A Training-Free Localized Video Style Transfer Method Based on Diffusion Models

Quanjian Song[1], **Zhengjia Zhong**[2]
Wen Chen[2], **Junwei Lin**[2]
[1]Xiamen University AI Class [2]Xiamen University Information Class
23020241154435@stu.xmu.edu.cn, 31520241154540@stu.xmu.edu.cn
10420221153228@stu.xmu.edu.cn, 23020241154421@stu.xmu.edu.cn

Abstract

This paper presents UniVST, a unified framework for localized video style transfer based on diffusion model. It operates without the need for training, offering a distinct advantage over existing diffusion methods that transfer style across entire videos. The endeavors of this paper comprise: (1) A point-matching mask propagation strategy that leverages the feature maps from the DDIM inversion. This streamlines the model's architecture by obviating the need for tracking models. (2) A training-free AdaIN-guided video style transfer mechanism that operates at both the latent and attention levels. This balances content fidelity and style richness, mitigating the loss of localized details commonly associated with direct video stylization. (3) A sliding-window consistent smoothing scheme that harnesses optical flow within the pixel representation and refines predicted noise to update the latent space. This significantly enhances temporal consistency and diminishes artifacts in stylized video. Our proposed UniVST has been validated to be superior to existing methods in quantitative and qualitative metrics. It adeptly addresses the challenges of preserving the primary object's style while ensuring temporal consistency and detail preservation.

Introduction

Video editing has greatly improved thanks to the use of diffusion methods (Cong et al. 2023; Jeong and Ye 2023; Yang et al. 2023). T2V-Zero (Khachatryan et al. 2023) changes self-attention mechanisms to cross-frame attention. Tune-A-Video customizes video editing by adjusting attention weights (Wu et al. 2023a). Fate-Zero (Qi et al. 2023) keeps the video content intact via information from the inversion process. Animate-Zero (Yu et al. 2023) and VideoBooth (Jiang et al. 2024) add layers that stress time.

Video stylization, a sub-area of video editing, is gaining popularity. It is about adding artistic styles to video content. Our review identifies two main video stylization methods. The first method uses image or text to direct the style. StyleCrafter (Liu et al. 2023) personalizes video style with extra training adapter; VideoBooth (Jiang et al. 2024) uses image for custom video creation prompts; while Diffutoon (Duan et al. 2024) focuses on cartoon coloring from text descriptions. The second method includes all-encompassing video editing frameworks like AnyV2V (Ku et al. 2024) and BIVDiff (Shi et al. 2024) that can perform various tasks, including style transfer. These usually apply a consistent style across the whole video, good for thematic expression but sometimes lacking in precision for specific uses. For example, as shown in Fig. 1, in film, scenes might be shot normally and later stylized to match different environments. This enables aried scene changes from one video take, which is also useful in advertising for efficiently applying local style changes to existing videos. Therefore, selectively styling parts of a video while keeping others unchanged is an important academic and commercial task.

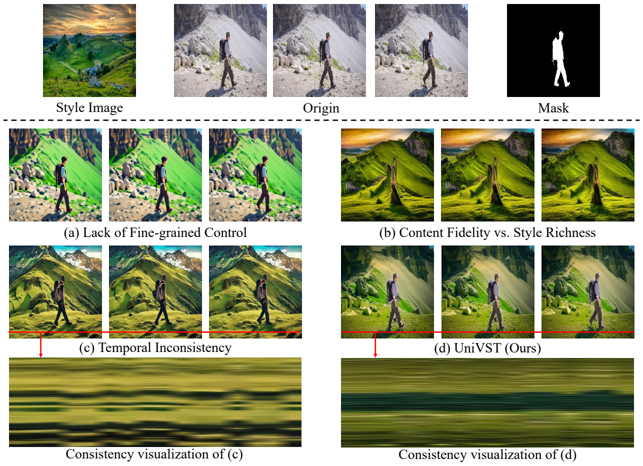
Style Image
Origin
Mask
(a) Lack of Fine-grained Control
(b) Content Fidelity vs. Style Richness
(c) Temporal Inconsistency
(d) UniVST (Ours)
Consistency visualization of (c)
Consistency visualization of (d)

Figure 1: Existing methods suffer from (a) lack of fine-grained control, (b) balance between content fidelity and style richness, and (c) temporal inconsistency. The last row shows the visualization of temporal consistency using MimicMotion's Y-T slices (Zhang et al. 2024a) for both the existing method (Shi et al. 2024) and our proposed UniVST.

We illustrate Fig. 1 and pinpoint three major limitations of current methods: *1. Lack of Fine-grained Control.* While some existing approaches (Qi et al. 2023; Wu et al. 2023a) leverage localized text descriptions to preserve the overall style transfer and maintain the primary object's style, the control is usually too coarse. This lack of precision can lead to the model failing to understand the nuances, potentially causing unintended style transfer effects on objects not explicitly mentioned in the text. *2. Content Fidelity vs. Style Richness*. In the realm of video stylization, there is a delicate balance to be struck between content fidelity and style richness. An overemphasis on style richness can result in a blurring and alter the original layout of the video, whereas an excessive focus on content fidelity might yield a stylized outcome that is indistinguishable from the original. *3. Temporal Inconsistency*. Unlike image stylization, video stylization requires careful consideration of temporal coherence between frames. As seen in (Shi et al. 2024; Ku et al. 2024), merely extending image stylization techniques to video can lead to inconsistencies across frames, manifesting as flickering and artifacts.

In this paper, we propose UniVST, solving the above issues in a unified framework for localized[1] video style transfer in a training-free manner. We inject the latent information from the content branch into the editing branch using mask, enabling the model to preserve the primary object's style. Unlike in the image domain, where masks can be readily obtained using additional models such as those in (Lugmayr et al. 2022; Mao et al. 2023), the video domain presents a more labor-intensive challenge due to the need for frame-by-frame mask generation. To address this, we introduce *a point-matching mask propagation strategy* leveraging feature maps from the DDIM inversion process (Song, Meng, and Ermon 2020) to capture correlations. Inspired by the image stylization model (Chung, Hyun, and Heo 2024), replacing the key and value in the editing branch with those from style branch can facilitate style transfer. However, we found that directly applying this method in video style transfer can lead to the loss and blurring of local details. Based on this foundation, we develop *a training-free AdaIN-guided style transfer mechanism*, which functions at both the latent and attention levels. It adeptly balances content fidelity with style richness. Detailed comparison of results is presented in Fig. 4.

Our survey shows that existing methods often utilize optical flow information to achieve temporal consistency. For instance, Ground-A-Video (Jeong and Ye 2023) uses optical flow to smooth initial noise, ensuring its temporal coherence, while Flatten (Cong et al. 2023) integrates optical flow into the attention process to maintain frame consistency. Building on these insights, we present *a sliding-window consistent smoothing scheme* that employs optical flow within the pixel representation and refines predicted noise to update the latent space. As illustrated in Fig. 1, our method reduce flickering and artifacts while enhances the temporal consistency of the edited video. Overall, our major contributions in this paper are as follows:

- We introduce, to the best of our knowledge, the first localized video style transfer framework, featuring a novel point-matching mask propagation strategy.
- We develop a training-free AdaIN-guided video style transfer mechanism that operates at both the latent and attention levels. It effectively harmonizes content fidelity and style richness throughout the transfer process.
- We present a sliding-window consistent smoothing scheme based on optical flow, which adeptly upholds temporal consistency during the video transfer process.
- Extensive experiments demonstrate that our framework outperforms several state-of-the-art methods in both qualitative and quantitative metrics.

Related Work

Image Style Transfer. RSCT (Ding et al. 2024) calls for the separate foreground and background style transfer, and combines them to ensure a consistent style. DiffStyle (Li 2024) employs LoRA (Hu et al. 2021) to refine image prompts for stylization, while ArtBank (Zhang et al. 2024c) enhances text embeddings through fine-tuning image prompts. StyleID (Chung, Hyun, and Heo 2024) leverages key-value replacement coupled with an initial latent AdaIN (Huang and Belongie 2017) to preserve content fidelity. Z* (Deng et al. 2023) rearranges attention to integrate content with style. InstanceStylePlus (Wang et al. 2024b) utilizes adapters to safeguard content retention and enhance stylistic expression.

Video Style Transfer. OCD (Kahatapitiya et al. 2024) integrates object-centric sampling and merging, expediting the editing process. StyleCrafter (Liu et al. 2023) refines adapters using image prompts based on the T2V model. Style-A-Video (Huang, Zhang, and Dong 2024) enhances the stylistic alignment through strategic style guidance. Diffutoon (Duan et al. 2024) pioneers primary and editing branches for multi-segment editing in video colorization, employing FastBlend (Duan et al. 2023) to uphold temporal continuity. BIVDiff (Shi et al. 2024) combines frame-by-frame image editing with temporal consistency modeling. AnyV2V (Ku et al. 2024) follows a similar initial editing strategy, subsequently leveraging an I2V model (Zhang et al. 2023) to propagate changes across the video sequence.

