves significantly faster inferencing than most diffusion-based approaches, such as InST and VCT. In conclusion, our approach maintains a favorable trade-off between stylization quality and speed.

5.3 Conditional control

Our method has the potential for extension to more sophisticated downstream applications, notably including conditional control utilizing binary masks. To facilitate this expansion, we introduce an auxiliary mapping function $\phi(\cdot)$ on $Q_c K_s^T / \sqrt{d}$, which enables more precise control over a designated region Ω during the image style transfer process. The augmented formulation can be expressed as

$$A' = \sigma \left(\left[\phi \left(\frac{Q_c K_s^T}{\sqrt{d}} \right), \frac{Q_c K_c^T}{\sqrt{d}} \right] \right) \quad (24)$$

Here, $\phi(x_{i,j})$ is defined as

$$\phi(x_{i,j}) = \begin{cases} -\infty, & \{i,j\} \notin \Omega \\ x_{i,j}, & \text{otherwise} \end{cases} \quad (25)$$

It is crucial to acknowledge that the direct assignment of $-\infty$ values may result in abrupt style transitions, potentially creating undesirable artificial boundaries. To mitigate this issue and achieve a more natural gradient effect, we implement a linear gradient function for $\phi(x_{i,j})$, which smoothly interpolates from $-\infty$ to $x_{i,j}$.

Moreover, the proposed framework demonstrates considerable flexibility, allowing for straightforward extension from the conventional one-to-one content-style image pair to a one-to-many configuration. This extension is readily achieved through a modification of Eq. (10). Specifically, when aiming to transfer the styles of N style images to a single content image, the equation is modified as Eq. (26):

$$A' = \sigma \left(\left[\frac{Q_c K_{s1}^T}{\sqrt{d}}, \dots, \frac{Q_c K_{sN}^T}{\sqrt{d}}, \frac{Q_c K_c^T}{\sqrt{d}} \right] \right) \quad (26)$$

Table 3 Quantitative quality and inferencing speed comparison. To evaluate the preservation of input content and style, we calculated the average values of content and style loss for the results obtained by various methods. The most favorable outcomes are highlighted in bold

	Ours	StyleID	InstantStyle-Plus	VCT	InST	QuantArt	CAST	StyTr ²	IEST	AdaAttN
$\mathcal{L}_c$	1.65	1.62	2.11	1.73	2.85	1.71	1.66	1.59	1.82	1.89
$\mathcal{L}_s$	3.63	3.73	4.18	3.98	4.70	5.50	4.18	3.79	3.86	3.78
Inference time (s/image)	44	36	19	1445	3835	0.121	0.011	0.087	0.1846	0.130

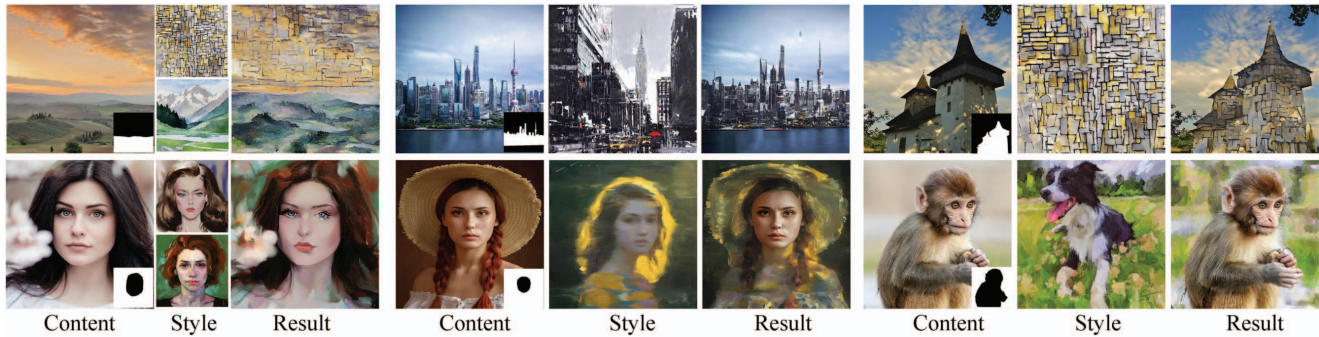


Fig. 10 Regional control results. The control map is represented by the binary mask located at the bottom right of the content reference. The first column represents multi-style regional control results; black pixels in the control map mark areas designated for style transfer using the first style image, while white pixels in the control map mark areas designated for style transfer using the second style image. Remaining columns represent single-style regional control results; white pixels mark areas designated for style transfer.

The effectiveness of our conditional control approach is visually demonstrated in Fig. 10. The leftmost column presents results obtained using the one-to-two content-style image configuration, showcasing the method's capability to blend multiple styles. Subsequent columns display outcomes from the conventional one-to-one content-style image setting, providing a basis for comparison. Notably, our conditional control strategy enables the generation of stylized results featuring smooth transitions between distinct regions. This gradual blending of styles across different areas of the image underscores the sophisticated nature of our approach, allowing for nuanced and aesthetically pleasing style transfers that avoid abrupt or artificial-looking boundaries.

