Highly Detailed and Temporal Consistent Video Stylization via Synchronized Multi-Frame Diffusion

Minshan Xie[1]   Hanyuan Liu[2]   Chengze Li[3]   Tien-Tsin Wong[1]

[1]The Chinese University of Hong Kong   [2]City University of Hong Kong

[3]Caritas Institute of Higher Education

{msxie,ttwong}@cse.cuhk.edu.hk, hy.liu@cityu.edu.hk, czli@cihe.edu.hk

Abstract

Text-guided video-to-video stylization transforms the visual appearance of a source video to a different appearance guided on textual prompts. Existing text-guided image diffusion models can be extended for stylized video synthesis. However, they struggle to generate videos with both highly detailed appearance and temporal consistency. In this paper, we propose a synchronized multi-frame diffusion framework to maintain both the visual details and the temporal consistency. Frames are denoised in a synchronous fashion, and more importantly, information of different frames is shared since the beginning of the denoising process. Such information sharing ensures that a consensus, in terms of the overall structure and color distribution, among frames can be reached in the early stage of the denoising process before it is too late. The optical flow from the original video serves as the connection, and hence the venue for information sharing, among frames. We demonstrate the effectiveness of our method in generating high-quality and diverse results in extensive experiments. Our method shows superior qualitative and quantitative results compared to state-of-the-art video editing methods.

1. Introduction

Video-to-video stylization or conversion, takes a source video (e.g. live action video) as input and converts it to a target one with the desired visual effects (e.g. cartoon style, or photorealistic one with the change of person's identity/hairstyle/dressing, etc). It can be regarded as a generalized rotoscoping, not only to produce cartoon animation, but more general ones. Due to its convenience and generality, there has be a large demand in the video content production, as observed in social platforms, such as YouTube and TikTok, even though the produced videos exhibit significant visual and temporal inconsistencies.

With the advances of large-scale data trained diffusion models, text-to-image (T2I) diffusion models [16, 32, 33] present the exceptional ability in generating diverse and high-quality images, and more importantly, its conformity to the text description given by users. Subsequent works based on T2I models [2, 9, 14, 20, 24] further demonstrate its image editing functionality. Therefore, it is natural to apply these T2I methods to the above video stylization task [5, 21, 50] by applying the *pretrained* T2I diffusion model on each frame individually (the second row of Fig. 1). However, even with per-frame constraints from the ControlNet [52], the direct T2I application cannot maintain the temporal consistency and leads to severe flickering artifacts (the third row of Fig. 1).

To maintain the temporal consistency, one can apply text-to-video (T2V) diffusion models [13, 16, 25], but with a trade-off of high computational training cost. This may not be cost effective. Some zero-shot methods [6, 21] imposes cross-frame constraints on the latent features for temporal consistency, but these constraints are limited to global styles and are unable to preserve low-level consistency, which may still exhibit flickering local structures (the fourth row of Fig. 1). A few methods utilize the optical flow to improve the low-level temporal consistency of the resultant videos. They typically warp from one frame to another, using the optical flow, patch the unknown region [5, 50], and followed by a post-processing smoothing [21, 50] for consistent appearance (warp-and-patch approach), which inevitably leads to alignment artifacts or over-blurriness (the fifth row of Fig. 1). It remains challenging to simultaneously achieve the highly detailed fidelity, the conformity to text prompt, and the temporal consistency throughout the entire video sequence.

In this paper, instead of using the optical flow for warp-and-patch, we utilize the correspondence sites, determined from the optical flow, as *portals for information sharing* among the frames. Such information sharing among frames is performed between each denoising step, hence we called it *synchronized multi-frame diffusion*. It is crucial for originally separated diffusion processes of frames to reach a

Figure 1. Our method can generate stylized frames with local visual consistency. From top to bottom: original video, SD [33], ControlNet [52], Text2Video-Zero [21], Rerender-A-Video [50] and ours. Text prompt: `"A cat with yellow eyes, oil painting."` Readers are encouraged to zoom in to better compare the fine details from different methods.

consensus, in terms of overall visual layout and color distribution, in the early stage of the diffusion process, before it is too late to fix. To achieve this, we design a multi-frame fusion stage on top of the existing diffusion model, which adds temporal consistency constraints to the intermediate video frames generated at each diffusion step. The visual content is unified among frames through consensus-based information sharing. We first propagate the content of each frame to overlapping regions in other frames. Then, each frame is updated (denoised) by fusing the propagated (shared) information received from all other frames. However, we observed that global-scale and medium-scale structure consensus can be achieved in the early denoising steps, but fine-scale detail consensus fails to be achieved with the misaligned detail generated at the later denoising steps. To prevent the generated details from being smoothed out, we propose an alternative propagation strategy that propagates the details of randomly selected frames to overwrite the overlapping regions in other frames. As each frame has an equal opportunity to propagate the details, a pseudo-equal sharing way is achieved.