Preliminary

Latent Diffusion Model. LDM (Rombach et al. 2022) adds and removes noise in a low-dim space by using the encoder $\mathcal{E}$ and decoder $\mathcal{D}$. Given a Gaussian noise Z_T, DDIM denoising can be formulated as (Yang et al. 2023):

$$Z_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{Z_{t\rightarrow 0}}_{\text{predicted } Z_0} + \underbrace{\sqrt{1-\alpha_{t-1}}\epsilon_\theta\left(Z_t, t, C\right)}_{\text{direction pointing to } Z_{t-1}}, \tag{1}$$

where $Z_{t\rightarrow 0}$ is an estimation of Z_0 at time step t:

$$Z_{t\rightarrow 0} = \left(Z_t - \sqrt{1-\alpha_t}\epsilon_\theta\left(Z_t, t, C\right)\right) / \sqrt{\alpha_t}, \tag{2}$$

where α_t is a parameter and $C = \phi$ for editing integrity.

DDIM Inversion. DDIM (Song, Meng, and Ermon 2020) transforms noise Z_T into Z_0. Given reversible ODE (Gear and Petzold 1984), its reverse process can be described by:

$$Z_{t+1} = A_t Z_t + B_t \epsilon_\theta\left(Z_t, t, \phi\right), \tag{3}$$

where A_t and B_t are functions of time t. During editing, the initial noise Z_T that aligns with the original video distribution Z_0 can be obtained using this equation.

[1] Here, "localized" indicates a particular part of the video.

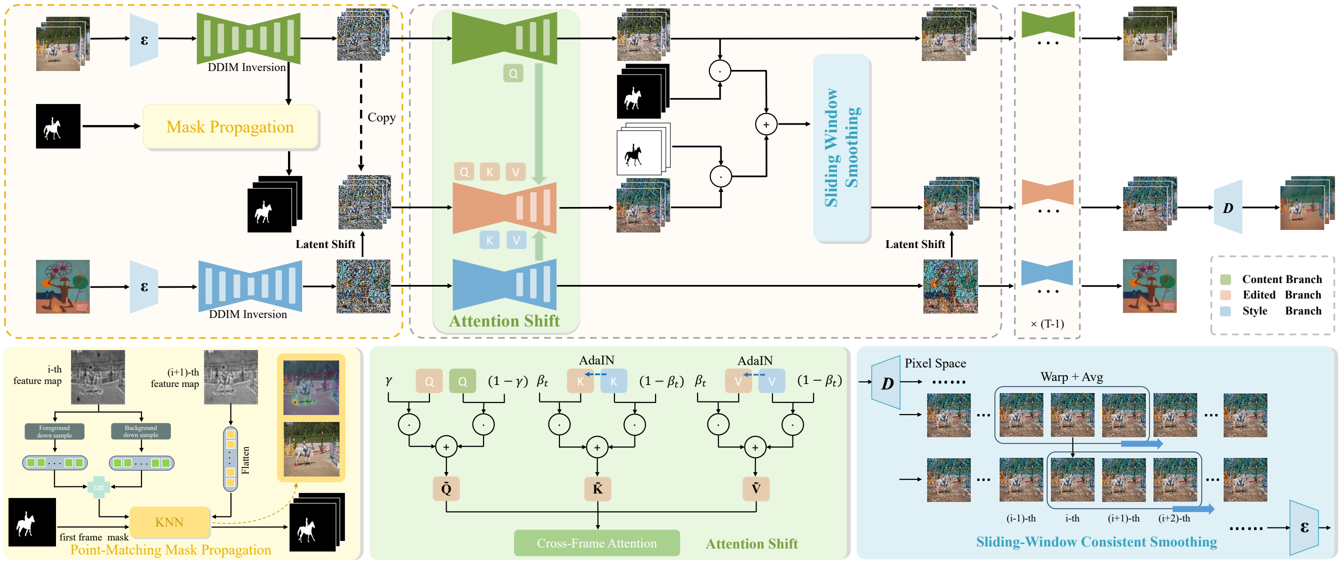

Figure 2: Overall framework. It is structured around three main components: (1) Point-Matching Mask Propagation, (2) AdaIN-Guided Video Style Transfer(Attention-shift and latent-shift) and (3) Sliding-Window Consistent Smoothing.

3D UNet Extension. For video tasks, LDM requires a 3D UNet extension. Following (Ho et al. 2022; Cong et al. 2023), the 3×3 convolutional kernel in the convolutional blocks is expanded to $1\times3\times3$. To achieve frame parallelization, the feature shape Z_t is transformed from $R^{b\times c\times f\times h\times w}$ to $R^{(bf)\times c\times h\times w}$ before self-attention and cross-attention operations. To facilitate frame interaction, we follow T2V-Zero (Khachatryan et al. 2023) to modify the self-attention to cross-frame attention :

$$\begin{aligned} Q^i = W^{query} Z^i, \quad K^i = W^{key}\left[Z^1, Z^{i-1}\right], \\ V^i = W^{value}\left[Z^1, Z^{i-1}\right], \end{aligned} \tag{4}$$

where Z^i denotes the feature of the i-th frame, and W^{query}, W^{key}, W^{value} represent the respective mapping matrices, with $[\cdot,\cdot]$ indicating concatenation.

Localized Video Style Transfer

Given an original video comprising N frames $\{I^i\}_{i=1}^N$ and a reference style image I^s, our objective is to transform the original video into a new sequence $\{\mathcal{I}^i\}_{i=1}^N$. To achieve this, we propose UniVST, which integrates the style of the reference image into the original video while preserving the style of the primary objects within the video unchanged. As shown in Fig. 2, it mainly includes *Point-Matching Mask Propagation*, *AdaIN-Guided Video Style Transfer*, and *Sliding-Window Consistent Smoothing* three components.

Point-Matching Mask Propagation

To preserve the primary style during the style transfer process, a potent technique is to apply a mask to blend the latent variables at several steps. This can be expressed as:

$$\mathcal{Z}_t = M \cdot Z_t + (1 - M) \cdot \mathcal{Z}_t, \tag{5}$$

where $\mathcal{Z}_t$ and Z_t denote the latent variables in the editing and content branch, respectively. However, in contrast to images, generating masks M for each video frame is laborious.

Algorithm 1: Point-Matching Mask Propagation

```
Input: $M_1$                                        Mask for the first frame.
Input: $F_{t_0}$                   Feature map from upsampling block-2.
Input: $r$                                             The sampling rates.
Input: $k$                                           The $k$-nearest points.
Input: $n$                               The number of reference frames.
Output: $\{M_i\}_{i=1}^N$                             Masks for all frames.
Initialize: previous_features ← [ ]   Initialize empty list for previous features.
Initialize: previous_masks ← [ ]       Initialize empty list for previous masks.
Initialize: first_feature ← [ ]          Initialize empty list for first feature.
for $i = 1, \ldots, N$ do
    $fore\_index \leftarrow$ where$(M_{i-1} = 1)$
    $back\_index \leftarrow$ where$(M_{i-1} = 0)$
    $fore\_index \leftarrow$ random_sample$(fore\_index, r \cdot \frac{|\text{fore}|}{|\text{fore}|+|\text{back}|})$
    $back\_index \leftarrow$ random_sample$(back\_index, r \cdot \frac{|\text{back}|}{|\text{fore}|+|\text{back}|})$
    $current\_index \leftarrow$ concat$(fore\_index, back\_index)$
    $current\_feature \leftarrow F^i_{t_0}[current\_index]$
    if $i = 1$ then
        first_feature.append($current\_feature$)                 Store first frame.
        $M_i \leftarrow M_1$
    else
        if |previous_features| $\geq n$ then
            previous_features.pop(0)          Remove the oldest frame feature.
            previous_masks.pop(0)                Remove the oldest frame mask.
        end if
        $f \leftarrow$ concat(previous_features, first_feature)
        $M_i \leftarrow$ KNN$(f, F^i_{t_0}$, concat(previous_masks, $M_1$), $k)$
        previous_features.append($current\_feature$)      Store current frame.
        previous_masks.append($M_i$)                       Store current mask.
    end if
end for
Return: $\{M_i\}_{i=1}^N$
```

Users typically prefer to control the video's main style by supplying a mask for only the initial frame. Extending the user-provided mask across all frames efficiently is neces-

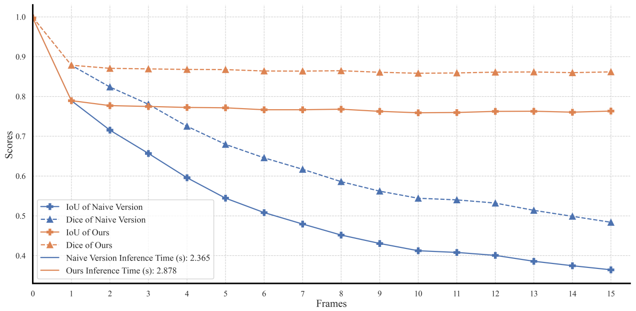

Figure 3: Comparison of accuracy and inference time under different mask propagation strategies. The introduction of anchor frames with the downsampling strategy significantly mitigates the accuracy degradation in subsequent frame propagation and also reduces inference time.

sary. Thus, we introduce point-matching mask propagation below and details are provided in Algorithm 1.

