Style Injection in Diffusion: A Training-free Approach for Adapting Large-scale Diffusion Models for Style Transfer

Jiwoo Chung*, Sangeek Hyun*, Jae-Pil Heo†
Sungkyunkwan University
{wldn0202, hsi1032, jaepilheo}@g.skku.edu

Abstract

Despite the impressive generative capabilities of diffusion models, existing diffusion model-based style transfer methods require inference-stage optimization (e.g. fine-tuning or textual inversion of style) which is time-consuming, or fails to leverage the generative ability of large-scale diffusion models. To address these issues, we introduce a novel artistic style transfer method based on a pre-trained large-scale diffusion model without any optimization. Specifically, we manipulate the features of self-attention layers as the way the cross-attention mechanism works; in the generation process, substituting the key and value of content with those of style image. This approach provides several desirable characteristics for style transfer including 1) preservation of content by transferring similar styles into similar image patches and 2) transfer of style based on similarity of local texture (e.g. edge) between content and style images. Furthermore, we introduce query preservation and attention temperature scaling to mitigate the issue of disruption of original content, and initial latent Adaptive Instance Normalization (AdaIN) to deal with the disharmonious color (failure to transfer the colors of style). Our experimental results demonstrate that our proposed method surpasses state-of-the-art methods in both conventional and diffusion-based style transfer baselines. Codes are available at github.com/jiwoogit/StyleID.

1. Introduction

Recent advances in Diffusion Models (DMs) have led to breakthroughs in various generative applications such as text-to-image synthesis [30, 33, 35] and image or video editing [2, 4, 6, 14, 19, 41, 48]. One of these efforts is also applied to the task of style transfer [10, 18, 45, 47, 53]; given style and content images, modifying the style of the content image to possess the given style.

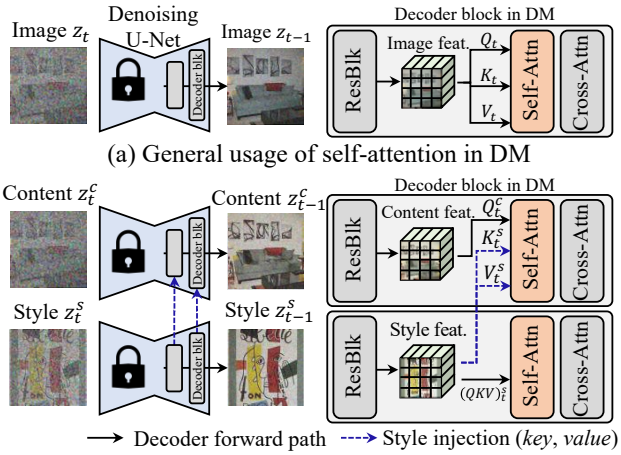
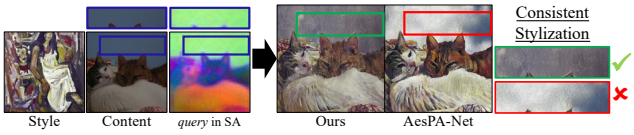


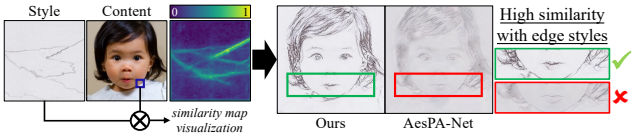
Figure 1. **Manipulation of self-attention features for style transfer.** (a) General self-attention (SA) deploys the *query*, *key*, and *value* features from a single image in both the training and inference phases. (b) At inference phase, we suggest that manipulating features of self-attention of pre-trained large-scale DM is an effective way to transfer the styles; injection of *key* and *value* of styles into SA of contents is a proper way for transferring styles. As a result, style-injected content z_{t-1}^c would maintain contents while modifying its style to resemble the target style.

General approaches for diffusion model-based style transfer leverage the generative capability of pre-trained DM. Some of these works focus on explicit disentangling style and content for interpretable and controllable style transfer [45], or inversion of the style image into the textual latent space of a large-scale text-to-image DM [53]. However, these methods additionally require gradient-based optimization for fine-tuning and textual inversion [34] for each style image, which is time-consuming. Without this issue, DiffStyle [18] introduces training-free style transfer, but they are known to be hardly applicable to Latent Diffusion Model [33] which is widely adopted for training large-scale text-to-image DM such as Stable Diffusion [33], hindering the users from taking advantage of the prominent generative ability of large-scale models.

* Equal contribution
† Corresponding author



(a) Similarity between *query* patches and style transfer results



(b) *query-key* similarity between patches of similar texture

Figure 2. **Desirable attributes of self-attention (SA) for style transfer.** (a) Visualization of *query* by PCA shows that *query* features well-reflect similarities among patches. That is, style transfer employing SA can preserve the original content, as content patches with similarities tend to receive similar attention scores from a corresponding style image patch. (b) We visualize a similarity map between the blue box (edge) *query* of the content image, and *key* of the style image. Thanks to the features representation of large-scale DM encompassing texture and semantics, a *query* exhibits higher similarity to *key*s that share a similar style, such as edges.

In this paper, we focus on extending the training-free style transfer to its application on large-scale pre-trained DM. We start from the observation of recent advances in image-to-image translation based on large-scale DM; they uncover the image editing capability of attention layers. Notably, Plug-and-play [41] shows that the residual block and the attention map of self-attention (SA) determine the spatial layout of generated images. Also, Prompt-to-Prompt [14] locally edits the image by replacing *key* and *value* of cross-attention (CA) obtained from text prompt, while keeping their original attention maps. That is, all these works suggest that 1) attention maps determine the spatial layout and 2) *key* and *value* of CA adjust the content to fill.