We conduct extensive qualitative and quantitative experiments to demonstrate the effectiveness of our method. Our method achieves outstanding performance compared with state-of-the-art methods in all evaluated metrics. It strikes a nice balance in terms of temporal consistency and semantic conformity to user prompts. Our contributions are summarized as follows:

- Instead of warp-and-patch approach, our zero-shot method is designed based on a consensus approach, in which all frames contribute to the generation of stylized content, in an equal and synchronized fashion.
- We propose to seamlessly blend the shared content from different frames using a novel Multi-Frame Fusion.

2. Related Work

Text-Driven Image Editing. Advancements in computer vision have led to significant progress in natural image editing. Before the rise of diffusion models [15, 40], various GAN-based approaches [11, 12, 27, 47] achieved commendable results. The emergence of diffusion models has elevated the quality and diversity of edited content even further. SDEdit [24] introduces noise and corruptions to an input image and then leverages diffusion models to reverse the process, enabling effective image editing. But, it suffers from the loss of fidelity. Prompt-to-Prompt [14] and Plug-and-Play [44] perform semantic editing by blending activations from original and target text prompts. Uni-Tune [45] and Imagic [20] focus on finetuning a single image for improved editability while maintaining fidelity. Researchers have also explored aspects like controllability [3, 18, 22, 36, 52] and personalization [10, 35] in diffusion-based generation, enhancing our understanding of how to tailor diffusion models to specific editing needs. Our proposed method builds upon existing image editing techniques [26, 28, 52] to preserve structural integrity and generate videos with temporal consistency.

Text-Driven Video Editing. Video editing poses unique challenges for diffusion-based methods compared to image editing, primarily due to the intricate requirements of geometric and temporal consistency. While image editing has seen significant progress, extending these advancements to videos remains a complex task. Text-to-Video (T2V) Diffusion Models [17] have emerged as a promising avenue. These models build upon the 2D U-Net architecture used in image models but extend it to a factorized space-time UNet [13, 16, 38, 51, 54]. Dreamix [25] focuses on motion editing by developing a text-to-video backbone while ensuring temporal consistency. Make-A-Video [38] leverages unsupervised video data and learns movement patterns to drive the image model. However, these methods require substantial video data for training. StableVideo [5] employs a compressed representation as the propagator for consistent video editing. It generates the appearance of the next frame based on warped information from the previous one. However, it requires additional training for the compressed representation and may involve suboptimal training for an

aggregation network to unify the edited foreground appearance.

Some recent efforts aim to make video editing more cost-effectively. Methods like Tune-a-Video [48], Uni-Tune [45], and Imagic [20] propose fine-tuning pre-trained T2I diffusion models on single videos to achieve consistent video editing. However, modeling complex motion remains a challenge. Some zero-shot methods [21], such as Text2Video-Zero [21] and ControlVideo [53], impose cross-frame constraints on latent features for temporal consistency and use ControlNet [52] for controllable video editing. However, these constraints are often limited to global styles and struggle to preserve low-level visual consistency.

Several methods have emerged to address the challenge of maintaining consistency across frames while preserving visual quality, relying on key frames [19, 43, 49] or optical flow [34] to propagate contents between frames. FLAT-TEN [6] introduces a flow-guided attention mechanism that leverages optical flow to guide the attention module during the diffusion process. However, as these methods operate in the latent domain, they may lead to low-level visual inconsistencies. Rerender-A-Video [50] utilizes optical flow to apply dense cross-frame constraints. It gradually inpaints the next frame by warping the overlapping region from the previous one. The fused regions combine to form the final output. However, the results tend to be blurry, as a smoothing operation is employed to avoid artifacts during fusion. Additionally, it may introduce inconsistent styles for disoccluded regions. Different with existing methods which follow a warp-and-patch strategy and a subsequent merging step, we propose to impose the temporal coherence with synchronized multi-frame diffusion to reach a consensus for all frames, in which all frames contribute more-or-less equally.