Point Matching. Inspired from DIFT (Tang et al. 2023), the three upsampling blocks in the UNet are replete with rich semantic information. Therefore, we can retain the feature map $\{F_{t_0}^i\}_{i=1}^N$ from upsampling block-2 at a given step t_0 during DDIM inversion in the source video, and use them to calculate the pixel correspondences between frames. Specifically, to pinpoint the pixel in the j-th frame that most closely resembles a given point p_i in the i-th frame, we calculate the cosine similarity between the corresponding point in the feature map $F_{t_0}^i$ and all points in the feature map $F_{t_0}^j$ as:

$$p_j = \underset{p_j \in F_{t_0}^j}{\arg\min}\, CosSim\big(p_i, p_j\big), \quad p_i \in F_{t_0}^i. \tag{6}$$

Utilizing this principle, we can propagate the mask from the first frame to all subsequent frames using the mask propagation strategy formally introduced in the following.

Mask Propagation. Given an initial object mask for the first frame from the user, we use k-NN (Cover and Hart 1967) to propagate this mask to subsequent frames. For each point in the i-th frame, we find the k most similar points in the $(i-1)$-th frame using point matching. If the majority of these k points are in the foreground, the corresponding point in the i-th frame is also classified as foreground; otherwise, it is background.

However, as shown in Fig. 3, using a naive autoregressive approach for mask propagation can lead to error accumulation, significantly decreasing accuracy for subsequent frames. To address this, we design the anchor frames mechanism that incorporates information from the first frame and the previous n frames.

Although introducing such a mechanism can improve the propagation accuracy of subsequent frames, it requires more similarity calculations, potentially reducing efficiency. To address this issue, we apply random downsampling at a rate of r to both foreground and background regions of the anchor frames to minimize costs. The sampling rates are then adjusted to reflect the proportions in the foreground and background. As illustrated in Fig. 3, the combination of anchor frames and random downsampling maintains final mask propagation accuracy while improving efficiency. More comprehensive information and in-depth analysis can be available in the experimental section.

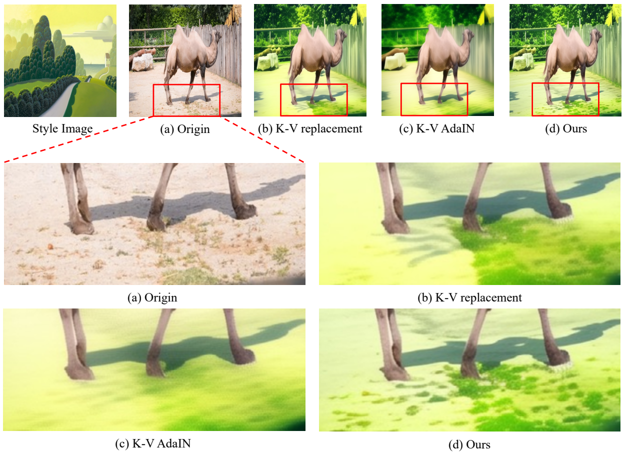

Figure 4: (a) Original video frame; (b) Key-value replacement (Chung, Hyun, and Heo 2024); (c) Key-value AdaIN; (d) Key-value AdaIN with gradual attention shift.

AdaIN-Guided Video Style Transfer

By using DDIM inversion (Zhang et al. 2024b) to the original video $\{I^i\}_{i=1}^N$ and the reference style image I^s, we obtain their respective noise latents Z_t and Z_t^s $(t = 1 \rightarrow T)$. Then, we establish the edited latents $\mathcal{Z}_t = Z_t$, and integrate a three-branch architecture with latent-shift and attention-shift mechanisms to accomplish video style transfer.

AdaIN-Guided Latent Shift. We have discovered that applying the AdaIN (Chung, Hyun, and Heo 2024) to the initial noise significantly enhances the transferred color style. Therefore, we extend the application of AdaIN to several steps during the denoising process. Within the interval $t \in [\tau_0, \tau_1]$, We consider applying it to the latent $\mathcal{Z}_t$ of the editing branch, shifting its mean and variance to the latent Z_t^s of the style branch:

$$\mathcal{Z}_t = \text{AdaIN}(\mathcal{Z}_t, Z_t^s), \tag{7}$$

$$\text{AdaIN}(\mathcal{Z}_t, Z_t^s) = \sigma\left(Z_t^s\right)\left(\frac{\mathcal{Z}_t - \mu\left(\mathcal{Z}_t\right)}{\sigma\left(\mathcal{Z}_t\right)}\right) + \mu\left(Z_t^s\right), \tag{8}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denoting mean and standard deviation.

AdaIN-Guided Attention Shift. To prevent the direct replacement of self-attention from causing content damage, (Chung, Hyun, and Heo 2024) blended the original video's query Q_t with the edited query $\mathcal{Q}_t$ as:

$$\begin{gathered}\tilde{\mathcal{Q}}_t = \gamma \cdot \mathcal{Q}_t + (1-\gamma) \cdot Q_t, \\ \tilde{\mathcal{K}}_t = K_t^s,\ \tilde{\mathcal{V}}_t = V_t^s,\end{gathered} \tag{9}$$

where γ controls the balance between content and style. However, as observed in Fig. 4(b), this approach still lead to a loss of localized details in video stylization. To mitigate the loss of details caused by direct replacement, we propose to shift the mean and variance of the distributions of $\mathcal{K}_t$ and $\mathcal{V}_t$ using AdaIN:

$$\tilde{\mathcal{K}}_t = \text{AdaIN}(\mathcal{K}_t, K_t^s),\ \tilde{\mathcal{V}}_t = \text{AdaIN}(\mathcal{V}_t, V_t^s). \tag{10}$$

Though this shift preserves localized details, as indicated in Fig. 4(c), it causes content blurring and artifacts. Upon analysis, we find that the early denoising stages contain less information, making key and value replacements inappropriate. Thus, within the interval $t \in [\tau_2, \tau_3]$, we gradually decrease the impact of the attention shift by combining the key and value replacements with the query blending:

$$\begin{aligned} \tilde{\mathcal{K}}_t &= \beta_t \cdot \text{AdaIN}(\mathcal{K}_t, K_t^s) + (1-\beta_t)\cdot K_t^s, \\ \tilde{\mathcal{V}}_t &= \beta_t \cdot \text{AdaIN}(\mathcal{V}_t, V_t^s) + (1-\beta_t)\cdot V_t^s, \\ \beta_t &= \frac{\beta_{\tau_3} - \beta_{\tau_2}}{\tau_3 - \tau_2} \cdot (t - \tau_2) + \beta_{\tau_2}. \end{aligned} \tag{11}$$

By combing key-value AdaIN with gradual attention shift in Eq. (11) and query blending in Eq. (9), the stylization results achieve an effective balance between content fidelity and style richness, as shown in Fig. 4(d) and Table 7.

Sliding-Window Consistent Smoothing

We address the common issues of flicker and artifacts in video stylization by focusing on temporal inconsistency. Following (Cong et al. 2023; Jeong and Ye 2023), we utilize optical flow with specialized sliding window smoothing.

Optical Flow Warping. Optical flow estimation is a pivotal technique, often used for object tracking (Kale, Pawar, and Dhulekar 2015; Shin et al. 2005). It predicts the motion direction and velocity of pixels by analyzing changes between consecutive frames. Using models like RAFT (Teed and Deng 2020), optical flow information can be predicted from one frame to the next, allowing for the reconstruction of the initial frame. In cases where occlusion occurs, bidirectional optical flow can estimate the mask and fill in occluded regions with the original image. This process, termed "Warp," is key to our following discussions.

Sliding Window Smoothing. We apply the warp operation during denoising process to achieve local smoothing. At time step t, we calculate the latent representation $\mathcal{Z}_{t\to 0}$ from the predicted noise ε_t according to Eq. (2). $\mathcal{Z}_{t\to 0}$ is then decoded by the decoder $\mathcal{D}$ into a pixel representation $\{P_t^i\}_{i=1}^N$. For each decoded frame P_t^i, we apply a sliding window of size $2m$ to warp its neighboring frames and smooth them to update the original pixel:

$$\bar{P}_t^i \leftarrow \frac{1}{2m+1} \cdot \sum_{j=i-m}^{i+m} \text{Warp}(P_t^i, P_t^j). \tag{12}$$

The updated values from the current window will be incorporated into the calculations for the next window, as shown in Fig. 2. Subsequently, the VAE's encoder $\mathcal{E}$ re-encodes the smoothed $\{\bar{P}_t^i\}_{i=1}^N$ into latent representations $\bar{\mathcal{Z}}_{t\to 0}$, which is an updated version of $\mathcal{Z}_{t\to 0}$. Next, the predicted noise ε_t is adjusted using the inversion of Eq. (2):

$$\bar{\varepsilon}_t \leftarrow \frac{\mathcal{Z}_t - \sqrt{\alpha_t}\bar{\mathcal{Z}}_{t\to 0}}{\sqrt{1-\alpha_t}}. \tag{13}$$

Finally, the latent representation is updated in accordance with the DDIM schedule in Eq. (1):

$$\mathcal{Z}_{t-1} \leftarrow \sqrt{\alpha_{t-1}}\bar{\mathcal{Z}}_{t\to 0} + \sqrt{1-\alpha_{t-1}}\bar{\varepsilon}_t. \tag{14}$$

It is important to note that we do not adopt a global smoothing strategy because extensive optical flow predictions would introduce significant computational overhead. Instead, by employing a sliding local window scheme, we achieved a balance between temporal consistency and efficiency. And this scheme is only performed within the interval $t \in [\tau_4, \tau_5]$.