Inspired by the aforementioned methods, we newly argue that manipulating the SA layer is an effective way to transfer the styles (Fig. 1). Specifically, similar to CA, we substitute the *key* and *value* of SA and observe that the generated images are still visually plausible and naturally incorporate the elements of the substituted image into the original image. This observation motivates us to propose a style transfer technique based on self-attention, which combines the styles (textures) of a specific image with the content (semantics and spatial layout) of different images. Furthermore, we highlight that SA layer has desirable characteristics for style transfer. First, as shown in Fig. 2 (a), in SA-based style transfer, the content image patches (*query*) that share semantic similarities engage with a similar style (*key*), thereby maintaining the relationship among these content image patches after the transfer. Next, thanks to the powerful feature representation of large-scale DM [49], each patch of the *query* reveals higher similarity to *key*s which has similar texture and semantics. For instance, in Fig. 2 (b), we can observe that the *query* feature of content within the blue box exhibits a high similarity to the *key* features of style with similar edge texture. This encourages the model to transfer style based on the similarity of local texture (e.g. edge) between content and style.

As a result, our method aims to transfer the textures of the style image to the content images by manipulating self-attention features of pre-trained large-scale DM without any optimization. To this end, we first propose an attention-based style injection method. The basic idea of it is substituting the content's *key* and *value* of SA with those of the style image, especially layers in the latter part of decoder which are relevant to the local textures. As mentioned in above paragraph, exchanged styles are well aligned with the content and texture of original image, exploiting the similarity-based attention mechanism. With the proposed style injection, we observe that the local texture patterns are successfully transferred, but there still are remaining problems such as disruption of original content and disharmonious colors. To handle these problems, we additionally propose the following techniques; query preservation, attention temperature scaling, and initial latent AdaIN. Query preservation makes the reverse diffusion process to retain the spatial structure of original content by preserving the *query* of the content image in the SA. Attention temperature scaling also aims to keep the structure of content by dealing with the blurred self-attention map introduced from the substitution of *key*. Lastly, initial latent AdaIN corrects inharmonious color problem, referring that the color distribution of style images is not properly transferred, by modulating the statistics of initial noise in the diffusion model.

Our main contributions are summarized as follows:

- We propose a style transfer method exploiting the large-scale pre-trained DM by simple manipulation of the features in self-attention; substituting *key* and *value* of content with those of styles without any requirements of optimization or supervision (e.g. text).
- We further improve the naive approach for style transfer to properly adapt the styles by proposing three components; query preservation, attention temperature scaling, and initial latent AdaIN.
- Extensive experiments on the style transfer dataset validate the proposed method significantly outperforms previous methods and achieves state-of-the-art performance.

2. Related Work

2.1. Diffusion Model-based Neural Style Transfer

Neural style transfer [11, 24–26, 29, 31, 43, 44, 52] is an example-guided image generation task that transfers the style of one image onto another while retaining the content of the original. In the realm of diffusion models, neural style transfer has evolved by leveraging the generative capability

of pre-trained diffusion models. For instance, InST [53] introduced a textual inversion-based approach, aiming to map a given style into corresponding textual embeddings. StyleDiffusion [45] aimed to disentangle style and content by introducing CLIP-based style disentanglement loss for fine-tuning DM for style transfer. Also, several approaches utilize the text input as a style condition or for determining the content to synthesize [10, 47].

Conversely, DiffStyle [18] proposed a training-free style transfer method that leverages *h-space* [23] and adjusts skip connections for effectively conveying style and content information, respectively. However, when DiffStyle is applied to Stable Diffusion [33, 42], their behavior is quite different from typical style-transfer methods; not only textures but also semantics such as spatial layout are also changed.

To address these limitations, we propose a novel algorithm that harmoniously merges style and content features within the self-attention layers of Stable Diffusion without any optimization process.

2.2. Attention-based Image Editing in DM

Following the remarkable advances achieved by pre-trained text-to-image DMs [32, 42], there have been numerous image editing works [2, 6, 19, 37] utilizing these DMs. Notably, Prompt-to-Prompt [14] proposed text-based local image editing by manipulating the cross-attention map. Specifically, they observe that cross-attention largely contributes to modeling the relation between the spatial layout of the image to each word in the prompt. Hence, they substitute the original words and cross-attention map with desirable ones, obtaining edited images matched with text conditions. Subsequently, Plug-and-play [41] introduces text-guided image-to-image translation method. They found that the spatial features (i.e. feature from residual block) and self-attention map determine the spatial layout of the synthesized image. Thus, while generating a new image with the given text condition, they guide the diffusion model with features and attention map from the original image for preserving the original spatial layout. Recently, MasaCtrl [3] proposes mutual self-attention control for consistent image editing using text prompts. In detail, they retain the source image's key and $value$ of the self-attention layers, while conditioning the model with desired text prompts.

Along with these works, we recognize the potential of attention maps in representing spatial information. However, different from the aforementioned methods concentrating on exploiting textual condition, we focus on conditioning by style and content images composed of two images from distinct styles. By combining the features in self-attention layers of both style and content images with precise adjustment of statistics in intermediate representations, we transfer the texture of the content image to the given style.

3. Background

Latent Diffusion Model (LDM) [33] is a type of diffusion model trained in the low dimensional latent space to focus on semantic bits of data and reduce computation costs. Given an image $x \in \mathbb{R}^{H\times W\times 3}$, the encoder $\mathcal{E}$ encodes x into the latent representation $z \in \mathbb{R}^{h\times w\times c}$ and decoder reconstructs the image from the latent.