3. Preliminary

Diffusion Models [39] are powerful probabilistic models that gradually denoise data, effectively learning the reverse process of a fixed Markov Chain [7, 15]. These models aim to learn the underlying data distribution $p(x_0)$ by iteratively denoising a normally distributed variable. The denoising process involves a sequence of denoising networks, denoted as $\epsilon_\theta(x_t, t)$; $t = 1, \ldots, T$. The model is trained to predict a denoised variant of its input x_{t-1} from x_t, where x_{t-1} and x_t represents the noisy version of the original input x_0. Besides, the problem can also be transformed to predict a clean version $x_{0|t}$ from x_t as we can sample x_{t-1} based on $x_{0|t}$ with a deterministic DDIM sampling [40, 41].

Latent Diffusion Models (LDMs) [33] employ perceptual compression through an autoencoder architecture, consisting of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. LDMs learn the conditional distribution $p(z|y)$ of condition y, where z represents the latent representation obtained from the encoder $\mathcal{E}$. The decoder $\mathcal{D}$ aims to reconstruct the original input x from this latent representation, i.e., $\mathcal{E}(x) = z$, $\mathcal{D}(\mathcal{E}(x)) \approx x$. The loss function quantifies the discrepancy between the noisy input and the output of the neural backbone. The neural backbone is generally realized as a denoising U-Net with cross-attention conditioning mechanisms [46] to accept additional conditions.

Conditional Generation. Natural language is flexible for global style editing but has limited spatial control over the output (the second row in Fig. 1). To improve spatial controllability, Zhang et al. [52] introduced a side path called *ControlNet* for Stable Diffusion to accept extra conditions, such as edges, depth, and human pose. ControlNet is often used to provide structure guidance from the input video to improve temporal consistency. However, ControlNet alone is insufficient to ensure medium- and fine-scale consistencies in terms of color and texture, across the frames (the third row in Fig. 1). To address this issue, cross-frame attention mechanisms [21] are further applied to all sampling steps for global style consistency on the latent features. These constraints are limited to global styles and lead to color jittering and fine-scale visual inconsistencies (the forth row in Fig. 1).

In contrast, we aim to generate a new video, in a style specified by text prompt, not just with temporal consistency, but also visual consistency in global, medium and fine scales. These consistencies are accomplished via sharing information among frames, using our proposed Synchronized Multi-Frame Diffusion process.

4. Synchronized Multi-Frame Diffusion

Given a video with N frames $\{\mathbf{I}_i\}_{i=0}^N$, our goal is to render it into a new video $\{\mathbf{I}_i'\}_{i=0}^N$ in a style specified by a text prompt. The stylized video shall mimic the motion of the original video, and maintain the temporal consistency and visual consistencies in all scales. To achieve this, we first assign a T2I diffusion process to each frame to generate the desired style. The major challenge here is on how to generate consistent frames in all visual scales. Instead of warping the generated content from one view to another and then smoothing as in previous approaches [5, 50], we propose a consensus-based approach in which all frames share their latent information among each other during each denoising time step. We call this method, *Synchronized Multi-Frame Diffusion* (SMFD).

As a frame must overlap with its neighboring frames, the generated content within the overlapping regions should be consistent. In other words, these overlapping regions (obtained via optical flow) can serve as a venue for latent information sharing among the frame diffusion processes. For each denoising time-step, the latent information from all frame diffusion processes are first combined before the next round of denoising. Fig. 2 shows our proposed video
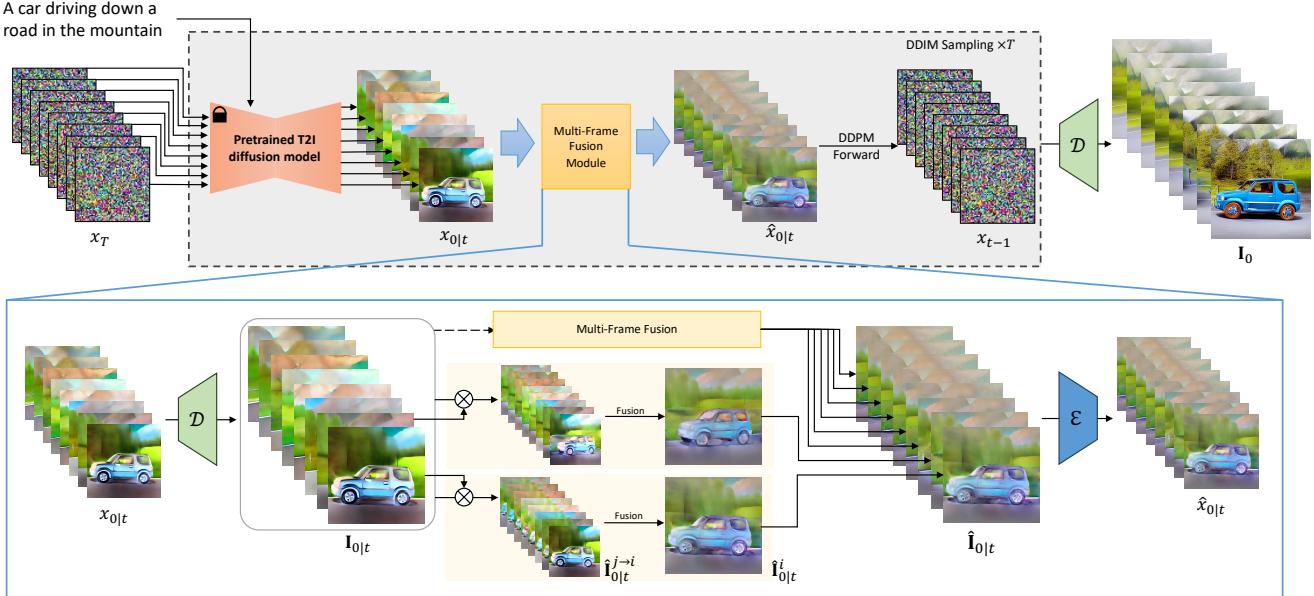
Figure 2. Framework of the proposed zero-shot text-guided video stylization. We first adopt a pretrained T2I model with cross-frame attention layers to generate stylized frames with global style consistency. The stylized frames are refined to render consistent frames in terms of visual content, color distribution, and temporal motion, using our Multi-Frame Fusion Module at each denoising step.