Experimentation

Experimental Settings

Our method is based on the Stable Diffusion V1.5 model, as referenced in the literature (Rombach et al. 2022). The hyperparameters are configured as follows: $\tau_0 = 0.1T$, $\tau_1 = 0.2T$, $\tau_2 = 0.4T$, $\tau_3 = 1.0T$, $\tau_4 = 0.5T$, $\tau_5 = 0.6T$, $t_0 = 0.4T$, with the total time period T set to 50. Additionally, we have $\gamma = 0.35$, $\beta_{\tau_2} = 0.1$, $\beta_{\tau_3} = 0.9$, $r = 0.3$, $k = 15$, $m = 2$, and $n = 9$. Videos are resized to a resolution of 512 by 512 pixels and are processed in batches consisting of 16 frames each. For optical flow estimation, we employ the RAFT model (Teed and Deng 2020). Our approach does not require training or fine-tuning and can be executed on a single RTX 3090 GPU, making it both efficient and practical for deployment.

Datasets. Building upon the foundational work referenced in (Chung, Hyun, and Heo 2024; Huang, Zhang, and Dong 2024), we have meticulously selected style images from two renowned databases: WikiArt (Phillips and Mackintosh 2011) and Laion-Aesthetics-6.5+ (Schuhmann et al. 2022), each contributing a collection of 20 distinct images. For the content videos, we have procured a dataset comprising 50 videos from the DAVIS2016 (Perazzi et al. 2016) and TGVE (Wu et al. 2023b) datasets, which we have amalgamated and designated as DAVTG. Through the integration of these datasets, we have successfully established two comprehensive subsets: DAVTG-WikiArt, which contains 1,000 curated pairs, and DAVTG-Laion, which also features 1,000 paired entries.

Metrics. (1) **Overall Transfer Performance:** Drawing inspiration from the methodologies outlined in (Chung, Hyun, and Heo 2024; Huang, Zhang, and Dong 2024), we assess the overall transfer performance by employing ArtFID (Wright and Ommer 2022), a metric that amalgamates LPIPS (Zhang et al. 2018) for gauging content fidelity and FID (Heusel et al. 2017) for evaluating the richness of style. (2) **Foreground Style Preservation:** To evaluate the preservation of texture, we utilize the SSIM (Wang et al. 2004) metric, while the CLIP-I score (Radford et al. 2021) is employed to assess the preservation of semantic content. During this evaluation, the background is neutralized to ensure accuracy. (3) **Temporal Consistency:** In line with the approaches detailed in (Huang, Zhang, and Dong 2024; Liu et al. 2023), we measure the similarity between consecutive frames using the CLIP-F score (Radford et al. 2021), which serves as a metric for evaluating the temporal coherence of

Table 1: Quantitative comparison of existing methods across various metrics, superscript * indicates non-diffusion methods. The **bold** values represent the best performance and the underscore stresses the second best.

Datasets	Methods	Overall Transfer Performance			Style Preservation		Consistency	Efficiency
		ArtFID↓	FID↓	LPIPS↓	SSIM↑	CLIP-I↑	CLIP-F↑	Inference Time(s)↓
DAVTG-WikiArt	Diffutoon (Duan et al. 2024)	43.150	26.268	0.582	0.961	0.954	**0.980**	44.263
	StyleCrafter (Liu et al. 2023)	39.119	21.008	0.777	0.928	0.895	0.978	111.225
	AnyV2V (Ku et al. 2024)	**33.774**	**19.104**	0.680	0.938	0.914	0.978	78.474
	BIVDiff (Shi et al. 2024)	46.071	25.468	0.740	0.921	0.894	0.955	158.367
	EFDM* (Zhang et al. 2022a)	37.407	22.350	0.602	0.936	0.935	0.963	**5.969**
	CAST* (Zhang et al. 2022b)	45.198	26.174	0.663	0.948	0.936	0.975	7.348
	UniVST (Ours)	37.152	26.134	**0.369**	**0.986**	**0.991**	**0.980**	162.745
DAVTG-Laion	Diffutoon (Duan et al. 2024)	38.610	23.700	0.563	0.959	0.953	0.979	44.263
	StyleCrafter (Liu et al. 2023)	43.161	23.246	0.780	0.926	0.897	0.978	111.225
	AnyV2V (Ku et al. 2024)	35.648	20.019	0.696	0.936	0.918	0.975	78.474
	BIVDiff (Shi et al. 2024)	36.676	20.156	0.733	0.921	0.897	0.956	158.367
	EFDM* (Zhang et al. 2022a)	35.673	21.549	0.582	0.933	0.931	0.964	**5.969**
	CAST* (Zhang et al. 2022b)	34.828	**19.974**	0.661	0.947	0.936	0.968	7.348
	UniVST (Ours)	**30.636**	21.146	**0.383**	**0.986**	**0.990**	**0.981**	162.745

the video. (4) **Inference Efficiency:** To gauge the time efficiency, we calculate the total inference time, thereby assessing their performance in terms of computational speed.

Quantitative Comparison

We have chosen a selection of video stylization techniques, including Diffutoon (Duan et al. 2024) and StyleCrafter (Liu et al. 2023), as well as two video editing frameworks: AnyV2V (Ku et al. 2024) and BIVDiff (Shi et al. 2024). Additionally, we have considered two traditional non-diffusion methods: EFDM (Zhang et al. 2022a) and CAST (Zhang et al. 2022b). Given that our specific task does not perfectly correspond with some of the existing methods, we have made necessary modifications to ensure a fair comparison. In the case of Diffutoon (Duan et al. 2024), which is designed to process text inputs for coloring tasks and does not natively support style images, we have overcome this limitation by leveraging Chat-GPT4 (Achiam et al. 2023) to generate descriptive captions for the style images. For the style generation task with StyleCrafter (Liu et al. 2023), which does not inherently support video editing, we employ inversion to obtain the initial noise of the original video, which is then utilized in the generation process. For the unified video editing frameworks, AnyV2V (Ku et al. 2024) and BIVDiff (Shi et al. 2024), we have integrated InstantStyle (Wang et al. 2024a) as the image editing model to enhance their capabilities. As for the traditional non-diffusion methods, EFDM (Zhang et al. 2022a) and CAST (Zhang et al. 2022b), which are well-aligned with our task, we have retained them in their original form without making any alterations.

We employ these six established methods as benchmarks to evaluate our proposed UniVST, with the quantitative comparisons presented in Table 1. The results demonstrate that our UniVST outperforms the others across all three performance metrics. On both the DAVTG-WikiArt and DAVTG-Laion datasets, UniVST consistently surpasses the competition in terms of foreground style preservation and temporal consistency, showcasing its robust capability for localized video style transfer. Regarding overall transfer performance, UniVST claims the top position on the DAVTG-Laion dataset and secures the second spot on DAVTG-WikiArt. Although localized transfer tasks might marginally affect overall performance, leading to a slightly higher FID score compared to other methods, these findings underscore UniVST's proficiency in effectively reconciling content fidelity with style richness.

Our inference efficiency does not stand out. This is because we opt not to adopt the temporal attention mechanisms in existing T2V frameworks (Liu et al. 2023; Huang, Zhang, and Dong 2024). Instead, we utilize the Sliding-Window Consistent Smoothing strategy, which requires dynamic optical flow estimation during the denoising process. While this approach improves temporal consistency, it also leads to increased computational overhead. In Fig 13, enlarging the sliding window size results in a proportional increase in the number of frames needed for optical flow estimation, thereby raising the time cost. This trade-off between temporal consistency and computational efficiency highlights the potential optimization in future work.