With the pretrained encoder, they encode the entire images in the dataset and train a diffusion model on latent space z, by predicting noise ϵ from the noised version of latent z_t at time step t. The corresponding training objective is

$$L_{\text{LDM}} = \mathbb{E}_{z,\epsilon,t}[\|\epsilon - \epsilon_\theta(z_t, t, y)\|_2^2], \tag{1}$$

where $\epsilon \in \mathcal{N}(0,1)$ is a noise, t is the number of time steps which uniformly sampled from $\{1, ..., T\}$, y is a condition, and ϵ_θ is a neural network which predicts the noise added to the z.

In our work, we utilize Stable Diffusion (SD) [33] which is the only publicized large-scale pre-trained DM. In the case of SD, y is a text, and ϵ_θ is a U-Net architecture in which a block for each resolution comprises a residual block, self-attention block (SA), and cross-attention block (CA), sequentially. Among these modules, we focus on the SA block to transfer the styles, as discussed in Sec. 1. Given a feature ϕ after the residual block, the self-attention block performs as follows:

$$\begin{aligned} Q = W_Q(\phi), K = W_K(\phi), V = W_V(\phi), \\ \phi_{\text{out}} = \text{Attn}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V, \end{aligned} \tag{2}$$

where d denotes the dimension of the projected *query*, and $W_{(\cdot)}$ is a projection layer. Note that, we don't use any text conditions, so the variable y is always an empty text prompt ("").

4. Method

In this paper, we aim to solve artistic style transfer by leveraging the generative capability of a pre-trained large-scale text-to-image diffusion model. Briefly, artistic style transfer is the task of modifying the style of a given content image I^c to that of style image I^s. Then, the stylized image I^{cs} would maintain the semantic content of I^c while its style (such as texture) is transferred from I^s. For simplicity, we skip the explanations about the encoding and decoding process of the autoencoder in the LDM. Instead, we focus on elaborating the proposed method in the aspect of the diffusion process. Thus, in the following sections, we regard the content, style, and stylized images same as their encoded counterparts z_0^c, z_0^s, and z_0^{cs}.

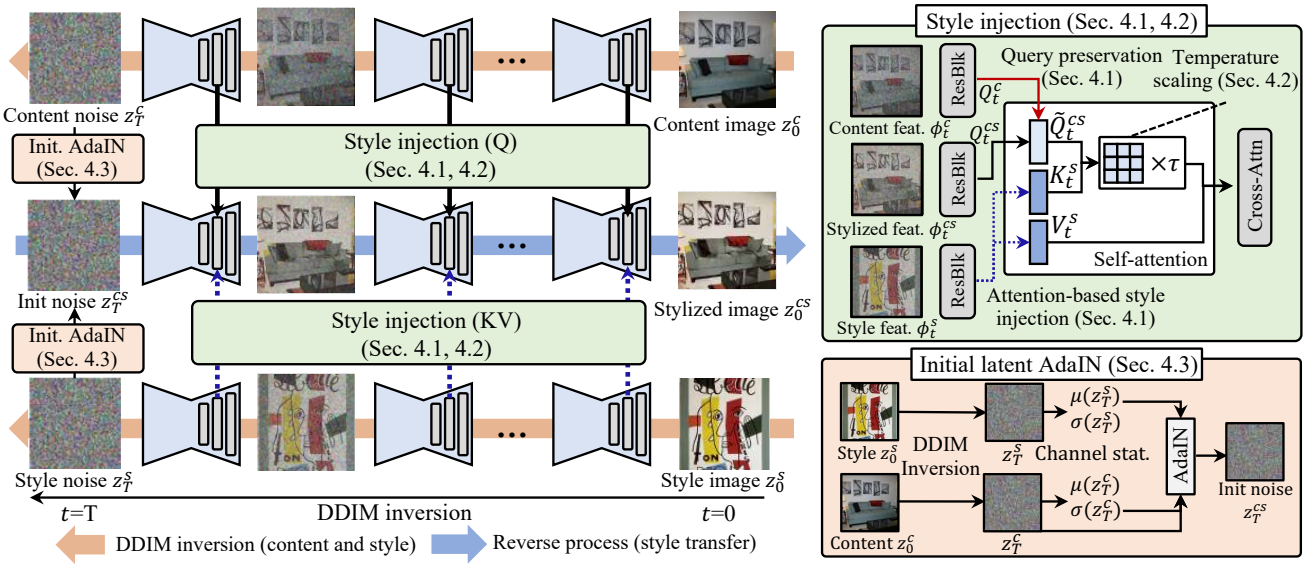


Figure 3. **Overall framework.** (Left) Illustration for the proposed style transfer method. We first invert content image z_0^c and style image z_0^s into the latent noise space as z_T^c and z_T^s, respectively. Then, we initialize the initial noise of stylized image z_T^{cs} from initial latent AdaIN (Sec. 4.3) which combines the content and style noise, z_T^c and z_T^s. While performing the reverse diffusion process with z_T^{cs}, we inject the information of content and style by attention-based style injection (Sec. 4.1) and attention temperature scaling (Sec. 4.2). (Right) Detailed explanation of style injection and initial noise AdaIN. Style injection is basically the manipulation of self-attention (SA) layer during the reverse diffusion process. Specifically, at time step t, we substitute the *key* (K_t^{cs}) and *value* (V_t^{cs}) in SA of stylized image with those of style features, K_t^s and V_t^s, from identical timestep t. At the same time, we preserve the content information by blending the *query* of content Q_t^c and *query* of stylized image Q_t^{cs}. Finally, we scale the magnitude of the attention map to deal with the magnitude decrease that the substitution of feature leads to. Initial latent AdaIN produces the initial noise z_T^{cs} by combining style noise z_T^s and content noise z_T^s. Specifically, we modify the channel statistics of z_T^c to resemble the statistics of z_T^s and regard it as z_T^{cs}. We observe this operation enables us to keep the spatial layout of content image while well-reflecting the color tones of a given style image.

