

Style-A-Video: Agile Diffusion for Arbitrary Text-based Video Style Transfer

Nisha Huang^{1,2} Yuxin Zhang^{1,2} Weiming Dong^{1,2}

¹School of Artificial Intelligence, UCAS ²MAIS, Institute of Automation, CAS

<https://github.com/haha-lisa/Style-A-Video>



Figure 1: We propose a text-driven video stylization method Style-A-Video based on diffusion models.

ABSTRACT

Large-scale text-to-video diffusion models have demonstrated an exceptional ability to synthesize diverse videos. However, due to the lack of extensive text-to-video datasets and the necessary computational resources for training, directly applying these models for video stylization remains difficult. Also, given that the noise addition process on the input content is random and destructive, fulfilling the style transfer task's content preservation criteria is challenging. This paper proposes a zero-shot video stylization method named Style-A-Video, which utilizes a generative pre-trained transformer with an image latent diffusion model to achieve a concise text-controlled video stylization. We improve the guidance condition in the denoising process, establishing a balance between artistic expression and structure preservation. Furthermore, to decrease inter-frame flicker and avoid the formation of additional artifacts, we employ a sampling optimization and a temporal consistency module. Extensive experiments show that we can attain superior content preservation and stylistic performance while incurring less consumption than previous solutions.

1 INTRODUCTION

Breakthroughs in the field of text-driven image generation have been made with the public availability of hundreds of millions of large-scale multimodal datasets [37, 44]. Text-conditional generation efforts, such as DALL-E 2 [38], Imagen [43], and Stable Diffusion [40], have both powerful image generation capabilities and enhanced user-friendliness, allowing even novices to generate high-quality images. It opens up a multitude of possibilities for editing real-world visual videos from current generative works. Yet, introducing diffusion models for editing real-world videos remains hard and demanding.

Because of the challenges in collecting large-scale data matching to text and video, advancement in the video domain is limited in comparison to that in the image domain. Training text-guided video generation paradigms [15, 45] from scratch is a time-consuming and resource-intensive task, and tough to acquire and generalize. Therefore, leveraging current text-image models to generate videos is more practical. Meanwhile, studies [2, 9, 15, 18, 45, 51, 56] have been implemented on the basis of text-image models for video aspects with promising achievements. Text2LIVE [2] propagates editors through the computing explicit correspondences according

to pre-trained Neural Layered Atlases(NLA) [25] models. Tune-A-Video [51] fine-tunes each input video on top of the existing image generation model. Despite these efforts, which have contributed to a reduction in resource waste, there is still a substantial computing and timing cost associated with training NLA and fine-tuning input. In contrast, we would like to achieve quick inference for arbitrary videos to circumvent the expensive computational consumption.

Furthermore, due to the destructive nature of the noise addition process to the content in the diffusion model and the random character of the denoising process, the input video content is frequently tough to retain. Moreover, the text prompt in the text-video model only tells the model what the user intends to change, but does not convey what the user prefers to retain. Thus, the text-guided video transfer effort leads to content modifications in the video [9, 51], which hardly meets the basic requirements of the stylization task [5, 6, 13].

To accomplish stylistic representation of text prompt, preservation of input video content, inter-frame consistency, and fast optimization of the inference process, we concentrate on the following essential components. First, we propose a new combination of control conditions including text, video frames, and attention maps; specifically, text for style guidance, video frames for content guidance, and attention maps for detail guidance. And we adapt the noise prediction process by a custom guidance method, which is combined with classifier-free guidance. The global content of the input video is well maintained while realizing the text-guided stylization. We achieve fast convergence during inference without fine-tuning or additional training, enabling arbitrary stylization tasks for text-video input. Finally, temporal consistency is also introduced to eliminate flicker and enhance temporal coherence.

In summary, we present the following contributions:

- In this work, we propose Style-A-Video, a novel framework for arbitrary text-driven video styling based on a diffusion model. This work is performed entirely in inference time without additional per-video training or fine-tuning.
- Novel noise prediction guiding formulas are proposed to achieve simultaneous control of style, content, and structure. Besides, we achieve the control of time and content consistency in the inference process.
- Various experiments and user studies demonstrate the higher visual quality and effectiveness of our method among corresponding baselines.

2 RELATED WORK

Image and video style transfer. Image style transfer has been widely studied in recent years, which enables the generation of artistic paintings without the expertise of a professional painter. Gatys *et al.* [13] find that the inner products of the feature maps in CNNs can be used to represent styles and propose a neural style transfer (NST) method through successive optimization iterations. However, the optimization process is time-consuming and difficult to be widely used.

A number of techniques [7, 24, 31, 55] align the second-order statistics of the style and content images to transfer styles arbitrarily. Adopting adaptive instance normalization (AdaIN), which normalizes content features using the mean and variance of style features,

Nisha Huang^{1,2} Yuxin Zhang^{1,2} Weiming Dong^{1,2}

¹School of Artificial Intelligence, UCAS ²MAIS, Institute of Automation, CAS
<https://github.com/haha-lisa/Style-A-Video>

Huang *et al.* [24] present this arbitrary style transfer approach. To produce domain-specific sequences for content and style, respectively, Deng *et al.* [7] introduce StyTr2, which has two separate transformer encoders. By comparing and contrasting various styles and taking into account the style distribution, Zhang *et al.* [55] present contrastive arbitrary style transfer (CAST), which enables users to learn style representations directly from image attributes.

The majority of image-style transfer methods [5, 12, 42] are employed for video-style transfer. The accuracy of the optical flow computation had a significant impact on the sequence-based approaches' effectiveness. Ghosting artifacts will result from using incorrect optical flows. Moreover, processing high-resolution or lengthy videos is challenging due to the computationally expensive nature of predicting optical fluxes. We intend to suggest a productive zero-shot video style transfer method that succeeds in both temporal consistency and stylistic expression well.