Qualitative Comparison

For a qualitative assessment, we have juxtaposed the editing results with those of six other baseline methods, as illustrated in Fig. 5. Diffutoon (Duan et al. 2024), which is renowned for its expertise in cartoon coloring, exhibits a limited responsiveness to a variety of image styles. This results in a final edited video that is notably monotonous, heavily inclined towards a cartoonish anime aesthetic. StyleCrafter (Liu et al. 2023), tailored for style-based video generation, demonstrates a high sensitivity to stylistic cues. However, it encounters difficulties in preserving content fidelity, leading to substantial distortion of background details and character features during the transfer process. BIVDiff (Shi et al. 2024) falls short in both content fidelity and temporal consistency, similarly altering background information and character traits during the transfer. AnyV2V (Ku et al. 2024) excels in maintaining temporal consistency but faces challenges with transfer tasks involving a central object. It often over-transfers the subject's style, resulting in a loss of its distinctive attributes. The traditional non-diffusion

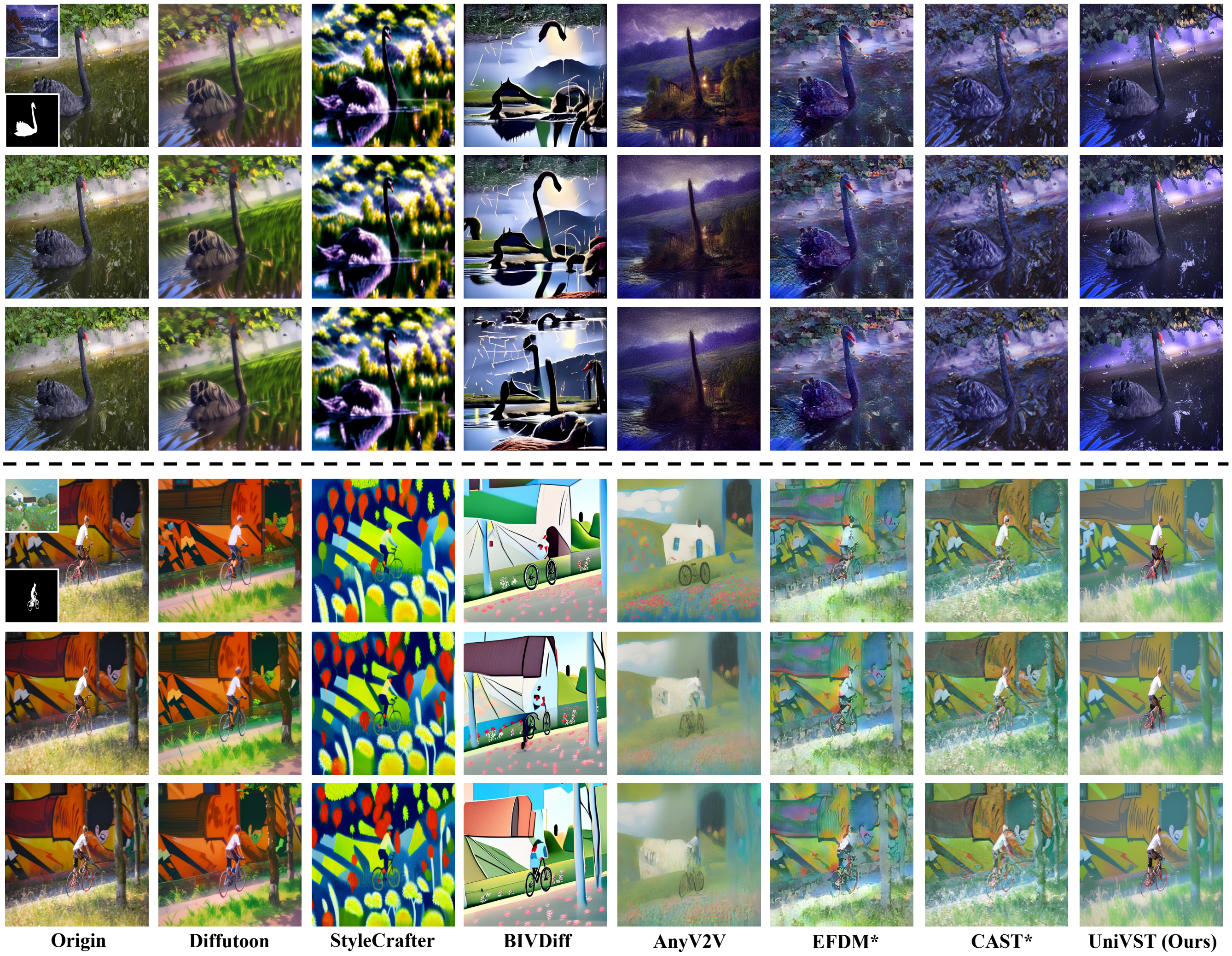

Figure 5: Qualitative comparison of our proposed UniVST with existing methods. superscript * indicates non-diffusion methods.

Table 2: Ablation study for proposed components. The **bold** values represent the best performance.

Datasets	Methods	Overall Transfer Performance			Style Preservation		Consistency
		AIFID↓	FID↓	LPIPS↓	SSIM↑	CLIP-I↑	CLIP-F↑
DAVTG-WikiArt	UniVST	**37.152**	26.134	0.369	**0.986**	**0.991**	0.980
	w/o mask-guidance	38.014	**26.050**	0.405	0.966	0.970	0.979
	w/o latent-shift	37.649	26.909	**0.349**	0.985	0.990	0.980
	w/o attention-shift	39.589	27.506	0.388	0.985	0.988	**0.983**
	w/o window-smoothing	38.626	26.907	0.384	0.982	0.989	0.970
DAVTG-Laion	UniVST	**30.636**	21.146	0.383	**0.986**	**0.990**	**0.981**
	w/o mask-guidance	30.848	**20.974**	0.404	0.961	0.967	0.979
	w/o latent-shift	31.001	21.766	**0.362**	0.985	0.989	0.980
	w/o attention-shift	32.534	22.623	0.377	0.985	0.989	**0.981**
	w/o window-smoothing	30.834	21.249	0.386	0.981	0.988	0.971

methods, EFDM (Zhang et al. 2022a) and CAST (Zhang et al. 2022b), compromise detail information during the

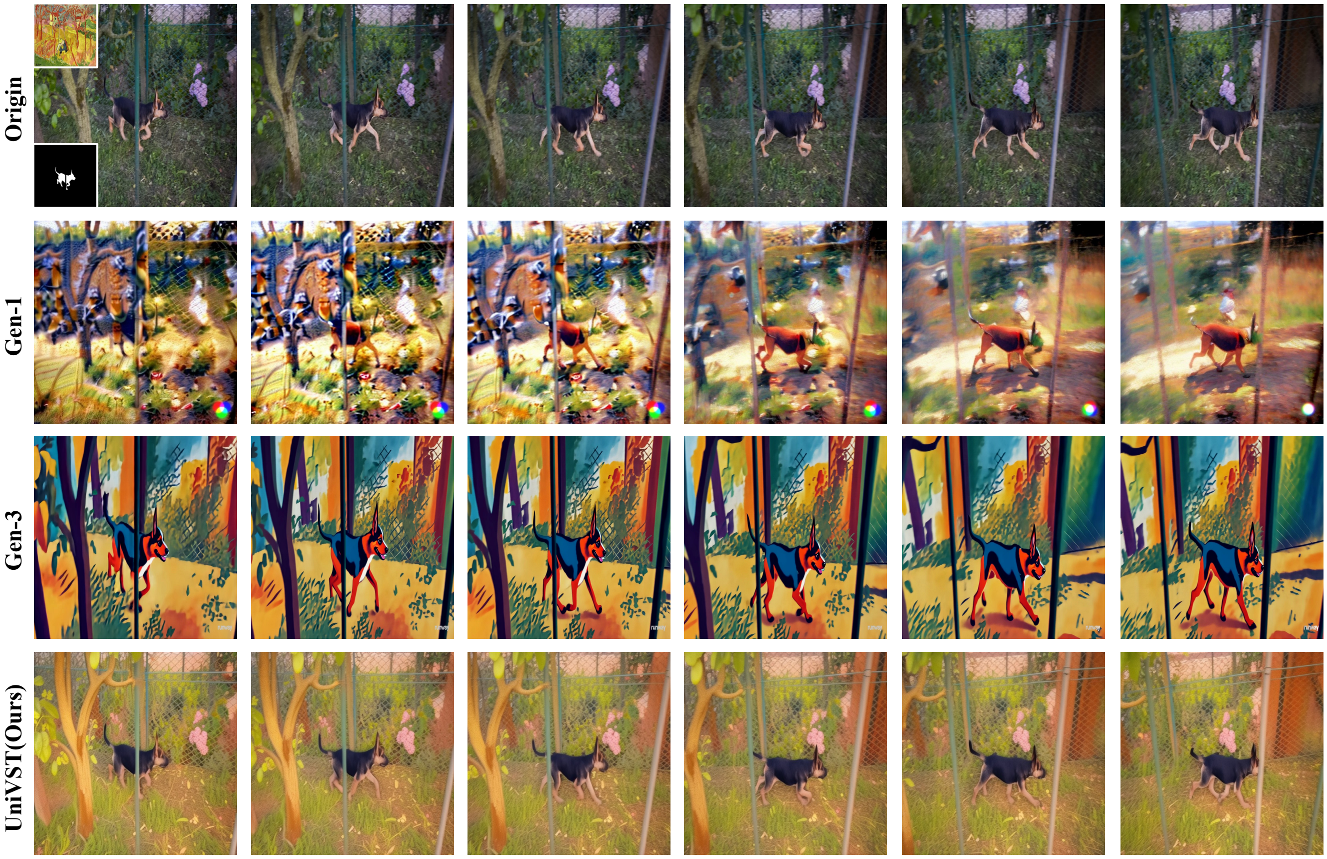

Figure 6: Qualitative comparisons of our proposed UniVST with some existing commercial models. The first row presents the original video, reference image, and mask, followed by the results of Gen-1 (Esser et al. 2023) in the second row, Gen-3 in the third row, and our proposed UniVST in the final row.

transfer process and introduce color artifacts. On the contrary, UniVST demonstrates a remarkable ability to balance content fidelity with style richness, achieving a harmonious integration of both. It ensures that the key structural and motion details of the original content are retained, while preserving the primary style of the object with great precision.