4.1. Attention-based Style Injection

We start from the observation in previous image-to-image translation methods, especially Prompt-to-Prompt [14]. The key idea of their method is changing the text condition for cross-attention (CA) while keeping the attention map. Since the attention map affects the spatial layout of output, substituted text conditions determine what to draw in the generated image, and these conditions are actually *key* and *value* in CA. Inspired by them, we manipulate the features in self-attention layer as like cross-attention, regarding the features from style image I^s as the condition. Specifically, in the generation process, we substitute the *key* and *value* of content image with those of style for transferring the texture of style image into the content image.

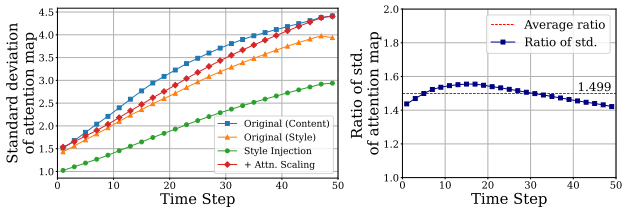
To this end, we first obtain the latent for content and style images with DDIM inversion [39], and then collect the SA features of style image over the DDIM inversion process. Specifically, for pre-defined timesteps $t = \{0, ..., T\}$, style and content images z_0^c and z_0^s are inverted from image ($t = 0$) to gaussian noise ($t = T$). During DDIM inversion, we also collect *query* features of content (Q_t^c) and *key* and *value* features of style (K_t^s, V_t^s) at every time steps.

After that, we initialize stylized latent noise z_T^{cs} by copying content latent noise z_T^c. Then, we transfer the target style to the stylized latent by injecting the *key* K_t^s and *value* V_t^s collected from the style into SA layer, instead of the original *key* K_t^{cs} and *value* V_t^{cs}, when performing the entire reverse process of stylized latent z_t^{cs}. However, only applying this substitution can lead to content disruption, since the content of stylized latent would be progressively changed as attended *value* changes. Hence, we propose query preservation to maintain original content. Simply, we blend *query* of stylized latent Q_t^{cs} and that of content Q_t^c for the entire reverse process. These style injection and query preservation processes at time step t are expressed as follows:

$$\tilde{Q}_t^{cs} = \gamma \times Q_t^c + (1 - \gamma) \times Q_t^{cs}, \tag{3}$$

$$\phi_{\text{out}}^{\text{cs}} = \text{Attn}(\tilde{Q}_t^{cs}, K_t^s, V_t^s), \tag{4}$$

where γ is degree of blending in range of $[0, 1]$. In addition, we apply these operations on the latter layers of decoder (7-12$^{\text{th}}$ decoder layers in SD) relevant to local textures. We also highlight that the proposed method can adjust the degree of style transfer by changing query preservation ratio γ. Specifically, higher γ maintains more content, while lower γ strengthens effects of style transfer.



(a) Std. of attention map (b) Ratio of std. of attention map

Figure 4. **Visualization of the standard deviation of attention map before softmax.** (a) Attention-based style injection reduces the standard deviation of self-attention map. Original denotes SA maps from the generation process without style injection. We use both style and content images for generation. (b) We compute the ratio between attention maps w/ and w/o style injection. For the std of original image, we use averaged std. of content and style.

4.2. Attention Temperature Scaling

Attention map is computed by scaled dot-product between *query* and *key* features. During training, *query* and *key* features in the SA layer originate from an identical image. However, if we substitute the key features with those of style images, the magnitude of similarity would be overall lowered as style and content are highly likely to be irrelevant. Thus, the computed attention map can be blurred or smoothed, and it would further make output images unsharp, which is detrimental to capturing both content and style information.

To quantify this issue, we measure the standard deviation of attention map, while ablating the attention-based style injection. In detail, we calculate the attention map before applying softmax, which is scaled-dot product between *query* and *key*. As shown in Fig. 4 (a), we validate that this style injection tends to lower the standard deviation of the attention map over the entire timesteps. That is, attention maps after softmax with style injection would be overly smooth.

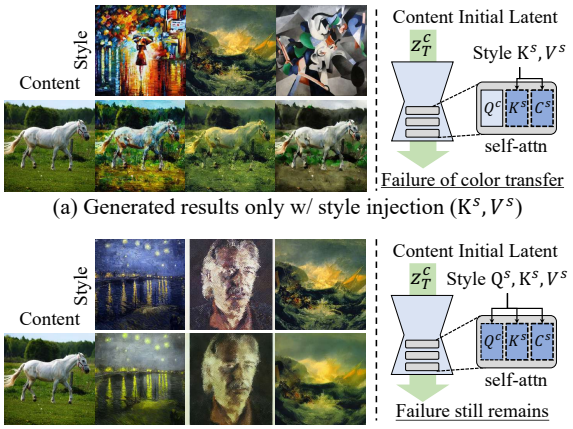
To rectify the attention map sharper, we introduce an attention temperature scaling parameter. In detail, we multiply the attention map before softmax by a constant temperature scaling parameter τ larger than 1. Thus, the attention map after softmax would be sharper than its original values. The modified attention process is represented as follows:

$$\text{Attn}_\tau(\tilde{Q_t^{cs}}, K_t^s, V_t^s) = \text{softmax}(\frac{\tau\, \tilde{Q_t^{cs}}(K_t^s)^T}{\sqrt{d}}) \cdot V_t^s, \tau > 1. \tag{5}$$

We use $\tau = 1.5$ as a default setting, which is the average ratio over entire timesteps. As reported in Fig. 4 (b), we confirm that it effectively calibrates the standard deviation of attention map similar to its original values.