Video style transfer takes a reference style image and statistically applies its style to an input video [5, 6, 12, 27, 42]. In comparison, our method applies a mix of style and content from an input text prompt or image while being constrained by the extracted structure data. By learning a generative model from data, our approach produces semantically consistent outputs instead of matching feature statistics.

Text-to-Image synthesis and editing. There has been remarkable progress since the use of conditional GANs in both text-guided image generation [30, 39]. ManiGAN [30] contains a text-conditioned GAN for editing an object's appearance while preserving the image content. However, such multi-modal GAN-based methods are restricted to specific image domains and limited in the expressiveness of the text. DALL-E [39] addresses this by learning a joint image-text distribution over a massive dataset. DALL-E produces text-to-images with remarkable accuracy, although it is not intended to edit already-existing images.

On the other hand, Denoising Diffusion Probabilistic Models (DDPMs) [16] are successfully leveraged for text-to-image generation. GLIDE [33] takes this approach further, supporting both text-to-image generation and inpainting. DALL-E 2 [38] leverages the CLIP [37] latent space and a prior model. To increase efficiency, Stable Diffusion [40] generates text from images in latent space rather than pixel space.

A recent influx of approaches makes use of a pre-trained generator [11, 35] and a pre-trained CLIP [37] to provide textual guidance during the generation process. StyleCLIP [35], StyleGAN-NADA [11], CLIPstyler [28], and LDAST [10] modify the content images based on CLIP optimizations. Since StyleCLIP and StyleGAN-NADA are limited by pre-trained generators and can only operate on specific data domains. CLIPstyler and LDAST can perform arbitrary content images, however, they still fall short in stylized representation. With the rapid development of generative work, diffusion methods have taken stylization a step further. Multiple works based on diffusions such as MGAD [22], DiffStyler [23], and InST [54] present impressive results and broaden application scenarios.

Text-to-video synthesis and editing. Although text-to-image generation has made notable strides, text-to-video generation is still an emerging area of study, primarily because there aren't ample pairs of high-quality text and video. By mapping text tokens to video

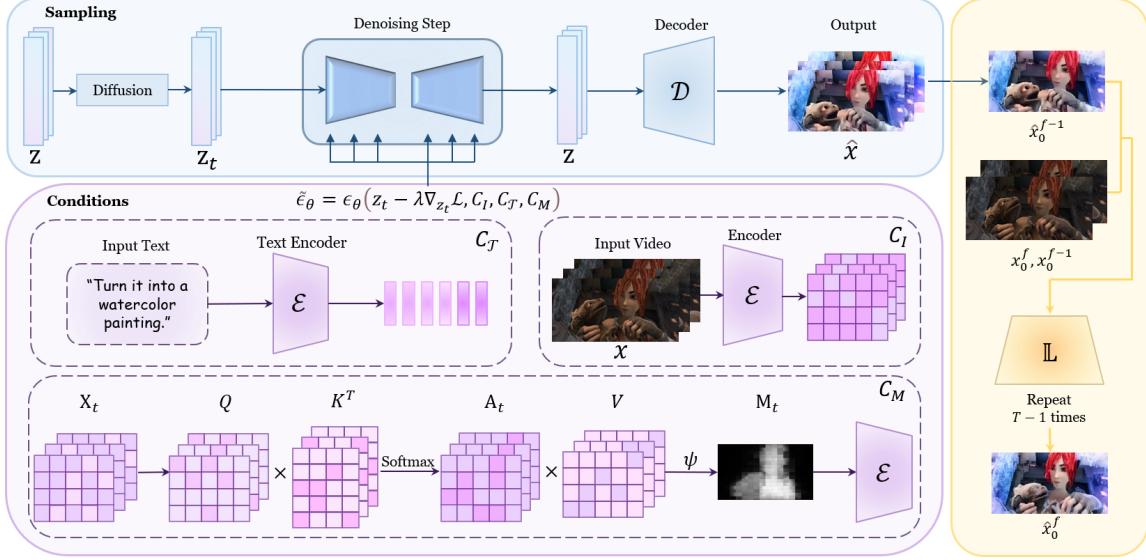


Figure 2: The overall framework of our method for text-driven video style transfer. The purple, blue, and yellow parts are the conditional representation, the sampling module, and the temporal consistency module, respectively. It generates the stylized video by maintaining the input video content while transferring the style through the textual description of the content.

tokens, GODIVA [49] is the first project to extend VQ-VAE [48] to text-to-video creation. Nüwa [50] suggests an integrated auto-regressive framework to handle tasks for generating text into both images and videos. CogVideo [20] extends CogView-2 [8] to text-to-video generation utilizing pre-trained text-to-image models and temporal attention modules to increase the quality of the video generation. A factorized space-time U-Net is suggested by Video Diffusion Models (VDM) [18] to carry out the diffusion process directly on pixels. For the purpose of producing high-resolution videos, Imagen Video [15] has more recently enhanced VDM using cascaded diffusion models and v-prediction parameterization. Similar in intent, Make-A-Video [45] intends to use considerable advancements in text-to-image generation to text-to-video generation. They blend the world movements from unsupervised video footage with the appearance-text information from text-image data.

The rapid advancement of text-guided video editing [2, 9, 34, 51] followed afterward. By breaking down a video into neural layers, Text2LIVE [2] enables the alteration of input movies using text instructions. A layered video representation offers constant propagation across frames once it is accessible. By fine-tuning a diffusion model on a single video, SinFusion [34] is able to produce variations and extrapolations of videos. Similar to this, Tune-A-Video [51] optimizes a video generated from an image model on a single video to allow for modification. However, the viability of these approaches in creative tools is constrained by pricey per-video training. Recently, Esser et al. introduce Gen-1 [9], which can edit any video with the use of text, however, the edited version loses the input video's content information to some extent.