We have also compared our method with existing commercial models, namely Gen-1 (Esser et al. 2023) and Gen-3, as illustrated in Fig. 6. Our approach demonstrates clear superiority in two critical aspects: achieving a closer style similarity to the reference image and maintaining higher content fidelity to the original video. Moreover, it uniquely excels in preserving the style consistency of local objects, ensuring a more cohesive and visually appealing result, even in complex or dynamic scenes.

Ablation Studies

We have developed five distinct variants to assess the contribution of its individual components: (1) The full UniVST model, (2) UniVST without mask guidance, (3) UniVST without latent-shift, (4) UniVST without attention-shift, and (5) UniVST without window smoothing. The quantitative outcomes are presented in Table 2, while the qualitative results are displayed in Fig. 8.

Our mask-guided strategy effectively maintains the foreground style throughout the transfer process, showing its efficacy both quantitatively and qualitatively, without necessitating additional model components. The latent-shift and attention-shift strategies are both integral to the style transfer process. It is noteworthy noteworthy that the absence of the attention-shift strategy results in an incomplete transfer and a blurring of local details, as exemplified in Fig. 8. Finally, our optical flow smoothing strategy significantly enhances the modeling of temporal consistency, effectively mitigating video artifacts. In addition, we conducted an in-depth study of the three proposed modules on the DAVIS-Laion dataset to further substantiate their individual contributions and overall impact.

Mask Propagation. We first present the visualization of the final mask propagation results, as shown in Fig. 7. Then, we further discuss the impact of downsampling rates on propagation accuracy and inference speed using the DAVIS dataset (Perazzi et al. 2016). Detailed results can be found in Fig. 9. As the sampling rate decreases, the efficiency of our mask propagation is significantly improved, while effectively maintaining the accuracy of the final propagation

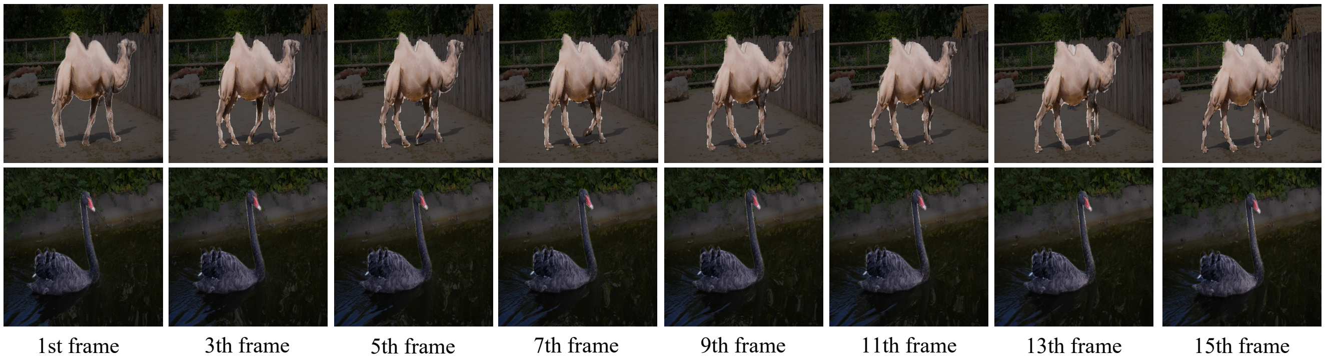

Figure 7: Visual results of mask propagation on the DAVIS dataset (Perazzi et al. 2016). The mask is blended with the original image to emphasize the subject, highlighting it while dimming the background.

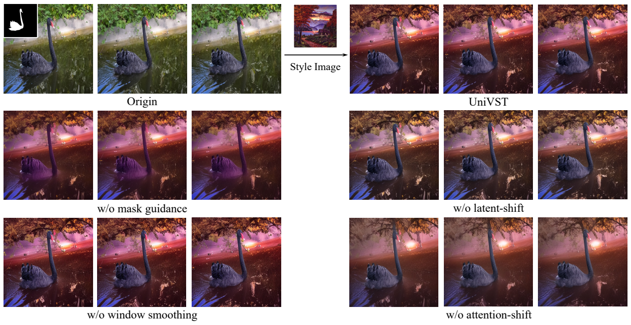

Figure 8: Visual results for ablations. Various components in our framework have all played important roles.

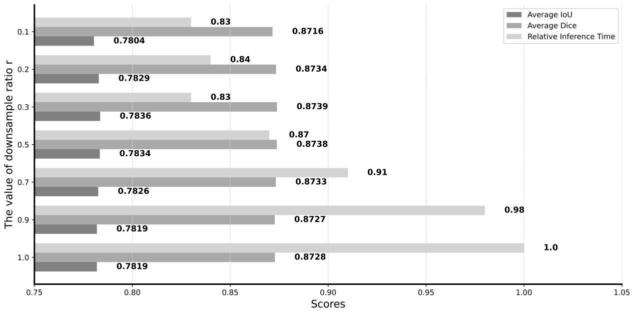

Figure 9: Quantitative results of mask propagation with different sampling ratio r on the DAVIS (Perazzi et al. 2016) dataset. Best view with zooming in.

Table 3: Performance and efficiency of mask propagation with different anchor frames. The **bold** fonts represent our final configuration.

Configuration	Avg IoU↑	Avg Dice↑	Inference Time(s)↓
Previous-1(Naive)	0.5321	0.6505	2.365
Previous-1+First	0.7799	0.8714	2.423(↓2.5%)
Previous-3+First	0.7792	0.8711	2.477(↓4.7%)
Previous-5+First	0.7800	0.8716	2.577(↓9.0%)
Previous-7+First	0.7809	0.8721	2.673(↓13.0%)
Previous-9+First	**0.7816**	**0.8726**	**2.742(↓15.9%)**
Previous-11+First	0.7820	0.8728	2.868(↓21.3%)
Previous-13+First	0.7820	0.8728	2.873(↓21.5%)
Previous-15+First	0.7820	0.8729	2.878(↓21.7%)

Table 4: Performance and efficiency of mask propagation with different values of k. The **bold** fonts represent our final configuration.

The value of k	Avg IoU↑	Avg Dice↑	Inference Time(s)↓
40	0.7806	0.8719	2.467
30	0.7836	0.8739	2.458
20	0.7836	0.8750	2.455
15	**0.7885**	**0.8772**	**2.452**
10	0.7875	0.8766	2.452

results. We also examine the impact of anchor frames selection on propagation accuracy and efficiency. As shown in Table 3, introducing more anchor frames improves propagation accuracy but also results in longer propagation times. Furthermore, we find that as the number of frames increases, the accuracy gains become less significant. After weighing these trade-offs, we select the first frame and the previous 9 frames as the final hyperparameter configuration. In addition, we perform an ablation study on the hyperparameter k of k-NN (Cover and Hart 1967). As shown in Table 4, the highest propagation accuracy is achieved at $k = 15$, without affecting propagation efficiency.

AdaIN-Guided Style Transfer. We investigate the optimal timestep interval for this method in Table 5. It means that applying attention-shift in the early and middle stages of denoising can yield better results. Futhermore, we explore the optimal timestep intervals for latent-shift, with quantitative results provided in Table 6. Applying latent-shift in the late stages yields better results. In addition, we present visual results of these ablation experiments, as shown in Fig. 10 and Fig. 11. For latent-shift, applying it in the early stages of denoising can cause issues such as content blurriness and semantic loss during the transfer process. In contrast, attention-shift produces better visual results when applied

Table 5: Quantitative results of attention-shift with different timestep intervals on the DAVIS-Laion dataset. The **bold** fonts represent our final configuration.