5.4 Photorealistic style transfer

Our method is applicable to photorealistic style transfer tasks, with its primary objective being the comprehensive transformation of the content image's visual appearance to match the stylistic characteristics of the reference image, beyond mere color manipulation. As evidenced in the first column of Fig. 11, our approach achieves sophisticated semantic-level style adaptation, which successfully transforms the oceanic regions in the content image into glacial formations that authentically reflect the wintry aesthetic of the style reference.

5.5 Video style transfer

We conducted a comprehensive comparative analysis of our proposed method against eight baseline approaches, including state-of-the-art video style transfer techniques such as UniST [42] and MCC [43].

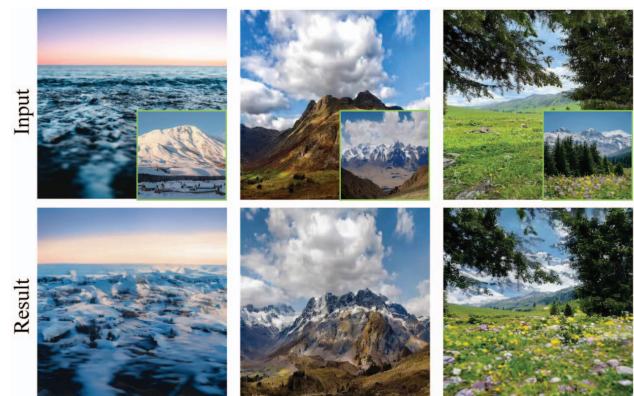


Fig. 11 Photorealistic style transfer results.

We show visual style transfer results and differences between adjacent stylized frames in Fig. 12. Without continuity constraints (w/o con.), our initial image style transfer method faces the problem of video flickering compared to contrast methods. By considering the attention of the i -th frame with the $(i - 1)$ -th frame and first frame, the consistency of stylized results slightly exceeds the baseline method, such as UniS, CAST, and IEST. Further refinement is achieved through the minimization of energy loss, which substantially reduces inter-frame differences. This approach yielded a marked improvement in temporal stability.

We also introduce the inter-frame differences $\delta = \|F_i - F_{i-1}\|$ and calculate its mean (mean_δ) and variance (var_δ) to quantitatively evaluate consistency. Table 4 presents results of this evaluation. The data demonstrate that our method consistently produces superior video results with high temporal coherence, outperforming the baseline approaches across the defined metrics.