3 METHODS

The purpose of our proposed approach is to stylize the input video for editing according to the text condition \mathcal{T} while preserving the

content and structure of the input video. Content is specifically defined as the appearance and semantic information of the input video. Structure refers to the shape and geometric features of the input video. To achieve that, we add the input frame information (denoted by I) with the self-attention information (denoted by M) to the generative model conditions. We infer the self-attention map M between encoded text \mathcal{T} and intermediate features of the denoiser ϵ_θ . The specific procedure is shown in Fig. 2. First, the implementation of our generative backbone is formulated as a conditional potential video diffusion model. Next, the guidance conditions are computed and encoded. Then, the guidance conditions and the classifier-free guidance are combined for noise prediction. And step-wise optimization is performed during the sampling process. Finally, the temporal consistency between frames is optimized.

3.1 Diffusion Models

Diffusion models [46] learn to reverse a fixed forward diffusion process. Normally-distributed noise is slowly added to each sample x_{t-1} to obtain x_t . The forward process models a fixed Markov chain and the noise is dependent on a variance schedule β_t where $t \in \{1, \dots, T\}$, with T being the total number of steps in our diffusion chain, and $x_0 := x$.

Denoising Diffusion Probabilistic Models (DDPMs) [16] are latent generative models trained to recreate a fixed forward Markov chain x_1, \dots, x_T . Given the data distribution $x_0 \sim q(x_0)$, the Markov transition $q(x_t | x_{t-1})$ is defined as a Gaussian distribution with a variance schedule $\beta_t \in (0, 1)$, that is,

$$q(x_t | x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbb{I}\right), \quad (1)$$

where $t = 1, \dots, T$. By the Bayes' rules and Markov property, one can explicitly express the conditional probabilities $q(x_t | x_0)$ and

$q(x_{t-1} | x_t, x_0)$ as

$$q(x_t | x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbb{I}\right), \quad (2)$$

$$q(x_{t-1} | x_t, x_0) = \mathcal{N}\left(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t\mathbb{I}\right), \quad (3)$$

$$\begin{aligned} \text{w.r.t. } t &= 1, \dots, T, \alpha_t = 1 - \beta_t, \\ \bar{\alpha}_t &= \prod_{s=1}^t \alpha_s, \tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t, \\ \tilde{\mu}_t(x_t, x_0) &= \frac{\sqrt{\bar{\alpha}_t} \beta_t}{1 - \bar{\alpha}_t} x_0 + \frac{\sqrt{\bar{\alpha}_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t. \end{aligned} \quad (4)$$

To generate the Markov chain x_1, \dots, x_T , DDPMs leverage the reverse process with a prior distribution $p(x_T) = \mathcal{N}(x_T; 0, \mathbb{I})$ and Gaussian transitions

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad (5)$$

where $t \in \{T, \dots, 1\}$. Learnable parameters θ are trained to guarantee that the generated reverse process is close to the forward process.

To this end, DDPMs follow the variational inference principle by maximizing the variational lower bound of the negative log-likelihood, which has a closed form given the KL divergence among Gaussian distributions. Empirically, these models can be interpreted as a sequence of weight-sharing denoising autoencoders $\epsilon_\theta(x_t, t)$, which are trained to predict a denoised variant of their input x_t . The objective can be simplified as

$$\mathbb{E}_{x, \epsilon \sim \mathcal{N}(0, 1), t} [\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2]. \quad (6)$$

Algorithm 1 Conditions Guidance Diffusion Sampling.

```

Input: The text prompt  $\mathcal{T}$ , video frame  $I$ ,  

frame number  $F$ , Diffusion model Model( $z_t$ ).  

Output: Stylized frames  $x_0^1, \dots, x_0^F$ .  

 $z_T \sim \mathcal{N}(0, \mathbb{I})$  a unit Gaussian random variable with specific seed  

S  

for  $f$  in  $1, \dots, F$  do  

 $x_t^f \leftarrow \text{noising}(x_0^f)$   

 $z_t = \mathcal{E}(x_t^f)$   

for  $t$  in  $T, \dots, 1$  do  

 $\epsilon, \Sigma, M \leftarrow \text{Model}(z_t)$   

 $\Delta z_t = \nabla_{z_t} \mathcal{L}_s$   

 $C_I, C_{\mathcal{T}}, C_M \leftarrow \text{CLIP embedding}(I, \mathcal{T}, M)$   

 $\tilde{\epsilon}_\theta = \epsilon_\theta(z_t - \lambda \Delta z_t, C_I, C_{\mathcal{T}}, C_M)$   

 $z_{t-1} \sim \mathcal{N}\left(\frac{1}{\sqrt{\bar{\alpha}_t}} \left(z_t - \frac{1 - \bar{\alpha}_t}{\sqrt{1 - \bar{\alpha}_t}} \tilde{\epsilon}\right), \Sigma\right)$   

end for  

return  $z_0$   

 $\hat{x}_0^f = \mathcal{D}(z_0)$   

if  $f = 1$   

 $\hat{x}_0^{f-1} = \emptyset$   

else  

 $\hat{x}_0^f \leftarrow x_0^f, \hat{x}_0^{f-1}$   

end for  

return  $x_0^f$ 

```

Latent Diffusion Models (LDMs) [40] are newly introduced variants of DDPMs that operate in the latent space of an autoencoder. LDMs improve the efficiency and quality of diffusion models

by operating in the latent space of a pre-trained variational autoencoder [26]. LDMs consist of two key components. First, the autoencoder is trained with patch-wise losses on a large collection of images. For an image x , the diffusion process adds noise to the encoded latent $z = \mathcal{E}(x)$ producing a noisy latent z_t where the noise level increases over timesteps $t \in T$. And a decoder \mathcal{D} learns to reconstruct the latent back to pixel space, such that $\mathcal{D}(\mathcal{E}(x)) \approx x$. The second component is a DDPM that is trained to remove the noise added to a latent representation of an image. This diffusion model can be conditioned on encoded embeddings of class labels. We learn a network θ that predicts the noise added to the noisy latent z_t given image condition c_I , text prompt condition $c_{\mathcal{T}}$, and attention map condition c_M . We minimize the following latent diffusion objective:

$$L = \mathbb{E}_{\mathcal{E}(x), c_I, c_{\mathcal{T}}, c_M, \epsilon \sim \mathcal{N}(0, 1), t} [\|\epsilon - \epsilon_\theta(z_t, t, c_I, c_{\mathcal{T}}, c_M)\|_2^2]. \quad (7)$$