4.3. Initial Latent AdaIN

In artistic style transfer, the color tone generally takes up a significant portion of the style information. In this context, we observe that the style transfer only with attention-based style injection often fails in terms of capturing the color tone of the given style. As shown in Fig. 5 (a), textures and local patterns are successfully transferred to the content image while the color tone of the content image still remains. Furthermore, even with injecting the *query*, *key*, and *value* of styles, the resulting images still preserve the color tone of the content, as shown in Fig. 5 (b).



(a) Generated results only w/ style injection (K^s, V^s)

(b) Generated results only w/ style injection (Q^s, K^s, V^s)

Figure 5. **Generated results only w/ style injection.** (a) We observe that generated images only with attention-based style injection do not harmonize with the given style in the aspect of color tone. (b) To identify the effects of every feature in SA on color tones, we additionally include *query* in the style injection process. However, color tones still resemble those of content, concluding features in self-attention have less effect on the color tones.

As substituting the self-attention features has less effects to color tone, we analyze the other vital part of DM; initial latent noise. One of the recent discoveries in DM is that the DM struggles to synthesize purely white or black images [13]. Instead, they tend to generate images of median color as the initial noise is sampled from zero mean and unit variance. Thus, we hypothesize the statistics of initial noise largely affect the colors and brightness of generated images.

Based on this assumption, we attempt to use the initial latent of style z_T^s for the style transfer process. However, if we simply start to generate the image from style latent z_T^s, the structural information of synthesized results also follows the style image and loses the structure of the content. To harness valuable information in both initial latents, we consider that the tone information is intricately connected with the channel statistics of the initial latent, following the principle underlying Style Loss [11] and AdaIN [17]. Thus, we employ AdaIN to modulate the initial latent for effective tone information transfer, represented as:

$$z_T^{cs} = \sigma(z_T^s)\left(\frac{z_T^c - \mu(z_T^c)}{\sigma(z_T^c)}\right) + \mu(z_T^s), \tag{6}$$

where $\mu(\cdot), \sigma(\cdot)$ denote channel-wise mean and standard deviation, respectively. Based on this, the initial latent z_T^{cs}

Metric	Ours	AesPA-Net	CAST	StyTR^2	EFDM	MAST	AdaAttn	ArtFlow	AdaConv	AdaIN	DiffuseIT	InST	DiffStyle
ArtFID ↓	**28.801**	31.420	34.685	30.720	34.605	31.282	30.350	34.630	31.856	30.933	40.721	40.633	41.464
FID ↓	**18.131**	19.760	20.395	18.890	20.062	18.199	18.658	21.252	19.022	18.242	23.065	21.571	20.903
LPIPS ↓	**0.5055**	0.5135	0.6212	0.5445	0.6430	0.6293	0.5439	0.5562	0.5910	0.6076	0.6921	0.8002	0.8931
CFSD ↓	**0.2281**	0.2464	0.2918	0.3011	0.3346	0.3043	0.2862	0.2920	0.3600	0.3155	0.3428	0.6759	0.2819

Table 1. Quantitative comparison with conventional (3$^{\text{rd}}$-11$^{\text{th}}$ columns) and diffusion model baselines (12$^{\text{th}}$-14$^{\text{th}}$ columns)

preserves content information from z_T^c while aligning the channel-wise mean and standard deviation with z_T^s.

5. Experiments

5.1. Experimental Settings

We conduct all experiments in Stable Diffusion 1.4 pre-trained on LAION dataset [36] and adopt DDIM sampling [39] with a total 50 timesteps ($t = \{1, ..., 50\}$). For default settings for hyperparameters, we use $\gamma = 0.75$ and $\tau = 1.5$, if they are not mentioned separately.

5.2. Evaluation Protocol

Conventional style transfer methods typically utilize Style Loss [11] as both training objective and evaluation metric, so their results tend to overfit the Style Loss. Thus, for a fair comparison, we employ a recently proposed metric, ArtFID [46] which evaluates overall style transfer performances with consideration of both content and style preservation and also is known as strongly coinciding with human judgment. Specifically, ArtFID is computed as $(\text{ArtFID} = (1 + \text{LPIPS}) \cdot (1 + \text{FID}))$. LPIPS [50] measures content fidelity between the stylized image and the corresponding content image, and FID [15] assesses the style fidelity between the stylized image and the corresponding style image.

Dataset. Our evaluations employ content images from *MS-COCO* [27] dataset and style images from *WikiArt* [40] dataset. All input images are center-cropped to 512 × 512 resolution. Also, for quantitative comparison, we randomly selected 20 content and 40 style images from each dataset, yielding 800 stylized images as StyTR^2 [9] has done.

Content Feature Structural Distance (CFSD). In the style transfer evaluation, the assessment of content fidelity often relies on the LPIPS distance. However, since LPIPS utilizes the feature space of AlexNet [20] pre-trained for classification task on ImageNet [7], which is known as texture-biased [12]. Thus, the style information of the images can affect the LPIPS score. To mitigate this style influence, we additionally introduce Content Feature Structural Distance (CFSD) which is a distance measure that only considers the spatial correlation between image patches.