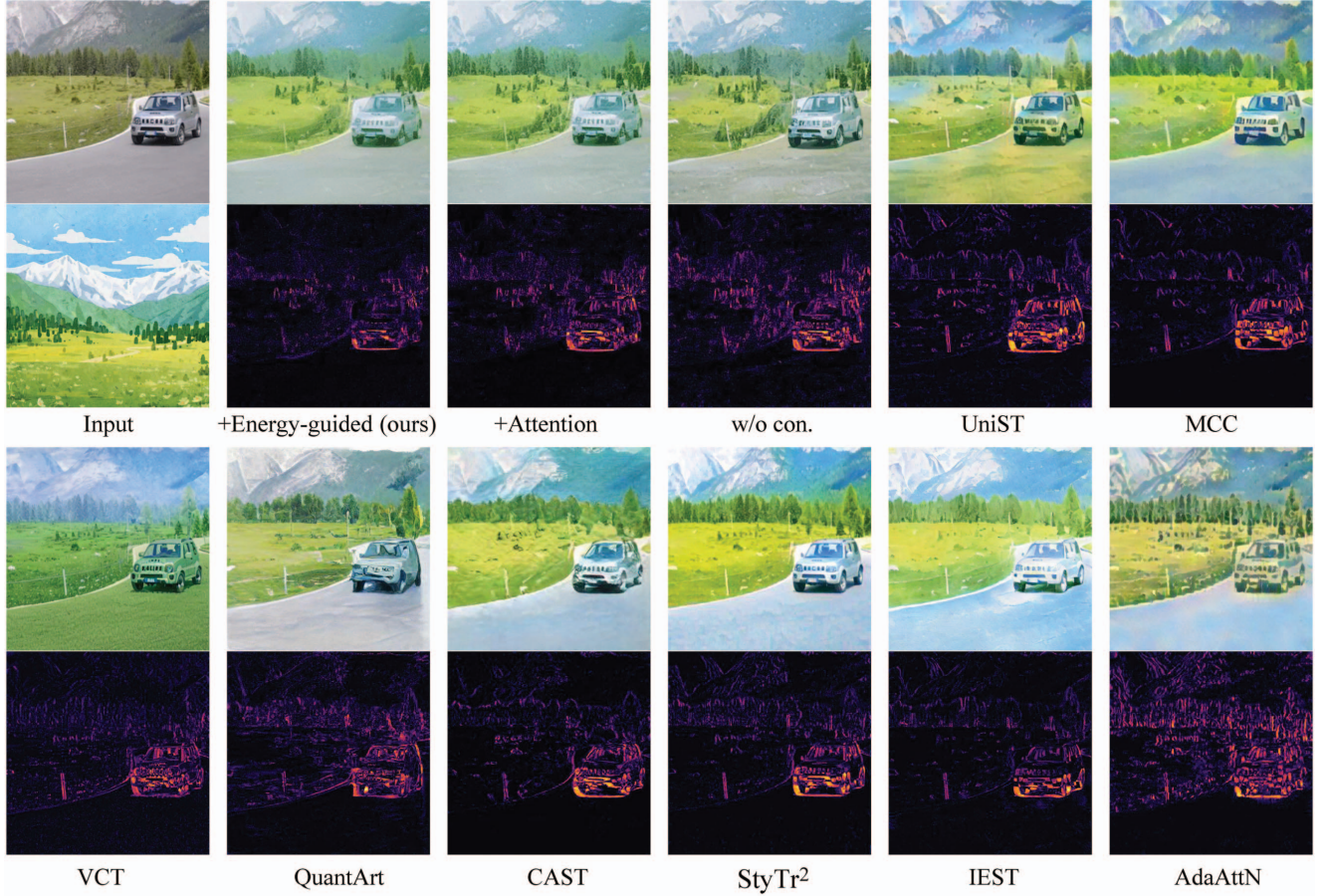


Fig. 12 Comparison of different video style transfer results. Heatmaps visualize inter-frame differences between two adjacent stylized frames.

Table 4 Quantitative consistency evaluation. Average mean_δ and var_δ of inputs and rendered clips for different methods

	Ours	+Attention	w/o con.	UniST	MCC	VCT	QuantArt	CAST	StyTr ²	IEST	AdaAttN
mean_σ	0.0362	0.0507	0.0723	0.0536	0.0491	0.0561	0.0718	0.0552	0.0605	0.0519	0.0711
var_σ	0.0041	0.0051	0.0086	0.0072	0.0075	0.0040	0.0054	0.0068	0.0055	0.0048	0.0066

5.6 Ablation and architectural study

5.6.1 Quantitative results

We conducted a comprehensive ablation study to clarify the individual contributions of each component to the overall efficacy of our method including the attention injection step/layer, cross-attention reweighting module, SAIN operation, and the style scaling factor. Table 5 presents a comprehensive quantitative comparison, offering insights into the relative impact of each component on the style transfer quality. The following sections provide a detailed qualitative analysis of each element, complementing the quantitative results and offering a more nuanced understanding of their respective roles in the style transfer process.

5.6.2 Influence of the attention injection step

This analysis focuses on the impact of the attention injection step on stylization outcomes. As Fig. 13 shows, the denoising step initiation significantly influences the stylization process. Commencing the denoising step prematurely results in a loss of content structure information, while delaying the start enhances content preservation (see the first three columns in Fig. 13). Conversely, extending the overall number of denoising steps accentuates style patterns without compromising content structure, as evidenced by the comparison between result columns 4 and 5. Furthermore, as discussed in Section 4.4, an increase in the number of denoising steps correlates with the emergence of more style visuals, contributing to more effective style rendering.

Table 5 Ablation study results

	$\mathcal{L}_c$	$\mathcal{L}_s$
Full method	1.65	3.63
Steps 0–10	2.63	4.98
Steps 10–20	1.19	4.48
Steps 20–30	1.13	4.96
Steps 5–20	1.42	4.01
Steps 15–30	1.34	4.47
Layers 0–5	2.06	5.07
Layers 5–10	1.11	6.05
Layers 0–15	2.57	3.76
Simple add	1.78	4.11
Style-across	1.93	3.55
w/o SAIN	1.61	3.70
Mean adjustment	1.88	3.59
$\lambda = 1.0$	1.13	4.61
$\lambda = 1.1$	1.47	4.13
$\lambda = 1.3$	2.23	3.49
$\lambda = 1.4$	2.29	3.45

**Fig. 13** Injecting attention module in different denoising steps.

Based on these observations, the optimal stylization results are achieved by initiating the denoising process at the 5-th step and concluding it at the 30-th step. This configuration strikes an ideal balance between preserving content structure and effectively incorporating style elements.

5.6.3 Influence of the attention injection layer

Figure 14 illustrates stylized outputs resulting from the integration of the attention module at various layers of the U-Net architecture. This experimental analysis reveals the different impacts of attention mechanisms across the network’s hierarchical structure. The indiscriminate application of attention mechanisms across all U-Net layers (layers 0–15) adversely affects the structural integrity of the content in the generated images. A more nuanced examination reveals that the low-resolution layers (layers 5–10) of the U-Net architecture effectively preserve content structure, albeit at the cost of limited style pattern transfer. In contrast, the high-resolution layers (layers 0–5 and 10–15) exhibit a superior capacity for style feature extraction.