3.2 Condition Representation

Style condition. Conditional Diffusion Models are well-suited to modeling conditional distributions such as $p(x | c)$. In this case, the forward process q remains unchanged while the conditioning variables c become additional inputs to the model. Our goal is to edit an input video based on a text prompt describing the desired edited video. Therefore, during sampling, we replace the category condition with the textual prompt description. The target style can be obtained from the textual prompts $c_{\mathcal{T}}$ and can be represented by the following equation:

$$z \sim p_\theta(z | c_{\mathcal{T}}), \quad x = \mathcal{D}(z). \quad (8)$$

Content condition. For the stylization task, besides style representation, content retention is another key issue. Due to the destructive nature of the noise addition process of the diffusion model on the content map itself, as well as the random nature of the denoising process, it is difficult to achieve better results on content retention. Previous works [1, 9], which used CLIP [37] image embedding to represent the content conditions, resulted in difficulties in achieving the traditional stylization requirements. In contrast, we add additional input channels to the first convolutional layer to concatenate z_t and c_I , so that the final generated results have more consistency in semantic content relative to the input video. The pre-trained checkpoints are used to initialize all of the diffusion model's available weights, and the weights that act on the newly added input channels are set to zero.

Self-features condition. Recent large-scale diffusion models [14, 22, 33] incorporate conditioning by augmenting the denoising U-Net ϵ_θ with the attention layer [14, 19]. The self-attention mechanism is a variant of the attention mechanism, which relies less on external information and is better at capturing the internal relevance of self-features. Specifically, given any feature map $X_t \in \mathbb{R}^{(HW) \times d}$ at a timestep t , for the height H and width W , the N -head self-attention is defined as:

$$Q_t^{(h)} = X_t W_Q^{(h)}, \quad K_t^{(h)} = X_t W_K^{(h)}, \quad (9)$$

where $W_Q^{(h)}, W_K^{(h)} \in \mathbb{R}^{C \times d}$ for $h = 0, 1, \dots, N - 1$.

$$A_t^{(h)} = \text{softmax}\left(Q_t^{(h)} \left(K_t^{(h)}\right)^T / \sqrt{d}\right). \quad (10)$$