In detail, we first define the correlation map between image patch features as follows. For a given image I, we obtain feature maps $F \in \mathbb{R}^{hw \times c}$, which is the output feature of *conv3* in VGG19 [38]. Then, we calculate the patch similarity map $M = F \times F^T, M \in \mathbb{R}^{hw \times hw}$, which is a similarity map between every pair of features in F. After that, for computing the distance between two patch similarity maps, we model the similarity between a single patch and the others as a probability distribution by applying softmax operation. Finally, the correlation map is represented as $S = [\text{softmax}(M_i)]_{i=1}^{hw}, S \in \mathbb{R}^{hw \times hw}$, where $M_i \in \mathbb{R}^{1 \times hw}$ is a similarity map between i^{th} patch and the other patches.

Then, CFSD is defined as KL-divergence between two correlation maps. In our case, we compute CFSD between the correlation map of the content (S^c) and stylized images (S^{cs}) as follows:

$$\text{CFSD} = \frac{1}{hw} \sum_{i=1}^{hw} D_{\text{KL}}(S_i^c || S_i^{cs}), \tag{7}$$

5.3. Quantitative Comparison

We evaluate our proposed method through comparison with twelve state-of-the-art methods, including nine conventional style transfer methods (AesPA-Net [16], CAST [52], StyTR^2 [9], EFDM [51], MAST [8], AdaAttN [28], ArtFlow [1], AdaConv [5], AdaIN [17]) and three diffusion-based style transfer methods (DiffuseIT [22], InST [53], DiffStyle [18]), which have a style image as input. We employ the publicly available implementations of all baselines, using their recommended configurations.

Comparison with Conventional Style Transfer. As shown in Tab. 1, our method largely surpasses the conventional style transfer methods in terms of ArtFID, which is known as coinciding the human preference. In addition, the proposed method records the lowest FID, which denotes that stylized images highly resemble the target styles. For content fidelity metrics, ours shows superior scores in both CFSD and LPIPS. We point out that ours achieves much lower CFSD compared to other methods, which is the metric to only consider the spatial correlation.

In addition, we also emphasize that the proposed method can arbitrarily adjust the degree of style transfer by changing the γ, and the proposed method significantly surpasses all the other methods in terms of FID (style), when we match the value of LPIPS (content) (Fig. 10).

Comparison with Diffusion-based Style Transfer. Our method demonstrates the best performance in terms of LPIPS, FID, and their combination (ArtFID) with a large

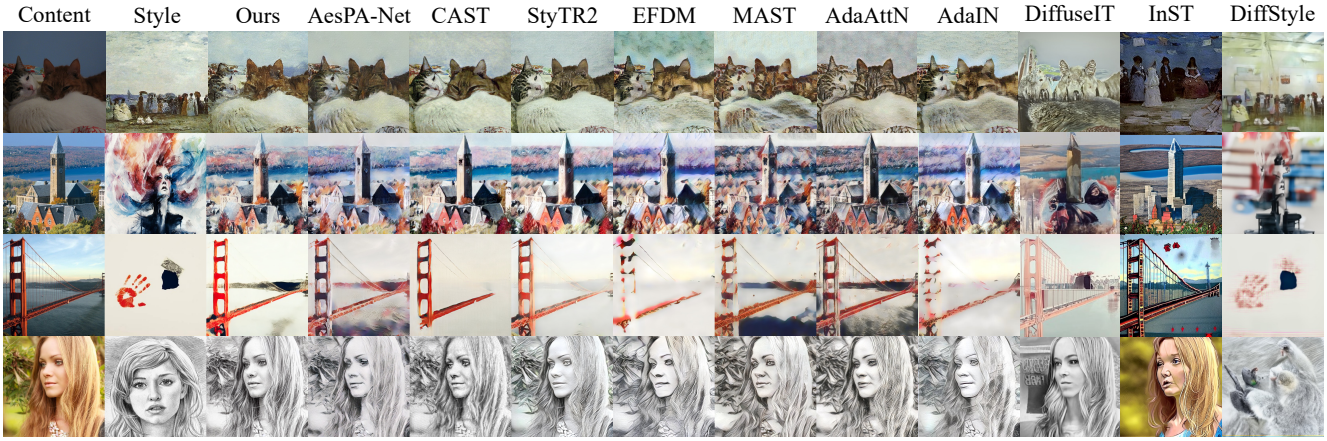


Figure 6. Qualitative comparison with conventional (4^{th}-10^{th} columns) and diffusion model baselines (11^{th}-13^{th} columns)

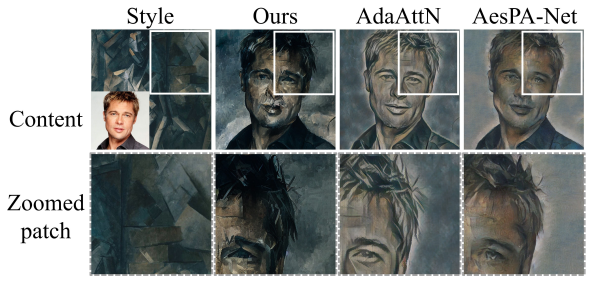


Figure 7. Qualitative comparison with best ArtFID (AdaAttN) and most recently proposed baselines (AesPA-Net) with additional zooming details

Metric	DiffuseIT	InST	DiffStyle	Ours
Time (sec)	792.8	816.9	355.9	**12.4**

Table 2. Comparison of inference time of diffusion-based methods for style transferring a given style and content pair

margin, as shown in Tab. 1. One significant factor for diffusion models is their running times, since they require several steps to synthesize a single image and it requires inevitable time cost. Hence, we measure the inference time for a pair of content and style images on a single TITAN RTX GPU, as shown in Tab. 2. Our method requires a total of 12.4 seconds, with 8.2 seconds for DDIM inversions and 4.2 seconds for sampling costs. As reported, we validate the proposed method significantly faster than other methods, even exploiting large-scale DM. This faster speed comes from the fact that the proposed methods can use the much smaller steps of DDIM inversion, because we additionally utilize the features collected during inversion steps, largely reducing the necessity for perfect inversion of content and style.