**Fig. 14** Injecting attention module in different U-Net layers.

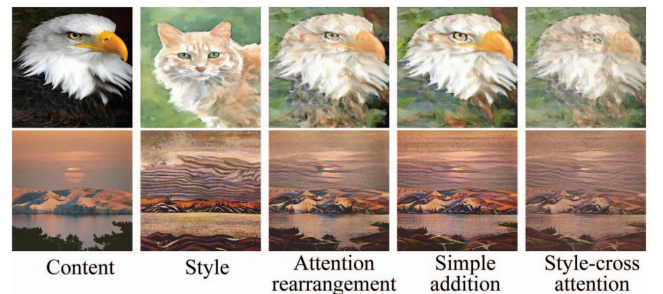
However, it is noteworthy that the high-resolution layers within the encoder section (layers 0–5) tend to compromise content structure and weaken style patterns.

Based on these observations, we conclude that the optimal strategy involves the selective injection of the attention module into the high-resolution layers of the decoder. This approach facilitates effective style transfer while maintaining the structural integrity of the content.

5.6.4 Influence of cross-attention reweighting

To evaluate the efficacy of the cross-attention reweighting module, we provide a visual representation of the stylized outcome in Fig. 15. We conducted a comprehensive ablation study to evaluate the efficacy of our proposed approach. Two key scenarios are examined: (i) the complete removal of content, as described in Eq. (7), and (ii) the utilization of only the summation operation, as in Eq. (8). For the latter, we assigned a value of 0.5 to the coefficient λ .

As evidenced in Fig. 15, the elimination of the content component results in a significant degradation of content structural preservation in the stylized outputs. Similarly, the sole incorporation of the query feature derived from the content image leads to a loss of content fidelity and inadequate style pattern transfer. In contrast, our proposed

**Fig. 15** Results of using different attention fusion strategies.

cross-attention reweighting mechanism demonstrates superior performance in achieving an optimal balance between content preservation and style transfer in the output images.

5.6.5 Influence of the SAIN module

The SAIN module aims to align global style distributions and mitigate color inconsistencies. To assess its effectiveness, we conducted experiments where the SAIN was either removed or replaced with a simple mean adjustment technique.

We show stylized results in Fig. 16 and place a color wheel in the lower right corner of the corresponding image to show the color distribution. The integration of SAIN results in stylized outputs with coherent color distributions that closely align with the target style. Conversely, the absence of SAIN leads to noticeable deviations from the desired style in terms of color fidelity. Substituting a simple mean adjustment technique for SAIN adversely affects content structure, underscoring the efficacy of our adaptive scaling approach using Kullback–Leibler divergence in preserving content details while achieving style transfer.

5.6.6 Influence of style scaling factor

We conducted a comprehensive visual analysis to evaluate the influence of varying the parameter λ in Eq. (11). Results are presented in Fig. 17, which shows the impact of different λ values on stylization outcomes. We note the robustness of the proposed



Fig. 16 Effect of the SAIN module.

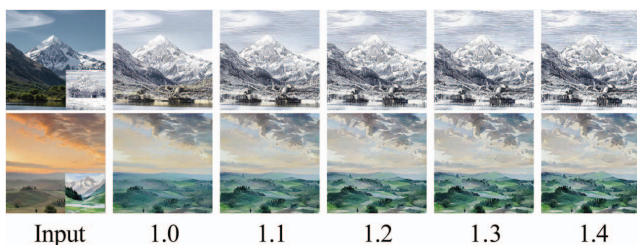


Fig. 17 Effect of varying λ in Eq. (11).

method across a spectrum of λ values. Notably, the approach exhibits consistent and satisfactory performance for λ values greater than or equal to 1.2. Based on these findings, we adopted a fixed value of $\lambda = 1.2$ for all other experiments in this study.

6 Limitations

Like a coin with two sides, evaluation of our attention-based zero-shot style transfer method presents both positive and negative outcomes. The favorable ones benefit from the utilization of a pre-trained stable diffusion model, which is also widely acknowledged to yield varying quality results based on the input data. Through our rigorous experimentation, we discovered a potential limitation of our method when transferring style information between portraits and landscape images. As Fig. 18 shows, the method struggles to adequately transfer stylistic patterns from portraits to landscape images, while conversely, when applying landscape styles to portraits, facial structural integrity is frequently compromised. These findings highlight the need for domain-specific adaptations in future iterations of this work.