Timesteps	$\leq 0.6T$	$\leq 0.5T$	$\leq 0.4T$	$\leq 0.3T$	$\leq 0.2T$	$\geq \mathbf{0.4T}$	$\geq 0.5T$	$\geq 0.6T$	$\geq 0.7T$	$\geq 0.8T$
ArtFID↓	31.843	32.520	33.298	34.912	35.655	**30.658**	30.835	30.868	31.019	31.893
FID↓	21.806	22.118	22.392	23.163	23.356	**20.529**	20.704	20.761	20.845	21.324
LPIPS↓	0.396	0.407	0.424	0.445	0.464	**0.424**	0.421	0.419	0.420	0.429

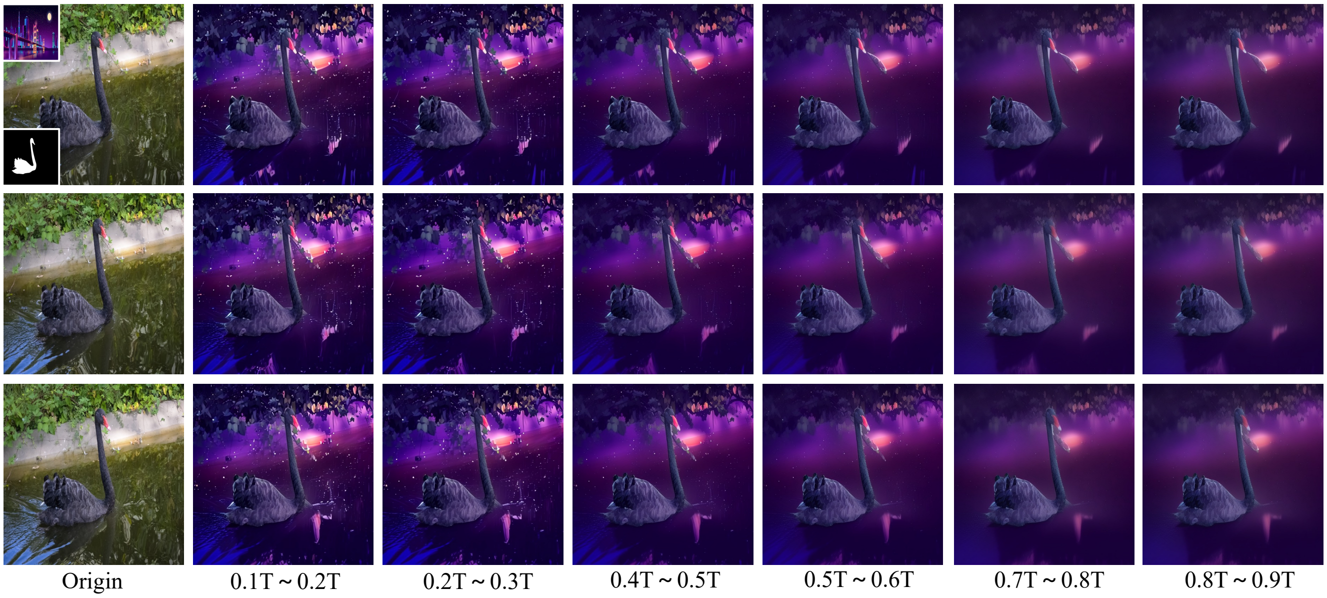

Figure 10: Visual results of latent-shift with different timestep intervals on the DAVIS-Laion.

Table 6: Quantitative results of latent-shift with different timestep intervals on the DAVIS-Laion dataset. The **bold** fonts show our final configuration.

Timesteps	**0.1T ~ 0.2T**	$0.2T \sim 0.3T$	$0.4T \sim 0.5T$	$0.5T \sim 0.6T$	$0.7T \sim 0.8T$	$0.8T \sim 0.9T$
ArtFID↓	**31.925**	32.288	32.637	32.83	33.097	32.955
FID↓	**21.722**	21.745	21.854	22.075	22.157	22.206
LPIPS↓	**0.405**	0.419	0.428	0.423	0.429	0.42

Table 7: Comparison between attention-shift and some existing training-free stylization methods on the DAVIS-Laion dataset. The **bold** fonts show our final configuration.

Configuration	The key-value replacement	The key-value AdaIN	Ours with Increasing	**Ours with Decreasing**
ArtFID↓	32.561	32.581	32.158	**30.658**
FID↓	21.389	21.829	21.620	**20.529**
LPIPS↓	0.454	0.427	0.422	**0.424**

in the early stages of denoising than in the later stages. Finally, to further investigate the effectiveness of our proposed attention-shift and the monotonicity of the time coefficient β_t, we compare the performance of four configurations: key-value replacement, key-value AdaIN, ours with increasing time coefficients, and ours with decreasing time coefficients. As shown in Table 7, using a decreasing time coefficient enables attention-shift to outperform other training-free methods.

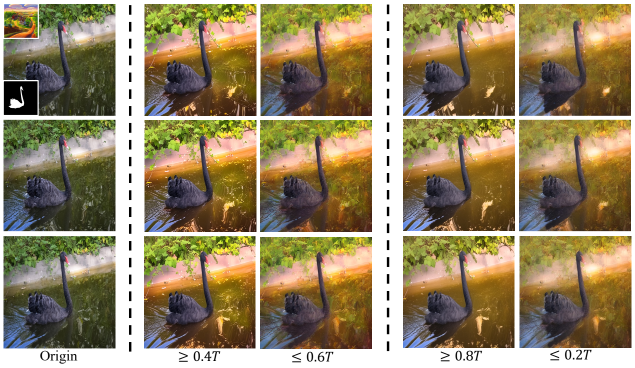

Figure 11: Visual results of attention-shift with different timestep intervals on the DAVIS-Laion. Best view with zooming in.

Sliding-Window Consistency Smoothing. To assess the impact of sliding window size on editing results and computational costs, we present the visual effects and corresponding inference times for different window sizes in Fig. 13. While larger window sizes yield slight visual improvements,

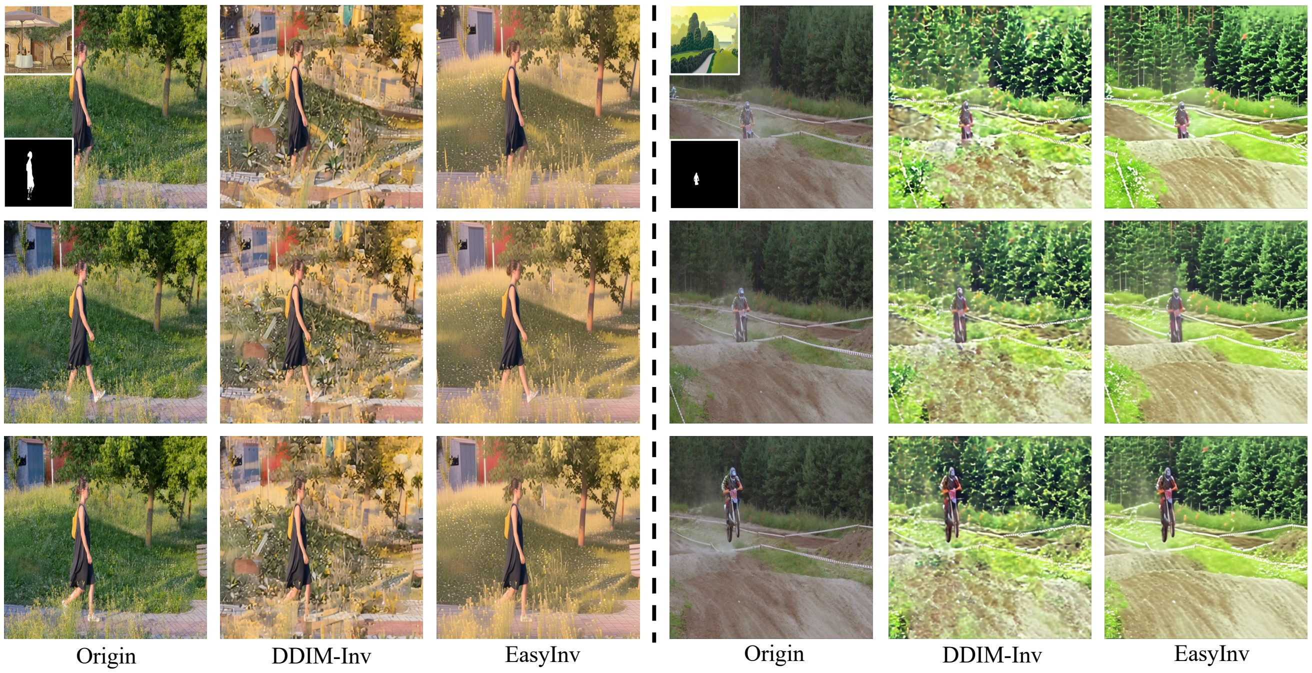

Figure 12: Qualitative comparisons of different inversion methods on transfer quality. EasyInv (Zhang et al. 2024b) demonstrates better transfer quality compared to DDIM-Inv (Song, Meng, and Ermon 2020).

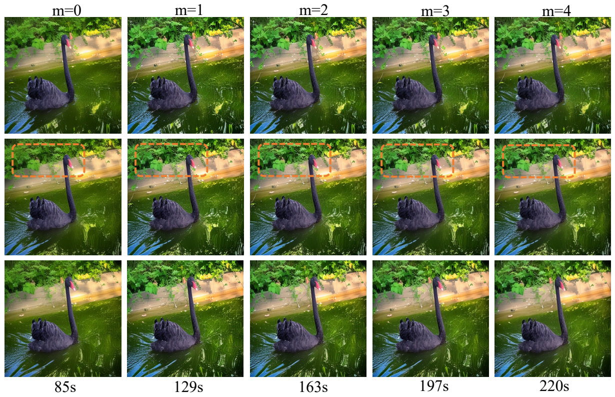

Figure 13: Visual results of sliding-window consistent smoothing with different window sizes. Best view with zooming in.