stylization framework.

To combine the contribution from all overlapped neighboring frames, we can warp the content from all involved neighboring frames to the current frame of interest and fuse them together using a Poisson solver [29, 42]. Disoccluded regions and image border in the warped content can be seamlessly handled in the gradient domain during the Poisson solving. Such fusion is performed for each frame with diffusion attached. This Multi-Frame Fusion Module is detailed in Sec. 4.1. With this information sharing among frames, the consensus in terms of color distribution and the overall structure can be reached in the early stage of the denoising process.

Although directly combining content from all involved frames can well unify the coarse-level visual content among frames during the early denoising steps (*semantic reconstruction stage*), it smoothens out the high-frequency details in the later denoising steps (*detail refinement stage*), leading to over-blurriness. To avoid smoothing out the fine details, we adopt an alternating propagating strategy during the detail refinement stage. We propagate the generated details of a randomly selected frame to overlapping region in other frames and overwrite (instead of fusing) the conflict details. A random frame is selected in each denoising step to encourage the contribution from involving frames. With such design, we can achieve both highly detailed fidelity and temporal consistency throughout the entire video sequence. In all our experiments, we treat the first half of denoising steps, $\frac{T}{2} < t < T$, as the semantic reconstruc-

tion stage, and the second half, $0 < t \leq \frac{T}{2}$, as the detail refinement stage.

4.1. Multi-Frame Fusion Module

In our framework, we adopt the pretrained T2I diffusion models with structure control [52] and cross-frame attention mechanism [21] to create stylized frames $\{\mathbf{I}_i^t\}_{i=0}^N$. In order to achieve pixel-level visual consistency, we perform multi-frame fusion in the image domain. We tackle the problem by updating each frame with the appearance information received from other frames, thereby achieving consensus among all frames. One important question is how to propagate the information of appearance across frames to achieve consistency. A simple way is to directly update the current frame using the overlapping region of other frames. However, it is obvious that there will be seams between the updated overlapping region and the rest of the region (Fig. 3(c)), due to the disocclusion.

Inspired by Ebsynth [19], we propose to blend the warped appearance from other frames in the gradient domain, and then solve for the images using Poisson equation. This generates multiple seamless candidates. These seamless candidates can further update the current frame without producing obvious seams. For every frame $\hat{\mathbf{I}}_{0|t}^j$, we can warp it to the pose of frame $\hat{\mathbf{I}}_{0|t}^i$, and yield a candidate image $\hat{\mathbf{I}}_{0|t}^{j \to i}$, in which its appearance follows $\hat{\mathbf{I}}_{0|t}^j$, but pose follows $\hat{\mathbf{I}}_{0|t}^i$. Fig. 4 shows all candidate images of a
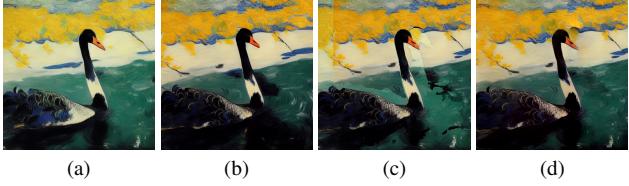
(a)   (b)   (c)   (d)