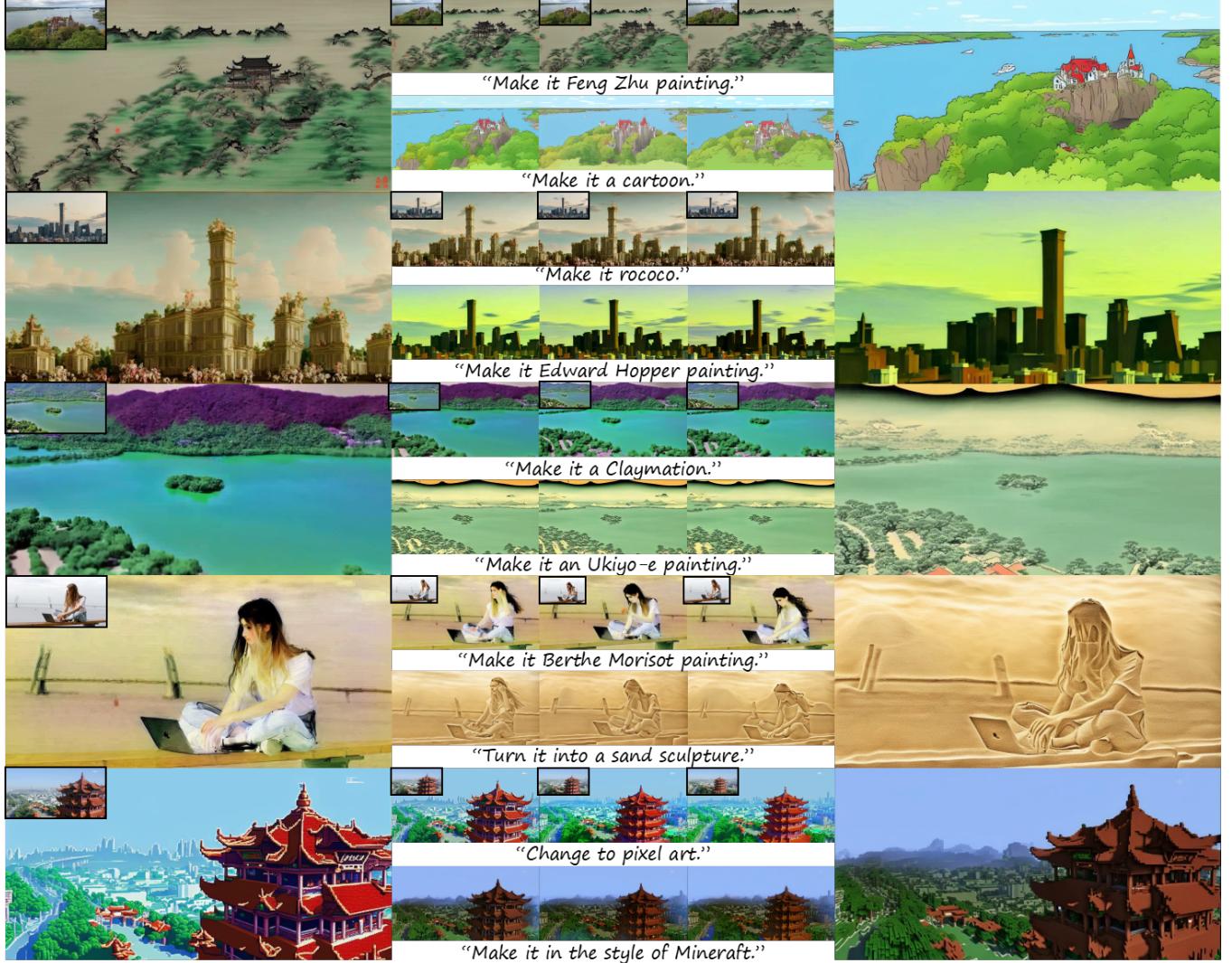


Figure 3: More video style transfer results by Style-A-Video. Our approach supports simple arbitrary style descriptions, without the necessity to describe the visual content in detail as [36, 51] require. The input contents for each result are shown as inset.

Each $A_t^{(h)}$ is then right multiplied by $V_t^{(h)} = X_t W_V^{(h)}$ where $W_V^{(h)} \in \mathbb{R}^{C \times d}$. Given a masking threshold ψ which is practically set to the mean value of A_t , the masked patches of x_t according to the self-attention map, and is formulated as follows:

$$M_t = \mathbb{1}(A_t > \psi) \quad (11)$$

3.3 Condition Guidance

Classifier-free guidance [17] is a method for trading off the quality and diversity of samples generated by a diffusion model. It is commonly used in class-conditional and text-conditional image generation to improve the visual quality of generated images and to make sampled images better correspond with their conditioning. Classifier-free guidance effectively shifts probability mass toward

data where an implicit classifier $p_\theta(z_t | c)$ assigns a high likelihood to the conditioning c . The implementation of classifier-free guidance involves jointly training the diffusion model for conditional and unconditional denoising, and combining the two score estimates at inference time. Training for unconditional denoising is done by simply setting the conditioning to a fixed null value $c = \emptyset$ at some frequency during training. At inference time, with a guidance scale $s \geq 1$, the modified score estimate $\tilde{\epsilon}_\theta(z_t, c)$ is extrapolated in the direction toward the conditional $\epsilon_\theta(z_t, c)$ and away from the unconditional $\epsilon_\theta(z_t, \emptyset)$:

$$\tilde{\epsilon}_\theta(z_t, c) = \epsilon_\theta(z_t, \emptyset) + s \cdot (\epsilon_\theta(z_t, c) - \epsilon_\theta(z_t, \emptyset)). \quad (12)$$

Guidance for conditions. For our task, the scoring network $\tilde{\epsilon}_\theta(z_t, c_I, c_T)$ has three conditions: the input image c_I , text prompt c_T , and self-attention map c_M . We find it beneficial to leverage

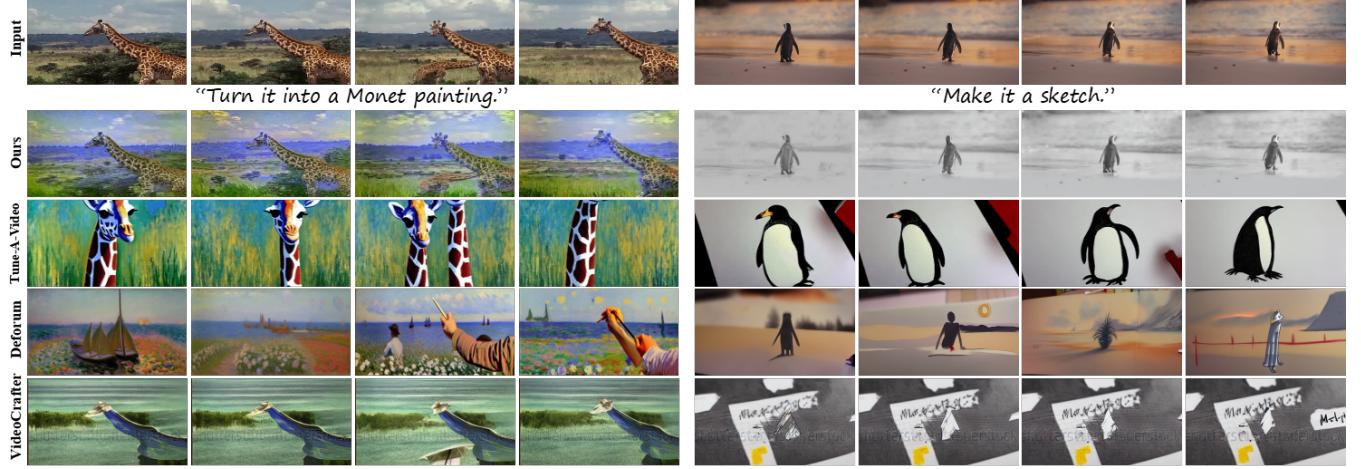


Figure 4: Qualitative comparison of Style-A-Video with other baselines. Our results have superior temporal consistency, content preservation, and style representation.

classifier-free guidance:

$$\epsilon'_\theta = \epsilon_\theta(z_t, \emptyset, \emptyset) \quad (13)$$

concerning each condition. Liu *et al.* [32] demonstrate that a conditional diffusion model can compose score estimates from multiple different conditioning values. We introduce three guidance scales, s_I , s_T , and s_M , which can be adjusted to trade off how strongly the generated samples correspond with the conditions. Our modified score estimate is as follows:

$$\begin{aligned} \tilde{\epsilon}_\theta(z_t, c_I, c_T, c_M) = & (1 - s_I - s_T - s_M) \cdot \epsilon'_\theta \\ & + s_I \cdot \epsilon_\theta(z_t, c_I, \emptyset) \\ & + s_T \cdot \epsilon_\theta(z_t, \emptyset, c_T) \\ & + s_M \cdot \epsilon_\theta(z_t, c_M, \emptyset). \end{aligned} \quad (14)$$