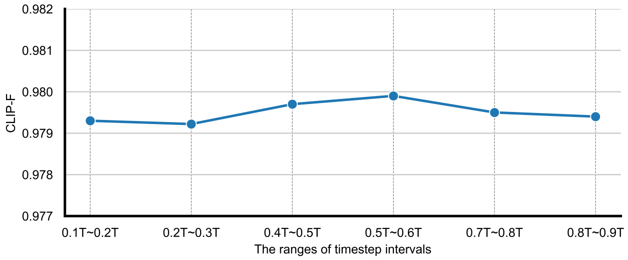

Figure 14: Quantitative results of sliding-window consistent smoothing with varying timestep intervals on the DAVIS-Laion.

they also result in significant inference costs. On the other hand, we further explore how applying this strategy at different timestep intervals affects final temporal consistency. As shown in Fig. 14, we find that the quality of smoothing strategy varies across the early, middle, and late stages of the denoising process, with the best results achieved when it is applied during the middle stage.

Impact of Inversion on UniVST. As a training-free stylization method, we need to obtain the initial noise of the original video, meaning that our method can not bypass inversion technique. Therefore, the performance of this method largely depends on the quality of the initial noise obtained through inversion. In Fig. 12, we find that naive inversion can lead to a loss of semantic details during the transfer process and adversely affect subsequent mask propagation. To minimize the errors introduced by the inversion process, we adopt EasyInv (Zhang et al. 2024b) and show better visual effects.

Conclusion

This paper presents UniVST, a unified framework for localized video style transfer that features a training-free approach, marking a significant advancement over traditional methods that require full video style transfer. Our contributions include: (1) A point-matching mask propagation strategy that eliminates the need for tracking models, enabling streamlined style transfer to specific video objects. (2) A training-free AdaIN-guided video style transfer mechanism that operates at both the latent and attention levels, ensuring a balance between content fidelity and style richness. (3) A sliding-window consistent smoothing scheme that uses optical flow to refine noise and update the latent space, improving temporal consistency and reducing artifacts. Extensive experiments show that UniVST outperforms existing methods in both qualitative and quantitative assessments, preserving the primary object's style while ensuring temporal consistency and detail preservation. Futhermore, it can be extended to a wider range of customized models.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chung, J.; Hyun, S.; and Heo, J.-P. 2024. Style Injection in Diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Cong, Y.; Xu, M.; Simon, C.; Chen, S.; Ren, J.; Xie, Y.; Perez-Rua, J.-M.; Rosenhahn, B.; Xiang, T.; and He, S. 2023. FLATTEN: Optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*.

Cover, T.; and Hart, P. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*.

Deng, Y.; He, X.; Tang, F.; and Dong, W. 2023. Z*: Zero-shot style transfer via attention rearrangement. *arXiv preprint arXiv:2311.16491*.

Ding, Z.; Li, P.; Yang, Q.; Shen, X.; Li, S.; and Gong, Q. 2024. Regional style and color transfer. *arXiv preprint arXiv:2404.13880*.

Duan, Z.; Wang, C.; Chen, C.; Qian, W.; and Huang, J. 2024. Diffutoon: High-resolution editable toon shading via diffusion models. *arXiv preprint arXiv:2401.16224*.

Duan, Z.; Wang, C.; Chen, C.; Qian, W.; Huang, J.; and Jin, M. 2023. FastBlend: A powerful model-free toolkit making video stylization easier. *arXiv preprint arXiv:2311.09265*.

Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J.; and Germanidis, A. 2023. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Gear, C. W.; and Petzold, L. R. 1984. Ode methods for the solution of differential/algebraic systems. *SIAM Journal on Numerical Analysis*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. *Advances in Neural Information Processing Systems*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, N.; Zhang, Y.; and Dong, W. 2024. Style-A-Video: Agile diffusion for arbitrary text-based video style transfer. *IEEE Signal Processing Letters*.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*.

Jeong, H.; and Ye, J. C. 2023. Ground-A-Video: Zero-shot grounded video editing using text-to-image diffusion models. *arXiv preprint arXiv:2310.01107*.

Jiang, Y.; Wu, T.; Yang, S.; Si, C.; Lin, D.; Qiao, Y.; Loy, C. C.; and Liu, Z. 2024. VideoBooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kahatapitiya, K.; Karjauv, A.; Abati, D.; Porikli, F.; Asano, Y. M.; and Habibian, A. 2024. Object-centric diffusion for efficient video editing. *arXiv preprint arXiv:2401.05735*.

Kale, K.; Pawar, S.; and Dhulekar, P. 2015. Moving object tracking using optical flow and motion vector estimation. In *Proceedings of the IEEE International Conference on Reliability, Infocom Technologies and Optimization*.

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Text2Video-Zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Ku, M.; Wei, C.; Ren, W.; Yang, H.; and Chen, W. 2024. AnyV2V: A plug-and-play framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*.

Li, S. 2024. DiffStyler: Diffusion-based localized image style transfer. *arXiv preprint arXiv:2403.18461*.

Liu, G.; Xia, M.; Zhang, Y.; Chen, H.; Xing, J.; Wang, X.; Yang, Y.; and Shan, Y. 2023. StyleCrafter: Enhancing stylized text-to-video generation with style adapter. *arXiv preprint arXiv:2312.00330*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mao, Q.; Chen, L.; Gu, Y.; Fang, Z.; and Shou, M. Z. 2023. MAG-Edit: Localized image editing in complex scenarios via mask-based attention-adjusted guidance. *arXiv preprint arXiv:2312.11396*.

Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; and Sorkine-Hornung, A. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Phillips, F.; and Mackintosh, B. 2011. Wiki Art Gallery, Inc.: A case for critical thinking. *Issues in Accounting Education*.

Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; and Chen, Q. 2023. FateZero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*.

Shi, F.; Gu, J.; Xu, H.; Xu, S.; Zhang, W.; and Wang, L. 2024. BIVDiff: A training-free framework for general-purpose video synthesis via bridging image and video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Shin, J.; Kim, S.; Kang, S.; Lee, S.-W.; Paik, J.; Abidi, B.; and Abidi, M. 2005. Optical flow-based real-time object tracking using non-prior training active feature model. *Real-Time Imaging*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tang, L.; Jia, M.; Wang, Q.; Phoo, C. P.; and Hariharan, B. 2023. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*.

Teed, Z.; and Deng, J. 2020. RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*.

Wang, H.; Wang, Q.; Bai, X.; Qin, Z.; and Chen, A. 2024a. InstantStyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*.

Wang, H.; Xing, P.; Huang, R.; Ai, H.; Wang, Q.; and Bai, X. 2024b. InstantStyle-Plus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image Quality Assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*.

Wright, M.; and Ommer, B. 2022. ArtFID: Quantitative evaluation of neural style transfer. In *Proceedings of the DAGM German Conference on Pattern Recognition*.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023a. Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Wu, J. Z.; Li, X.; Gao, D.; Dong, Z.; Bai, J.; Singh, A.; Xiang, X.; Li, Y.; Huang, Z.; Sun, Y.; et al. 2023b. Cvpr 2023 text guided video editing competition. *arXiv preprint arXiv:2310.16003*.

Yang, S.; Zhou, Y.; Liu, Z.; and Loy, C. C. 2023. Rerender A Video: Zero-shot text-guided video-to-video translation. In *Proceedings of the SIGGRAPH Asia Conference Papers*.

Yu, J.; Cun, X.; Qi, C.; Zhang, Y.; Wang, X.; Shan, Y.; and Zhang, J. 2023. AnimateZero: Video diffusion models are zero-shot image animators. *arXiv preprint arXiv:2312.03793*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhang, S.; Wang, J.; Zhang, Y.; Zhao, K.; Yuan, H.; Qin, Z.; Wang, X.; Zhao, D.; and Zhou, J. 2023. I2VGen-XL: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*.

Zhang, Y.; Gu, J.; Wang, L.-W.; Wang, H.; Cheng, J.; Zhu, Y.; and Zou, F. 2024a. MimicMotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*.

Zhang, Y.; Li, M.; Li, R.; Jia, K.; and Zhang, L. 2022a. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Zhang, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Lee, T.-Y.; and Xu, C. 2022b. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 conference proceedings*.

Zhang, Z.; Lin, M.; Yan, S.; and Ji, R. 2024b. Easy-Inv: Toward fast and better ddim inversion. *arXiv preprint arXiv:2408.05159*.

Zhang, Z.; Zhang, Q.; Xing, W.; Li, G.; Zhao, L.; Sun, J.; Lan, Z.; Luan, J.; Huang, Y.; and Lin, H. 2024c. ArtBank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In *Proceedings of the AAAI Conference on Artificial Intelligence*.