Figure 3. We use poisson image editing to seamlessly blend the overlapping region. (a) I_t^i, (b) I_t^j, (c) Copy-and-paste exhibits obvious seams, (d) Poisson blending.
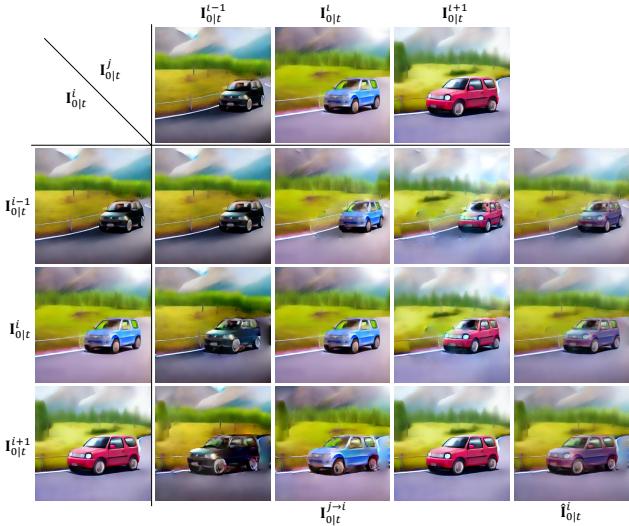

Figure 4. For each frame, we can generate multiple candidates following similar color distribution with the other frames. Thus, the fused frames can have similar appearance among all frames.

3-frame video. Each of the black, blue and red cars are warped to all possible poses (Fig. 4, middle 3×3 table). By combining all candidates, the fused frame can have similar appearance (car with a mixture appearance of black, blue and red) among all frames (Fig. 4, the rightmost column), and thereby achieving the visual consistency.

While the overall semantic structure and color distribution can be preserved by above fusion, the details may be damaged due to misalignment of fine textures from different frames (Fig. 5). To generate consistent frames with high fidelity, we adopt a pseudo-equal sharing way by alternatively propagating the details of randomly selected frames to overwrite the conflict textures during the later denoising steps.

Shared information propagation. Each predicted frame $\mathbf{I}_{0|t}^i$ is firstly warped to other frames using optical flow and generates candidate edited frames for combination. However, directly copying the overlapped region from other frames and pasting it onto the current frame leads to large abrupt intensity changes or seams. Thus, we propose to
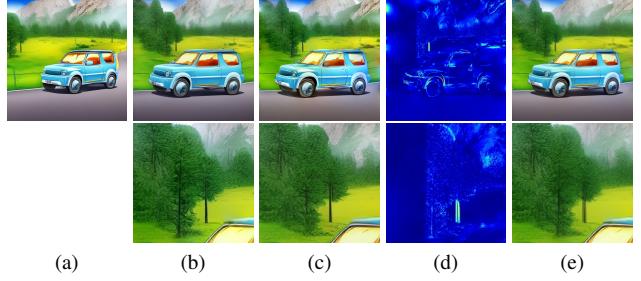

(a) (b) (c) (d) (e)

Figure 5. Details in different frames with misalignment can lead to blurriness after averaging. (a) Frame 1, (b) Frame 2, (c) Poisson blended image, (d) difference of (b)&(c), (e) fused image of (b)&(c).

seamlessly blend the occluded regions to the warped frame using a Poisson solver [29, 42]. The idea is to reconstruct pixels in the blending region such that the boundary of warped content owns a zero gradient. Fig. 3(c) shows the obvious seam of the warped image boundary if we simply copy-and-paste the warped content, while no seam is observed if we adopt the Poisson blending in Fig. 3(d). Then we can generate a candidate (warped) frames $\hat{\mathbf{I}}_{0|t}^i$ at timestep t with

$$\hat{\mathbf{I}}_{0|t}^{j \to i} = \text{PIE}(\mathbf{I}_{0|t}^i, w_j^i(\mathbf{I}_{0|t}^j), M_j^i), \qquad (1)$$

where w_j^i and M_j^i denote the optical flow and occlusion mask from $\mathbf{I}_j$ to $\mathbf{I}_i$, respectively. $\text{PIE}(\cdot, \cdot, \cdot)$ donates the Poisson solver [29, 42] which seamlessly blends the masked region of $\mathbf{I}_{0|t}^i$ into $w_j^i(\mathbf{I}_{0|t}^j)$. Thus, $\hat{\mathbf{I}}_{0|t}^{j \to i}$ can follow the color appearance of $w_j^i(\mathbf{I}_{0|t}^j)$.