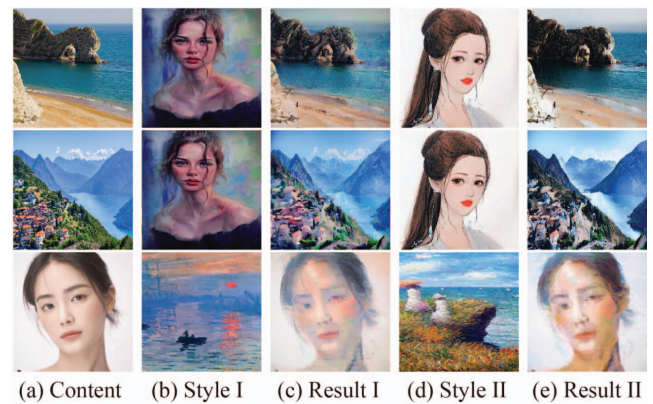


Fig. 18 Limitations in cross-domain style transfer between landscape and portrait imagery. Portrait-to-landscape fails to preserve style patterns, while landscape-to-portrait disrupts facial structures.

7 Conclusions

This study has introduced a novel zero-shot style transfer approach that adjusts the local and global style distribution in the latent space of a pre-trained diffusion model. We proposed a cross-attention reweighting module as a replacement for the conventional self-attention mechanism in diffusion

models for local style adjustment. This innovative module enables the utilization of local query features to extract style-related information from the key and value features, thereby enhancing the style transfer process. Furthermore, we introduced a scaled adaptive instance normalization technique to modulate the initial global style distribution while mitigating content leakage issues. Our method has been extended to video style transfer by considering correlation between frame features, demonstrating compatibility with both image and video diffusion models. Comprehensive experimental evaluations substantiate the superiority of our proposed approach in terms of stylization quality: it surpasses current state-of-the-art methods in the field.

Appendix

A Algorithm pipeline

We present the complete style transfer pipeline in Algorithm 1 to elucidate implementation details.

Algorithm 1 Style transfer pipeline

Input: Style initial latent noise map z_T^s , content initial latent noise map z_T^c , stylized latent noise map z_T , and null prompt P_{null} .

Output: Latent map z_0^s , z_0^c and stylized latent map z_0 .

for $t = T, T-1, \dots, 1$ **do**

$\epsilon_c, \{Q_c, K_c, V_c\} \leftarrow \epsilon_\theta(z_t^c, P_{\text{null}}, t)$

$z_{t-1}^c \leftarrow \text{Sample}(z_t^c, \epsilon_c)$

$\epsilon_s, \{Q_s, K_s, V_s\} \leftarrow \epsilon_\theta(z_t^s, P_{\text{null}}, t)$

$z_{t-1}^s \leftarrow \text{Sample}(z_t^s, \epsilon_s)$

$\{Q^*, K^*, V^*\} \leftarrow$

$\text{Style}(\{Q, K, V\}, \{Q_c, K_c, V_c\}, \{Q_s, K_s, V_s\})$

$\epsilon = \epsilon_\theta(z_t, P_{\text{null}}, t; \{Q^*, K^*, V^*\})$

$z_{t-1} \leftarrow \text{Sample}(z_t, \epsilon)$

end for

return z_0^s, z_0^c, z_0

B Proof of Eq. (12)

$$\hat{f}_c = \frac{1}{2} \text{Attn}(Q_c, K_s, V_s) + \frac{1}{2} \text{Attn}(Q_c, K_c, V_c) \quad (\text{A.1})$$

$$= \frac{1}{2} \text{Softmax} \left(\frac{Q_c K_s^T}{\sqrt{d}} \right) V_s + \frac{1}{2} \text{Softmax} \left(\frac{Q_c K_c^T}{\sqrt{d}} \right) V_c \quad (\text{A.2})$$

Note that the matrix $\hat{f}_c$ can be expressed as a combination of individual elements. To simplify the

derivation process, we focus on each element $\hat{f}_c^{i,j}$:

$$\hat{f}_c^{i,j} = \frac{1}{2} \text{Softmax} \left(\frac{[Q_c]_{i,\cdot} K_s^T}{\sqrt{d}} \right) [V_s]_{\cdot,j} \quad (\text{A.3})$$

$$+ \frac{1}{2} \text{Softmax} \left(\frac{[Q_c]_{i,\cdot} K_c^T}{\sqrt{d}} \right) [V_c]_{\cdot,j} \quad (\text{A.4})$$

where $[Q]_{i,\cdot} \in \mathbb{R}^{1 \times N}$ corresponds to the i -th row in matrix Q and $[V]_{\cdot,j} \in \mathbb{R}^{N \times 1}$ denotes the j -th column in matrix V . To facilitate the derivation, we introduce two intermediate variables:

$$x^{cs} = \frac{1}{\sqrt{d}} \cdot [Q_c]_{i,\cdot} K_s^T, \quad x^{cc} = \frac{1}{\sqrt{d}} \cdot [Q_c]_{i,\cdot} K_c^T \quad (\text{A.5})$$

where $x^{cs}, x^{cc} \in \mathbb{R}^{1 \times N}$. Expanding the softmax formulation, we express $\hat{f}_c^{i,j}$ as a weighted sum over N elements:

$$\hat{f}_c^{i,j} = \frac{1}{2} \cdot \sum_n \frac{\exp(x_n^{cs})}{\sum_m \exp(x_m^{cs})} [V_s]_{k,j} + \frac{1}{2} \cdot \sum_n \frac{\exp(x_n^{cc})}{\sum_m \exp(x_m^{cc})} [V_c]_{k,j} \quad (\text{A.6})$$

$$= \frac{1}{2} \cdot \sum_n \frac{\exp(x_n^{cs})}{\sum_m \exp(x_m^{cs})} [V_s]_{k,j} + \frac{\exp(x_n^{cc})}{\sum_m \exp(x_m^{cc})} [V_c]_{k,j} \quad (\text{A.7})$$

$$= \frac{1}{2} \cdot \sum_n \frac{\exp(x_n^{cs}) [V_s]_{k,j} \sum_m \exp(x_m^{cc})}{\sum_m \exp(x_m^{cc}) \sum_m \exp(x_m^{cs})} + \frac{\exp(x_n^{cc}) [V_c]_{k,j} \sum_m \exp(x_m^{cs})}{\sum_m \exp(x_m^{cc}) \sum_m \exp(x_m^{cs})} \quad (\text{A.8})$$

$$= \sum_n \frac{\exp(x_n^{cs}) [V_s]_{k,j} \sum_m \exp(x_m^{cc}) + \exp(x_n^{cc}) [V_c]_{k,j} \sum_m \exp(x_m^{cs})}{2 \sum_m \exp(x_m^{cc}) \sum_m \exp(x_m^{cs})} \quad (\text{A.9})$$

$$= \sum_n \frac{\exp(x_n^{cs}) [V_s]_{k,j} \sum_m \exp(x_m^{cc}) + \exp(x_n^{cc}) [V_c]_{k,j} \sum_m \exp(x_m^{cs})}{\sum_m \exp(x_m^{cc}) \sum_m \exp(x_m^{cs}) + \sum_m \exp(x_m^{cc}) \sum_m \exp(x_m^{cs})} \quad (\text{A.10})$$

$$= \sum_n \frac{\frac{\exp(x_n^{cs}) [V_s]_{k,j} \sum_m \exp(x_m^{cc})}{\sum_m \exp(x_m^{cc})} + \frac{\exp(x_n^{cc}) [V_c]_{k,j} \sum_m \exp(x_m^{cs})}{\sum_m \exp(x_m^{cs})}}{\frac{\sum_m \exp(x_m^{cc})}{\sum_m \exp(x_m^{cc})} + \frac{\sum_m \exp(x_m^{cs})}{\sum_m \exp(x_m^{cs})}} \quad (\text{A.11})$$

Upon replacing $\ln \frac{\sum_m \exp(x_m^{cc})}{\sum_m \exp(x_m^{cs})}$ by C , we obtain the resulting expression



$$\hat{f}_c^{i,j} = \sum_n \frac{\exp(x_n^{cs} + C)[V_s]_{k,j} + \exp(x_n^{cc})[V_c]_{k,j}}{\sum_m \exp(x_m^{cs} + C) + \sum_m \exp(x_m^{cc})} \quad (\text{A.12})$$

$$= \text{Softmax}([x^{cs} + C, x^{cc}]) * \begin{bmatrix} [V_s]_{\cdot,j} \\ [V_c]_{\cdot,j} \end{bmatrix} \quad (\text{A.13})$$

By substituting the derived expression for $\hat{f}_c^{i,j}$ into the matrix form, we obtain the formulation presented in Eq. (8).

C Visualization of content style correlation

Figure 19 depicts results obtained from the content-style correlation analysis, focusing specifically on the outcomes derived from $\text{Softmax}(Q_c K_s)$. This visualization capitalizes on the semantic correspondence between content queries and style keys, where higher correlation indicates shared semantic attributes. This inter-domain alignment mechanism enables precise content-style fusion through cross-attention weighting.

Our method exploits the inherent relationship between content and style by incorporating a more advanced attention mechanism, which enables the network to selectively focus on relevant style information and effectively transfer it to the content images. By leveraging the insights gained from the

content-style correlation analysis, we anticipate that our proposed approach will lead to enhanced style transfer results, surpassing the limitations of existing methods.

Availability of data and materials

We do not require training data; however, our test data are available upon request from the authors.

Author contributions

Yingying Deng contributed to the methodology and writing; Xiangyu He handled the investigation and editing; Fan Tang was responsible for validation and revision; and Weiming Dong provided financial support and further revision.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Acknowledgements

This work was supported in part by the Beijing Science and Technology Plan Project (Z231100005923033), and in part by Beijing Natural Science Foundation (L221013).