3.4 Sampling Optimization

Loss. We want to allow substantial texture and appearance changes while preserving the objects' original spatial layout, shape, and perceived semantics. While various perceptual content losses have been proposed in the context of style transfer, most of them use features extracted from a pre-trained VGG model. Instead, we define our loss in CLIP feature space. This allows us to impose additional constraints on the resulting internal CLIP representation of Io. Inspired by classical and recent works [47], we adopt the self-similarity measure. Specifically, we feed an image into CLIP's ViT encoder and extract its spatial tokens from the deepest layer. The structure loss is denoted by \mathcal{L}_s . The network is optimized by minimizing the similarity between the input frame x_0^f and the predicted frame x_t^f :

$$\mathcal{L}_s = 1 - \mathcal{D}_{\text{cos}}(x_0^f, x_t^f), \quad (15)$$

where $f \in \{1, \dots, F\}$, F being the frame number of the input video. We then take a gradient step according to the z gradient ∇_{z_t} :

$$\Delta z_t = \nabla_{z_t} \mathcal{L}_s, \quad (16)$$

Table 1: Quantitative evaluation and user study I for baselines. The best results are highlighted in bold while the second best results are marked with an underline.

	Ours	Tune-A-Deforum	Text2LIVE	VideoCrafter	Video
Fra-Con↑	0.987	0.882	0.908	0.969	<u>0.973</u>
Pro-Con↑	0.304	0.235	0.263	<u>0.272</u>	0.266
Fra-Acc↑	<u>0.983</u>	0.75	0.872	0.987	0.945
Preference↑	-	0.157	0.086	0.229	0.286

which optimizes the denoising network together with the guidance condition, as in the following equation:

$$\tilde{\epsilon}_\theta = \epsilon_\theta(z_t - \lambda \Delta z_t, C_I, C_T, C_M). \quad (17)$$

3.5 Temporal Consistency

The video frames $\{x_0^f\}_{f=1}^F$ are consistent with each other globally in both the short term and the long term. However, it might contain local flicker due to the misalignment between input and atlas-based frames. Hence, we use an extra local deflicker network to refine the results further. Prior work has shown that local flicker can be well addressed by a flow-based regularization. Hence, we choose a lightweight pipeline [27] with modification. As shown in Fig 2, we predict the output frame \hat{x}_0^f by providing two consecutive frames x_0^f, x_0^{f-1} and previous output \hat{x}_0^{f-1} to local refinement network \mathbb{L} . Two consecutive frames are firstly followed by a few convolution layers and then fused with the \hat{x}_0^{f-1} . The local flickering network is trained with a temporal consistency loss to remove local flickering artifacts [29].

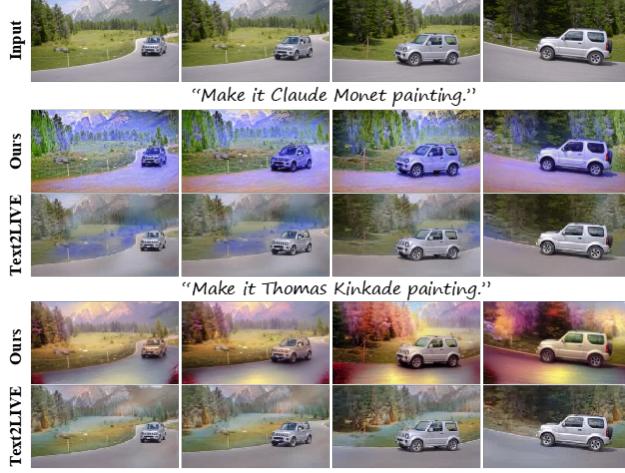


Figure 5: More qualitative comparison. We conducted it utilizing the pre-trained model provided by Text2LIVE [2].

4 EXPERIMENTS

4.1 Implementation Details

In practice, we implement our approach through latent diffusion models (LDMs) [40]. Our development is based on the capabilities of two large-scale pre-trained models operating [3] on different modalities - the large-scale language model GPT-3 [4] and the image generation model Stable Diffusion [40]. Style-A-Video takes 1 second and consumes 3925 MiB to generate a 512×256 frame on a single NVIDIA GeForce RTX 3090. The U-Net [41] architecture we utilize is based on Wide-ResNet [52]. To ensure the quality of the results and to maintain consistency of the parameters, the diffusion step and the time step used for the experiments in our work are set to 30.

4.2 Qualitative Evaluation

Results analysis. We test our approach on a variety of videos and stylized texts. These videos are sourced from the web and contain various object categories, such as people, animals, landscapes, etc. Figs. 1 and 3 show the results obtained by using different input videos and text prompts to guide. Our method can handle arbitrary shots, such as large wide-angle landscape shots with close-ups of people. In addition, we can generate stable results for both still shots and moving shots without additional subject tracking.

Comparison with baselines. We compare Style-A-Video with SOTA text-driven video editing methods, including Text2LIVE [2], Tune-A-Video [51], Defroum [21], and VideoCrafter [53]. Tune-A-Video requires additional training for each input video and text. The results in Fig. 4 show that Tune-A-Video has difficulty reproducing the content, shape, motion, and position of the input video; for example, the size and pose of the penguin and giraffe do not match the input video. And some content and artifacts are generated that do not match the input content, and from the results, backgrounds are generated that do not match the input content. Defroum generates high quality in the resulting images, however, there is no coherence between frames. In the giraffe example, the giraffe turns into a