Candidates Fusion at Semantic Reconstruction Stage. We then need to fuse these candidate frames to guarantee consistent geometric and appearance among all stylized frames. For frame $\mathbf{I}_{0|t}^i$, we can obtain $N-1$ candidate frames $\hat{\mathbf{I}}_{0|t}^{j \to i}$ which has the same geometric structure but different color appearances. The updated frames is the simply average value of all candidate frames.

$$\hat{\mathbf{I}}_{0|t}^i(p) = \frac{1}{N} \sum_{j=0}^{N} \hat{\mathbf{I}}_{0|t}^{j \to i}(p), \qquad (2)$$

where p is the position. With this, every frame overlapping with the current frame can contribute to the denoising process of the current frame. Consensus in overall structure and color appearance can be reached quickly in the early semantic reconstruction stage of the denoising process.

Candidates Fusion at Detail Refinement Stage. However, the above fusion by averaging may smooth out the

5

high-frequency details generated during the detail refinement stage due to misalignment (Fig. 5). To generate consistent high-frequency details for corresponding regions, we propagate the generated detail with alternating sampling strategy during the detail refinement stage. We randomly anchored one stylized frame $\hat{\mathbf{I}}_{0|t}^{j} = \mathbf{I}_{0|t}^{j}$ at each timestep and propagate the details to overlapping regions in other frames $\hat{\mathbf{I}}_{0|t}^{i} = \hat{\mathbf{I}}_{0|t}^{j \to i}$ to overwrite conflict textures. With this pseudo-equal sharing way, we can generate consistent appearance with highly-detail fidelity.

5. Experimental Results

5.1. Experimental Settings

In practice, we implement our approach over stable diffusion v1-5 [33]. We use VideoFlow [37] for optical flow estimation and compute the occlusion masks by forward-backward consistency check [23]. We choose the canny edge condition branch from [52] as the structure guidance in our method. We apply our method on several videos from DAVIS [30]. The image resolution is set to 512×512. We employ DDPM [15] sampler with 20 steps. All experiments are conducted on an NVIDIA GTX3090 GPU. In terms of running time, a 512×512 video clip with 8 frames takes about 45 seconds to generate.

5.2. Comparison with State-of-the-Art Methods

In this section, we compare our editing results with three recent zero-shot methods: FateZero [31], Text2Video-Zero (T2V-Zero) [21], and Rerender-A-Video (RAV) [50], and two methods with extra training: AnimateDiff (AD) [13] and StableVideo [5]. Besides, we also select Control-Net [52] as a competitor to evaluate the geometric constraint. As the official code of AnimateDiff [13] does not support ControlNet [52], it fails to generate video with similar geometry as the original video. Thus, we re-implement it to support ControlNet [52] for comparison, named AnimateDiff+.

Figures 6 and 7 present the visual results. FateZero [31] will fail to edit the input video when it fail to extract correct cross-attention map for the user text prompt, leading to stylized frames similar to the input video. While each frame generated by Text2Video-Zero [21] is of high quality and generate consistent global style, they may suffer from color jittering and lack of consistency in medium- and fine-scale details/texture. Because Rerender-A-Video [50] follows a continuous generation, stylized frames may suffer from over-blurring in later frames (readers are encouraged to blow up the figure for better inspection). Animate-Diff+ can produce frames with rich textures, but it does not follow the motion of the original movie. For example in the Fig. 6(f) camel example, the panned background in the original video becomes static in their stylized output.

Table 1. Quantitative comparison. The best score in **bold** and the first runner-up with underline.

Methods	Fram-Acc ↑	Feat-Con ↑	Mont-MSE ↓
StableDiffusion	0.9104	0.8545	167.5751
Controlnet	0.7478	0.8828	93.1104
FateZero	0.2133	**0.9814**	**12.0448**
T2V-Zero	0.7502	0.9443	50.2440
Rerender-A-Video	0.5319	0.9556	43.6998
AnimateDiff+	**0.7940**	0.9707	23.3980
Ours	<u>0.7891</u>	<u>0.9785</u>	<u>20.0540</u>

This negligence of motion is also reflected in our quantitative evaluation of temporal consistency in Table 1 (metrics Mont-MSE). Although StableVideo [5] can produce temporally consistent video, it can produce noticeable seams along background and foreground objects (Fig. 7). In contrast, our proposed method shows clear superiority on generating frames with temporal consistency and clear texture details.