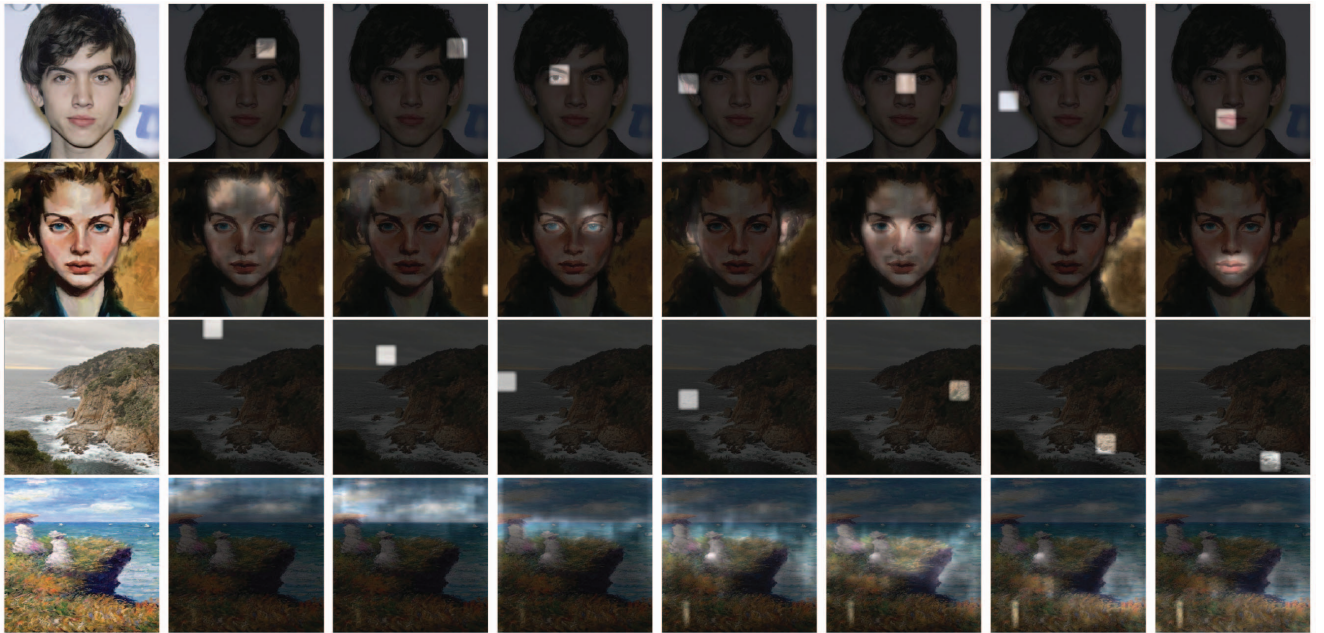


Fig. 19 Cross-attention results. The first and third rows pertain to content images, whereas the second and last rows pertain to style images. The highlighted blocks in the content images denote the content queries, while the highlighted regions in the style images correspond to the relevant regions of the corresponding content query blocks, as determined by the softmax function in cross-attention having high values.

References

- [1] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10146–10156, 2023.
- [2] Cheng, B.; Liu, Z.; Peng, Y.; Lin, Y. General image-to-image translation with one-shot image guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 22679–22689, 2023.
- [3] Rowles, C.; Vainer, S.; De Nigris, D.; Elizarov, S.; Kutsy, K.; Donné, S. IPAdapter-instruct: Resolving ambiguity in image-based conditioning using instruct prompts. In: *Computer Vision – ECCV 2024. Lecture Notes in Computer Science, Vol. 15631*. Del Bue, A.; Canton, C.; Pont-Tuset, J.; Tommasi, T. Eds. Springer Cham, 54–70, 2025.
- [4] Wang, H.; Xing, P.; Huang, R.; Ai, H.; Wang, Q.; Bai, X. InstantStylePlus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788*, 2024.
- [5] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10674–10685, 2022.
- [6] Deng, Y.; He, X.; Tang, F.; Dong, W. Z*: Zero-shot style transfer via attention reweighting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6934–6944, 2024.
- [7] Gatys, L. A.; Ecker, A. S.; Bethge, M. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2414–2423, 2016.
- [8] Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, 1510–1519, 2017.
- [9] Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; Yang, M. H. Universal style transfer via feature transforms. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 386–396, 2017.
- [10] Deng, Y.; Tang, F.; Dong, W.; Sun, W.; Huang, F.; Xu, C. Arbitrary style transfer via multi-adaptation network. In: Proceedings of the 28th ACM International Conference on Multimedia, 2719–2727, 2020.
- [11] Park, D. Y.; Lee, K. H. Arbitrary style transfer with style-attentional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5873–5881, 2019.
- [12] Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Sun, Z.; Li, Q.; Ding, E. AdaAttN: Revisit attention mechanism in arbitrary neural style transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 6629–6638, 2021.
- [13] An, J.; Huang, S.; Song, Y.; Dou, D.; Liu, W.; Luo, J. ArtFlow: Unbiased image style transfer via reversible neural flows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 862–871, 2021.
- [14] Xu, Y.; Wei, H.; Lin, M.; Deng, Y.; Sheng, K.; Zhang, M.; Tang, F.; Dong, W.; Huang, F.; Xu, C. Transformers in computational visual media: A survey. *Computational Visual Media* Vol. 8, No. 1, 33–62, 2022.
- [15] Wu, X.; Hu, Z.; Sheng, L.; Xu, D. StyleFormer: Real-time arbitrary style transfer via parametric style composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 14598–14607, 2021.
- [16] Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Pan, X.; Wang, L.; Xu, C. StyTr2: Image style transfer with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11316–11326, 2022.
- [17] Zhang, Y.; Tang, F.; Dong, W.; Huang, H. Domain enhanced arbitrary image style transfer via contrastive learning. In: Proceedings of the ACM SIGGRAPH Conference Proceedings, Article No. 12, 2022.
- [18] Chen, H.; Zhao, L.; Wang, Z.; Zhang, H.; Zuo, Z.; Li, A.; Xing, W.; Lu, D. Artistic style transfer with internal-external learning and contrastive learning. In: Proceedings of the 35th Conference on Neural Information Processing Systems, 26561–26573, 2021.
- [19] Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [20] Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, S. K. S.; Lopes, R. G.; Ayan, B. K.; Salimans, T.; Ho, J.; Fleet, D. J.; Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. In: Proceedings of the 36th International Conference on Neural Information Processing Systems, 36479–36494, 2022.
- [21] Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.