5.4. Qualitative Comparison

Comparison with Conventional Style Transfer. As shown in Fig. 6, we observe that our method tends to highly preserve the structural information of the content image, while also transferring the style well. For instance, as shown in the third row, ours retains the structure of the bridge, but the baselines struggle to preserve structure or transfer the style. We also provide the qualitative comparison with zooming details in Fig. 7 and Supplementary.

Comparison with Diffusion-based Style Transfer. We also compare our method with recent diffusion-based style transfer baselines [18, 23, 53]. As shown in Fig. 6, we observe the proposed technique transfers the style to the content well. On the other hand, baselines often lose the structure of content or fail to transfer the style, when an arbitrary content style pair is given. For instance, DiffuseIT and DiffStyle suffer from generating shape and visually plausible images or drop the original content. Differently, InST synthesizes the realistic images, while struggling to transfer style (1^{st} row) or change content of image (2^{nd}, 3^{rd} rows).

5.5. Ablation Study

To validate the effectiveness of the proposed components, we conduct ablation studies in both quantitative and qualitative ways. As shown in Fig. 8 and Tab. 3, style injection is significant for guiding the style and content of given images (Config. B). Besides, initial latent AdaIN has a large portion of transferring the color tone of style (Config. D). Attention temperature scaling is in charge of enhancement of quality in synthesized results such as sharpening details and resolving blurriness. For instance, this scaling jointly reduces the FID and LPIPS (Config. A* vs. C in Tab. 3). For more detailed analysis, we provide quantitative metrics with the style-content trade-off, while changing the attention scaling parameter τ in Fig. 10 (b). As reported, attention scaling effectively reduces both FID and LPIPS, proving its effects on the preservation of content and capability of style transfer ($\tau = 1.0$ vs. $\tau = 1.5$).

5.6. Additional Analysis

Content-Style Trade-Off. Our proposed method offers flexible control of the trade-off relation between content and style fidelity by adjusting the parameter γ, as discussed in

	Configuration	ArtFID	FID	LPIPS
A	Ours ($\gamma = 0.75$, default)	28.80	18.13	0.505
A*	Ours ($\gamma = 0.6$)	**27.97**	**17.21**	0.535
B	- Style Inject. (Sec. 4.1)	43.72	27.13	0.554
C	- Attn. Scaling (Sec. 4.2)	29.02	17.81	0.542
D	- Init. AdaIN (Sec. 4.3)	29.26	20.05	**0.390**

Table 3. Ablation study on proposed components

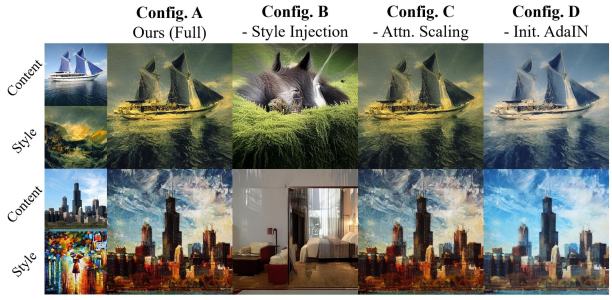


Figure 8. Qualitative comparison with ablation studies

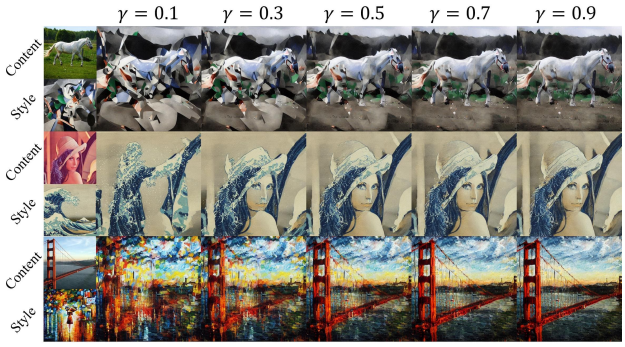
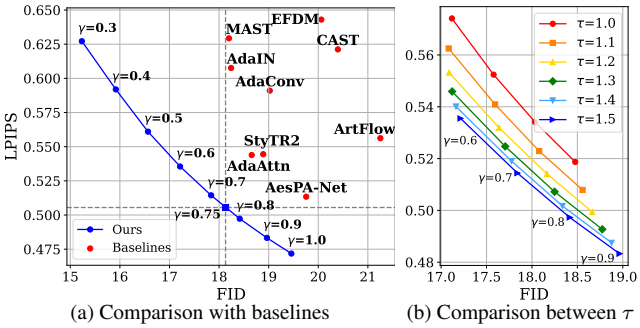


Figure 9. Visualization of effects of query preservation ratio γ

Sec. 4.1. In detail, we compute the FID and LPIPS while varying γ within the range $[0.3, 1]$ with a step size of 0.1. As shown in Fig. 10 (a), our method surpasses baseline methods across all ranges of content and style fidelity. This result implies the proposed method significantly outperforms the other methods, when we match style or content metric to the compared model by adjusting the γ of ours. Note that, dotted lines refer to our model reported in Tab. 1.