For quantitative evaluation, we follow other methods [4, 31, 50] to compute CLIP-based frame-wise editing accuracy (Fram-Acc), and CLIP-based frame-wise cosine similarity between consecutive frames (Feat-Con). Fram-Acc evaluates whether the generated frames align with the target text prompt, while the Feat-Con evaluates whether consecutive frames shares similar image features. Additionally, we employ the motion consistency of dense optical flow (Mont-MSE) of the edited video frames from Stable-Video [5]. The Farneback algorithm [8] in OpenCV [1] is employed to calculate the average L2 distance of dense optical flow between the edited and original videos. We manually collect 50 video clips, each with 8 frames, and generate stylized videos with 11 artistic styles, e.g. water coloring style, oil painting style, Chinese ink painting style, Pixar style, etc. We additionally compare with the pretrained T2I diffusion model [33] for baseline. As StableVideo requires extra training for compressed representation of a video, we did not quantitaively compare it due to the limited resource.

Table 1 lists the evaluation scores. As results of FateZero [31] closely resemble the input video and may ignore the user text prompt, the method therefore obtains the lowest Fram-Acc score. On the other hand, although AnimateDiff+ highly respects the user text prompt and obtains the highest Fram-Acc score, it receives a lower Feat-Con and Mont-MSE scores, i.e. weaker temporal consistency, as it sometimes ignores the motion of the original video, as demonstrated by the relatively static background in their camel and car results of Fig. 6(f). In sharp contrast, our method highly respects the user prompt (first runner-up Fram-Acc), and faithfully follows the motion in the input video and never comes up with a static background (first runner-up in both temporal consistency scores Feat-Con and

Figure 6. Stylized results comparison. Our method can generate consistent results with more details. Text prompts: `"A camel is walking in the dirt, Van Gogh style."` and `"A small car driving down a road in the mountains, water coloring."` Readers are encouraged to zoom in to better compare the fine details and visual content consistency of different methods.

Mont-MSE). Note that even FateZero obtains highest Feat-Con and Mont-MSE, it is too similar to the input video to be useful. In other words, our method strikes a nice balance in both the semantic conformity to the user prompt and the motion of the input video, while producing highly detailed texture content.

5.3. Ablation Study

Multi-Frame Fusion Module As the core of our research, we evaluate the impact of the Multi-Frame Fusion Module. Its objective is to allow information sharing among frames, and hence, ensure the visual consistency among frames in all scales. Fig. 8 shows an example, where

color and structure inconsistency exists without our proposed Multi-Frame Fusion Module.

Poisson Image Editing Fig. 9 illustrates the effectiveness of Poisson solver in blending candidates to achieve information sharing across frames. For evaluation, we generate candidate regions by directly merging overlapping regions with disoccluded regions. We can see that there are noticeable seams in the final results. This is because the generated appearance of the cat between two frames may not match, leading to abrupt intensity changes along the merged boundaries.

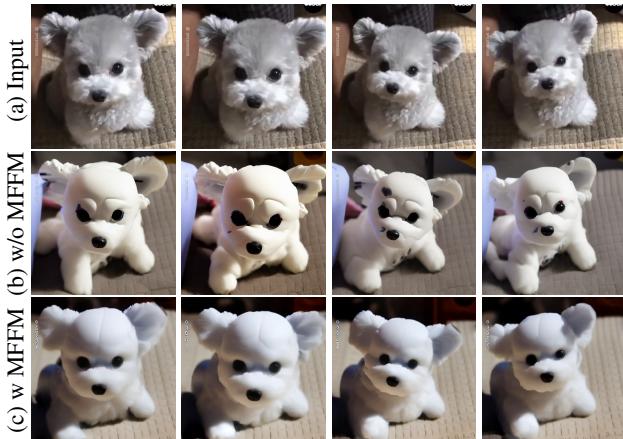Figure 7. Stylized results comparison to StableVideo [5]. Text prompt: "A duck in winter snowy scene."

Figure 8. Ablation study of Multi-Frame Fusion Module (MFFM). Without MFFM, the appearances of frames are very inconsistent. This evidences the importance of information sharing via our MFFM. Text prompt: "A robotic dog."
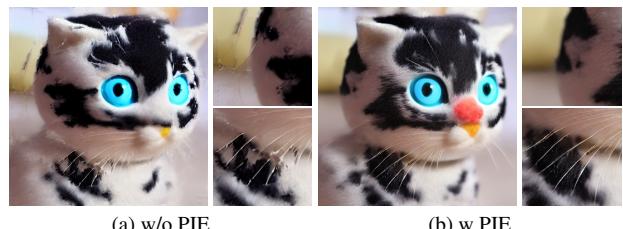

Figure 9. Ablation study of Poisson Image Editing (PIE). Poisson solving effectively avoids the obvious seams/fragmentation at the overlapping regions. Text prompt: "A detailed woolen toy cat."