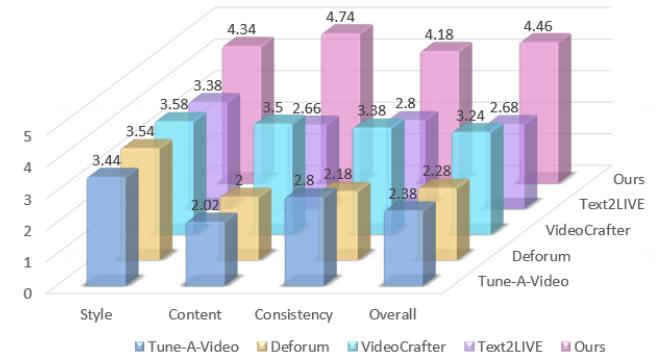


Figure 6: User study II. The average user rating results under the four metrics.

boat and the arm of a person who is drawing. VideoCrafter wreaks havoc on content in terms of presentation style. Unlike the above baseline, our method does a good job of maintaining the content and motion of the input video during editing, while performing stylistic expressions and achieving a high-quality video stylization. In addition, we compare our model with the publicly available ‘car’ based pre-trained model of Text2LIVE [2], which maintains the content of the input video rather well but struggles to meet the editing requirements imposed by the text. As demonstrated in Fig. 5, Text2LIVE is more limited in terms of style expressiveness than our results.

4.3 Quantitative Evaluation

We conduct the quantitative evaluation using the trained CLIP [37] model as previous methods [2, 21, 51, 53]. We randomly select 25 videos generated by Style-A-Video and each SOTA method. We quantify the trade-off between time consistency, cue consistency, and frame accuracy using the following three metrics, respectively.

Temporal consistency. We measure the temporal consistency of frames by first calculating the CLIP image embedding on all frames of the output video and then calculating the cosine similarity between all consecutive frame pairs measured the temporal consistency of frames.

Prompt consistency. We first calculate the CLIP image embedding on the output frames of the output video and the CLIP text embedding for the stylized text prompt. Besides, we calculate the average cosine similarity between the stylized frame embedding and the text embedding. This is thus used to measure the consistency of text and stylization.

Frame accuracy. We calculate the CLIP image embedding for the input video frames, with their stylized frames. The average cosine similarity between the two is calculated. This is used to measure the degree of content preservation for the input video.

As can be seen in Table 1, the Style-A-Video method achieves superior results in terms of temporal consistency and prompt consistency compared to the baselines and shows comparable content preservation accuracy to the TextLIVE [2] method.

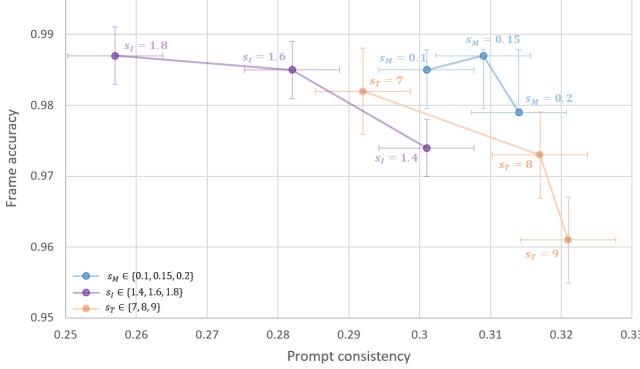


Figure 7: Ablation study. This figure reflects the effect of different condition guidance parameters scales on frame accuracy and prompt consistency. Better viewed for zoom-in.

4.4 User Study

User Study I. We compare our approach with several open-source SOTA text-guided video editing methods, including Text2LIVE [2], VideoCrafter [53], Deforum [21], and Tune-A-Video [51]. All baselines are trained and reasoned using publicly available implementations with default configurations. We randomly display 20 sets of content-style pairs, each pair containing the results of Style-A-Video with one of the other methods. We required 70 participants to select the more favored result and collected 1400 votes. The last row of Table 1 reports the percentage of votes each method received compared to Style-A-Video.

User Study II. We design a novel and detailed user study to better evaluate the performance of various aspects of the generated results. For each participant, 40 text-video pairs are randomly provided. We requested volunteers to score the following criteria: “how well the results agree with the style of the text description”, “how well the results agree with the content of the input video”, “temporal consistency of the results”, and “overall effects”. Finally, we collect 11200 scores from 70 participants. Fig. 6 shows the average scores of each method. It shows that our methods achieved good results on all indicators.

4.5 Ablation Study

Fig. 7 shows the effect of each guidance condition C_I , C_T , and C_M on the model we proposed. Larger guidance coefficients s_I , s_T , and s_M imply an increase in guidance weights. We observe a slight trade-off for increasing the intensity parameter in the baseline model. A larger s_I implies stronger agreement with the input video but reduces agreement with the style guide. Similarly, a larger s_T implies a stronger stylization effect and thus more deviation from the content of the input video.

To assess the impact of the self-attention mask condition and temporal consistency module on the Style-A-Video findings, we conduct an ablation study. As shown in Fig. 8, each design is ablated individually to analyze its impact. The second column investigates the effectiveness of the self-attention mask condition. Removal of the self-attention map condition leads to the loss of fine details (e.g., towns, temples, wings). The third column investigates the effect of

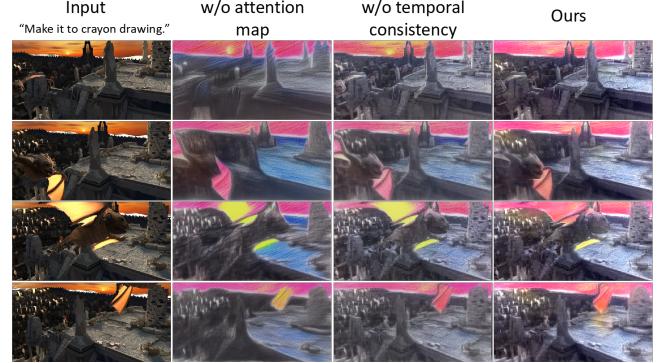


Figure 8: Ablation study. This figure presents the visualization of the impact of attention maps and temporal consistency on Style-A-Video.

the temporal consistency module. In the absence of this module, differences in color and illumination are produced between frames, resulting in the flickering of the video. These results show that all the above key designs contribute to the successful results of our approach.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we propose a new text-driven video stylization framework Style-A-Video. It is based on the latent diffusion model and is capable of stylizing videos based on given style and content information. We implement to provide temporal consistency guidance, style guidance, and content retention guidance in each denoising process. We further propose the cross-attention module as a guiding condition for structural information to improve the performance of our framework in terms of structural retention and overall quality. We try to use text, video, and cross-attention graphs as conditions simultaneously for video stylization applications. Our framework is capable of generating arbitrary video, and arbitrary style, in a faster inference process. We believe this will pave the way for the wide application of video synthesis and editing. In the future, we plan to study the effect of other conditions such as three parameters and pose estimation on video stability.