We also visualize the effects of the style-content trade-off by synthesizing images by adjusting γ. As shown in Fig. 9, the lower γ highly reflects the style while losing the content of the given image, and vice versa. This characteristic of the proposed method suggests that the users can adjust the degree of style by following their personal preferences.

Study on the value of τ. We observe that the gradual increase of τ enhances the performance of style transfer, although its effects on enhancement become smaller as τ goes larger, as shown in Fig. 10 (b). This result implies that the attention temperature scaling effectively works with a simple modification of the magnitude of the attention map.



(a) Comparison with baselines (b) Comparison between τ

Figure 10. Style-content trade-offs

Figure 11. Comparison with text-guided style transfer methods

Comparison with text-guided style transfer. We additionally compare the proposed method with the style transfer methods [21, 41] which are conditioned on the textual inputs. As text-guided methods tend to modify the style largely, we use $\gamma = 0.3$ for this experiment. Since the text condition hardly contains all the information in the style image such as texture and color tones, the transferred results less resemble the target style, as shown in Fig. 11. Differently, we validate that the proposed method successfully transfers the style with high fidelity.

6. Conclusion

Our work addresses the challenges associated with diffusion model-based style transfer methods, which often require time-consuming optimization steps or struggle to leverage the generative potential of large-scale diffusion models. To this end, we propose the method of adapting the pre-trained large-scale diffusion model on style transfer in a training-free way. Our method focuses on manipulating the features of self-attention layers, akin to the cross-attention mechanism, by substituting the *key* and *value* during the content generation process with those of the style. Furthermore, we propose the query preservation and attention temperature scaling to mitigate the issue of disruption of content, and initial latent AdaIN to handle the disharmonious color (failure to transfer the colors of style). Experimental results show the superiority of our proposed method over state-of-the-art techniques in previous baselines.

Acknowledgments

This work was supported in part by MSIT/IITP (No. 2022-0-00680, 2019-0-00421, 2020-0-01821, 2021-0-02068), and MSIT&KNPA/KIPoT (Police Lab 2.0, No. 210121M06).

References

[1] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 862–871, 2021. 6

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 1, 3

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, 2023. 3

[4] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023. 1

[5] Prashanth Chandran, Gaspard Zoss, Paulo Gotardo, Markus Gross, and Derek Bradley. Adaptive convolutions for structure-aware style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7972–7981, 2021. 6

[6] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR 2023 (Eleventh International Conference on Learning Representations)*, 2023. 1, 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[8] Yingying Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. Arbitrary style transfer via multi-adaptation network. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2719–2727, 2020. 6

[9] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11326–11336, 2022. 6

[10] Martin Nicolas Everaert, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, and Radhakrishna Achanta. Diffusion in style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2251–2261, 2023. 1, 3

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2, 5, 6

[12] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 6

[13] Nocholas Guttenberg. Diffusion with offset noise. https://www.crosslabs.org/blog/diffusion-with-offset-noise, 2023. 5

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2, 3, 4

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[16] Kibeom Hong, Seogkyu Jeon, Junsoo Lee, Namhyuk Ahn, Kunhee Kim, Pilhyeon Lee, Daesik Kim, Youngjung Uh, and Hyeran Byun. Aespa-net: Aesthetic pattern-aware style transfer networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22758–22767, 2023. 6

[17] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 5, 6

[18] Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. Training-free style transfer emerges from h-space in diffusion models. *arXiv preprint arXiv:2303.15403*, 2023. 1, 3, 6, 7

[19] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Conference on Computer Vision and Pattern Recognition 2023*, 2023. 1, 3

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 6

[21] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 8

[22] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022. 6

[23] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 7

[24] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 2

[25] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017.

[26] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image

stylization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 453–468, 2018. 2

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[28] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6649–6658, 2021. 6

[29] Ming Lu, Hao Zhao, Anbang Yao, Yurong Chen, Feng Xu, and Li Zhang. A closed-form solution to universal style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5952–5961, 2019. 2

[30] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

[31] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5880–5888, 2019. 2

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. 3

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3

[34] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 6

[37] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023. 3

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 4, 6

[40] Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1):394–409, 2019. 6

[41] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 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*, pages 1921–1930, 2023. 1, 2, 3, 8

[42] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022. 3

[43] Huan Wang, Yijun Li, Yuehai Wang, Haoji Hu, and Ming-Hsuan Yang. Collaborative distillation for ultra-resolution universal style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1860–1869, 2020. 2

[44] Zhizhong Wang, Lei Zhao, Haibo Chen, Lihong Qiu, Qihang Mo, Sihuan Lin, Wei Xing, and Dongming Lu. Diversified arbitrary style transfer via deep feature perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7789–7798, 2020. 2

[45] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7677–7689, 2023. 1, 3

[46] Matthias Wright and Björn Ommer. Artfid: Quantitative evaluation of neural style transfer. In *DAGM German Conference on Pattern Recognition*, pages 560–576. Springer, 2022. 6

[47] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. Zero-shot contrastive loss for text-guided diffusion image style transfer. *arXiv preprint arXiv:2303.08622*, 2023. 1, 3

[48] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 1

[49] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *arXiv preprint arXiv:2305.15347*, 2023. 2

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[51] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8035–8045, 2022. 6

[52] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–8, 2022. 2, 6

[53] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10156, 2023. 1, 3, 6, 7