- [22] Liu, F.; Chang, X. IIDM: Image-to-image diffusion model for semantic image synthesis. *Computational Visual Media* Vol. 11, No. 2, 423–429, 2025.
- [23] Lee, S. H.; Kim, S.; Byeon, W.; Oh, G.; In, S.; Park, H.; Yoon, S. H.; Hong, S. H.; Kim, J.; Kim, S. Audio-guided implicit neural representation for local image stylization. *Computational Visual Media* Vol. 10, No. 6, 1185–1204, 2024.
- [24] Ma, H.; Yang, J.; Huang, H. Taming diffusion model for exemplar-based image translation. *Computational Visual Media* Vol. 10, No. 6, 1031–1043, 2024.
- [25] Deng, Y.; He, X.; Tang, F.; Dong, W. Z-magic: Zero-shot multiple attributes guided image creator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18390–18400, 2025.
- [26] Kavar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; Irani, M. Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6007–6017, 2023.
- [27] Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y. Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.
- [28] Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; Cohen-Or, D. Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6038–6047, 2023.
- [29] Li, S.; van de Weijer, J.; Hu, T.; Khan, F. S.; Hou, Q.; Wang, Y.; Yang, J.; Cheng, M. M. StyleDiffusion: Prompt-embedding inversion for text-based editing. *arXiv preprint arXiv:2303.15649*, 2023.
- [30] Tumanyan, N.; Geyer, M.; Bagon, S.; Dekel, T. Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1921–1930, 2023.
- [31] Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; Zheng, Y. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 22503–22513, 2023.
- [32] Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
- [33] 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. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 38, No. 5, 4296–4304, 2024.
- [34] Xu, X.; Guo, J.; Wang, Z.; Huang, G.; Essa, I.; Shi, H. Prompt-free diffusion: Taking “text” out of text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8682–8692, 2024.
- [35] Zhang, L.; Rao, A.; Agrawala, M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 3813–3824, 2023.
- [36] Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [37] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems, 5998–6008, 2017.
- [38] Ceylan, D.; Huang, C. P.; Mitra, N. J. Pix2Video: Video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 23149–23160, 2023.
- [39] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10674–10685, 2022.
- [40] Huang, S.; An, J.; Wei, D.; Luo, J.; Pfister, H. QuantArt: Quantizing image style transfer towards high visual fidelity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5947–5956, 2023.
- [41] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8795–8805, 2024.
- [42] Gu, B.; Fan, H.; Zhang, L. Two birds, one stone: A unified framework for joint learning of image and video style transfers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 23488–23497, 2023.
- [43] Deng, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Xu, C. Arbitrary video style transfer via multi-channel correlation. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 35, No. 2, 1210–1217, 2021.





Yingying Deng received her B.Sc. degree in automation from Beijing University of Science and Technology in 2017, and received her Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2022. She is an assistant professor in Beijing University of Science and Technology. Her research

interests include computational visual media and machine learning.



Xiangyu He received his B.E. degree in information security from Beijing University of Posts and Telecommunications in 2017. He earned his Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2022. His current research

interests include deep learning, image retrieval, and high-performance computing.



Fan Tang (Member, IEEE) received his B.Sc. degree from North China Electric Power University, Beijing, in 2013, and his Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2019. He is an assistant professor in the Institute of Computing Technology, Chinese

Academy of Sciences. His research interests include computer graphics, computer vision, and machine learning.



Weiming Dong (Member, IEEE) is a professor in the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences. He received his B.Sc. and M.Sc. degrees in 2001 and 2004, both from Tsinghua University. He received his

Ph.D. degree in computer science from the University of Lorraine, France, in 2007. His research interests include image synthesis, image recognition, and computational creativity.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made.

The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

To submit a manuscript, please go to <https://jcvm.org>.