REFERENCES

- [1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. 2022. edifi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324* (2022).
- [2] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. 2022. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*. Springer, 707–723.
- [3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2022. Instructpix2pix: Learning to follow image editing instructions. *arXiv:2211.09800* (2022).
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [5] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. 2017. Coherent online video style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*. 1105–1114.
- [6] Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. 2021. Arbitrary video style transfer via multi-channel correlation. In *in Proc. 35th AAAI Conf. Artif. Intell. (AAAI)*, Vol. 35. 1210–1217.

- [7] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. 2022. Stytr2: Image style transfer with transformers. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 11326–11336.
- [8] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217* (2022).
- [9] Patrick Esser, Johnathan Chiu, Parmida Atighelchian, Jonathan Granskog, and Anastasis Germanidis. 2023. Structure and Content-Guided Video Synthesis with Diffusion Models. *arXiv:2302.03011* (2023).
- [10] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. 2022. Language-Driven Artistic Style Transfer. In *European Conference on Computer Vision (ECCV)*.
- [11] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)* (2022), 1–13.
- [12] Wei Gao, Yijun Li, Yihang Yin, and Ming-Hsuan Yang. 2020. Fast video multi-style transfer. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 3222–3230.
- [13] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *IEEE/CVF Conferences on Computer Vision and Pattern Recognition (CVPR)*. 2414–2423.
- [14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).
- [15] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)* (2020), 6840–6851.
- [17] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv:2207.12598* (2022).
- [18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *arXiv:2204.03458* (2022).
- [19] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. 2022. Improving Sample Quality of Diffusion Models Using Self-Attention Guidance. *arXiv preprint arXiv:2210.00939* (2022).
- [20] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868* (2022).
- [21] Eddy Hu. 2023. *Deforum Stable Diffusion V0.7*. <https://github.com/HelixNGC7293/DeforumStableDiffusionLocal>
- [22] Nisha Huang, Fan Tang, Weiming Dong, and Changsheng Xu. 2022. Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion. In *Proc. ACM Int. Conf. Multimedia. (MM)*. 1085–1094.
- [23] Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Yong Zhang, Weiming Dong, and Changsheng Xu. 2022. DiffStyler: Controllable Dual Diffusion for Text-Driven Image Stylization. *arXiv preprint arXiv:2211.10682* (2022).
- [24] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*. 1501–1510.
- [25] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. 2021. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–12.
- [26] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
- [27] Xiaoyu Kong, Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Yongyong Chen, Zhenyu He, and Changsheng Xu. 2023. Exploring the Temporal Consistency of Arbitrary Style Transfer: A Channelwise Perspective. *IEEE Trans. Neural Networks Learn. Syst.* (2023).
- [28] Gilhyun Kwon and Jong Chul Ye. 2022. CLIPstyler: Image Style Transfer with a Single Text Condition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [29] Chenyang Lei, Xuanchi Ren, Zhaoxiang Zhang, and Qifeng Chen. 2023. Blind Video Deflickering by Neural Filtering with a Flawed Atlas. *arXiv preprint arXiv:2303.08120* (2023).
- [30] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. 2020. Manigan: Text-guided image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7880–7889.
- [31] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal style transfer via feature transforms. *Advances in neural information processing systems* 30 (2017).
- [32] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022*.
- [33] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).
- [34] Yaniv Nikankin, Niv Haim, and Michal Irani. 2022. SinFusion: Training Diffusion Models on a Single Image or Video. *arXiv preprint arXiv:2211.11743* (2022).
- [35] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2085–2094.
- [36] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. 2023. FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. *arXiv:2303.09535* (2023).
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. ICML*. PMLR, 8748–8763.
- [38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125* (2022).
- [39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.
- [40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 10684–10695.
- [41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. Springer, 234–241.
- [42] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. 2016. Artistic style transfer for videos. In *Pattern Recognition: 38th German Conference, GCPR 2016, Hannover, Germany, September 12–15, 2016, Proceedings 38*. Springer, 26–36.
- [43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasedipour, Burcu Karagol Ayan, Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv:2205.11487* (2022).
- [44] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv:2210.08402* (2022).
- [45] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv:2209.14792* (2022).
- [46] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*. 2256–2265.
- [47] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2022. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10748–10757.
- [48] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).
- [49] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. 2021. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806* (2021).
- [50] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. 2022. Niwu: Visual synthesis pre-training for neural visual world creation. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*. Springer, 720–736.
- [51] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2022. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. *arXiv preprint arXiv:2212.11565* (2022).
- [52] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. *arXiv:1605.07146* (2016).
- [53] Yong Zhang, Yingqing He, Haoxin Chen, and Menghan Xia. 2023. *VideoCrafter: A Toolkit for Text-to-Video Generation and Editing*. <https://github.com/VideoCrafter/VideoCrafter>
- [54] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2022. Inversion-Based Creativity Transfer with Diffusion Models. *arXiv preprint arXiv:2211.13203* (2022).
- [55] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. 2022. Domain enhanced arbitrary image style transfer via contrastive learning. In *Proc. SIGGRAPH*. ACM, 12:1–12:8.
- [56] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018* (2022).