Alternating Detail Propagation In addition, we also conducted experiments on the alternating detail propagation as shown in Fig. 10. Merging all candidates can guarantee consistency but it may smooth out the fine details when

Figure 10. Ablation study of alternating detail propagation (ADP). Without ADP, fine details are smoothened out at overlapping regions. Text prompt: "A black swan is swimming on the water, Van Gogh style."

Figure 11. **Failure case.** Inconsistent stylized regions or undesirable deformation (e.g. sunglasses) may be resulted with inaccurate optical flow. Text prompt: "A cat with sunglasses is eating a strawberry on the beach."

conflict textures appear among frames during denoising steps. We can see that the feathers of the swan and flower in the background are blurred. In contrast, our pesudo-sharing strategy can help generate consistent appearance across frames while preserving the high-frequency details.

5.4. Limitations

Firstly, our multi-frame fusion steps rely on optical flow for information sharing. Therefore, inaccurate optical flow estimation may lead to inconsistent appearance. Moreover, our proposed method may fail to change the geometry of the original video as we rely on the Canny edge condition. In Fig. 11, when changing the rabbit to a cat, the optical flow at the area with geometry changes will be incorrect, resulting in distortion and blurriness at the ears and sunglasses.

6. Conclusion

We propose a zero-shot text-driven approach for video stylization. We design a multi-frame fusion module to generate stylized videos with high-detailed fidelity and temporal consistency. We utilize the optical flow of the original video as a correspondence site to share information among edited frames. Our extensive experiments and demonstrate that our approach achieves outstanding qualitative and quantitative results compared to state-of-the-art methods. Unlike the previous methods which may exhibit serious visual arti-

facts of certain forms, our method produce high-quality results that highly respects the user text prompt semantically, and simultaneously,respects the motion in the given video.

References

[1] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 6

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1

[3] Shidong Cao, Wenhao Chai, Shengyu Hao, and Gaoang Wang. Image reference-guided fashion design with structure-aware transfer by diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3524–3528, 2023. 2

[4] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 6

[5] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023. 1, 2, 3, 6, 8

[6] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023. 1, 3

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[8] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, pages 363–370. Springer, 2003. 6

[9] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 1

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[11] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1, 2, 6

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

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3, 6

[16] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1, 2

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 2

[18] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 2

[19] Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. Stylizing video by example. *ACM Transactions on Graphics (TOG)*, 38(4):1–11, 2019. 3, 4

[20] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 1, 2, 3

[21] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 1, 2, 3, 4, 6

[22] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2

[23] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 6

[24] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1, 2

[25] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 1, 2

[26] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning

adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[27] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2

[28] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[29] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Transactions on Graphics (TOG)*, 22 (3):313–318, 2003. 4, 5

[30] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 6

[31] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv:2303.09535*, 2023. 6

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 6

[34] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. In *Pattern Recognition: 38th German Conference, GCPR 2016, Hannover, Germany, September 12-15, 2016, Proceedings 38*, pages 26–36. Springer, 2016. 3

[35] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[36] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[37] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. *arXiv preprint arXiv:2303.08340*, 2023. 6

[38] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[41] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020. 3

[42] Jian Sun, Jiaya Jia, Chi-Keung Tang, and Heung-Yeung Shum. Poisson matting. In *ACM SIGGRAPH 2004 Papers*, pages 315–321. 2004. 4, 5

[43] Ondřej Texler, David Futschik, Michal Kučera, Ondřej Jamriška, Šárka Sochorová, Menclei Chai, Sergey Tulyakov, and Daniel Sýkora. Interactive video stylization using few-shot patch-based training. *ACM Transactions on Graphics (TOG)*, 39(4):73–1, 2020. 3

[44] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2

[45] Dani Valevski, Matan Kalman, Eyal Molad, Eyal Segalis, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning a diffusion model on a single image. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2, 3

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[47] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2

[48] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 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*, pages 7623–7633, 2023. 3

[49] Yiran Xu, Badour AlBahar, and Jia-Bin Huang. Temporally consistent semantic video editing. In *European Conference on Computer Vision*, pages 357–374. Springer, 2022. 3

[50] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 1, 2, 3, 6

[51] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18456–18466, 2023. 2

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 2, 3, 4, 6

[53] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 3

[54] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2