StableVideo: Text-driven Consistency-aware Diffusion Video Editing

Wenhao Chai[1] * Xun Guo[2]✉ Gaoang Wang[1] Yan Lu[2]

[1] Zhejiang University [2] Microsoft Research Asia

{wenhaochai.19, gaoangwang}@intl.zju.edu.cn, {xunguo, yanlu}@microsoft.com

Figure 1: **Editing results of StableVideo.** The input video (top row) contains long range motion and viewpoint changing. Our approach performs stable editing according to the prompt of "*Orange SUV in sunny snow winter*" on foreground and background. In the edited video (bottom row), the "*orange SUV*" maintains high geometric and temporal consistency, although the viewpoints keep changing.

Abstract

Diffusion-based methods can generate realistic images and videos, but they struggle to edit existing objects in a video while preserving their appearance over time. This prevents diffusion models from being applied to natural video editing in practical scenarios. In this paper, we tackle this problem by introducing temporal dependency to existing text-driven diffusion models, which allows them to generate consistent appearance for the edited objects. Specifically, we develop a novel inter-frame propagation mechanism for diffusion video editing, which leverages the concept of layered representations to propagate the appearance information from one frame to the next. We then build up a text-driven video editing framework based on this mechanism, namely StableVideo, which can achieve consistency-aware video editing. Extensive experiments demonstrate the strong editing capability of our approach. Compared with state-of-the-art video editing methods, our approach shows superior qualitative and quantitative results. Our code is available at this https URL.

1. Introduction

Recent years have witnessed significant progress in extensive computer vision tasks taken by deep learning. Nevertheless, natural video editing, which aims at manipulating the appearance of target objects and scenes, still faces two essential challenges that are deterministic to the editing quality: the *generator* equipped with rich prior knowledge that consistently produces high-fidelity edited contents adhering faithfully to the original geometry of the target objects, and the *propagator* that disseminates the edited contents throughout the entire video while keeping highly temporal consistency.

The flourish of text-driven generative diffusion models pre-trained on large-scale image and language data [34, 14, 53, 16, 41, 5] provides impressive generation quality. Several diffusion-based methods achieve good performance in image editing [2, 31], but few methods have tried to apply diffusion models in video editing, since it is challenging to modify existing objects while preserving their appearance over the entire video [49, 12, 26]. Dreamix [27] proposes a solution to generate consistent video according to input image/video and prompts. However, it focuses more on generating smooth motions, *e.g.*, pose and camera movements, rather than maintaining geometric consistency of the objects across time. Moreover, such video diffusion models often suffer from huge computing complexity which is not friendly for practical applications.

Neural layered atlas (NLA) [24, 23] tries to tackle the temporal continuity problem by decomposing the video into a set of atlas layers, each of which describes one target object to be edited. For each atlas layer, the positions of the

*The work was done when the author was with MSRA as an intern.

video are mapped into the corresponding 2D positions in it, so that semantically correspondent pixels over the whole video can be represented by the same atlas position. Instead of frame-by-frame editing, NLA edits atlas layers to ensure that the modifications can be precisely mapped back to video frames for temporal smoothness. Text2LIVE [1] provides a text-driven appearance manipulation solution of adding additional edit layers on atlases, in which a specific generator for the edit layers is trained. Although it achieves good results with strict structure preserved, it is not able to apply thorough editing. Moreover, the specifically trained generator also limits the richness of the generated contents.

This brings up the question: *Could text-driven diffusion video editing achieve high temporal consistency*? Intuitively, employing text-driven diffusion models to edit the atlases corresponding to the target objects could reach such goal. However, this gives rise to drawbacks rather than benefits. Being the summary of the whole video, atlases always have distorted appearance due to the viewpoint and camera movement, which are required to be specifically pretrained and generated as in [1]. Diffusion models may fail in generating satisfied atlas pixels in many cases, so that the corresponding edited frames will also be contaminated. To answer the question, we present two concepts for utilizing diffusion models in video editing. Firstly, instead of editing the atlases directly, we propose to update the atlases via editing key video frames. Secondly, we introduce temporal dependency constraints for diffusion models to generate objects with consistent appearance across time.

Based on analysis above, we present a novel diffusion video editing approach, StableVideo, to perform consistency-aware video editing. In specific, we propose two effective technologies for this purpose. Firstly, to edit the objects with consistent appearance, we design an inter-frame propagation mechanism on top of the existing diffusion model [55], which can generate new objects with coherent geometry across time. Secondly, to achieve temporal consistency by leveraging NLA, we design an aggregation network to generate the edited atlases from the key frames. We then build up a text-driven diffusion-based framework, which provides high-quality natural video editing. We conduct extensive qualitative and quantitative experiments to demonstrate the capability of our approach. Compared with state-of-the-art methods, our approach achieves superior results with much lower complexity.

In summary, we present the following contributions:

- To our best knowledge, we are the first to solve the consistency problem of diffusion video editing by considering the concept of layered atlas approaches, which provides an efficient and effective way for this topic.
- We present a new video editing framework which can manipulate the appearance of the objects with high geometry and appearance consistency across time. Our method can be easily applied to other text-driven diffusion models.
- We conduct extensive experiments on a variety of natural videos, which shows superior editing performance compared with state-of-the-art methods.

2. Related Work

2.1. Diffusion for Image Editing

The editing of natural images is an important task in the field of computer vision that has been widely studied. Prior to the emergence of diffusion models [42, 15], many GAN-based approaches [11, 10, 30, 32, 48] have achieved good results. There are also some works focus on low-level editing [52, 6, 51]. The advent of diffusion models has made it possible to achieve even higher quality and more diverse edited contents. SDEdit [26] adds noise and corruptions to an input image and uses diffusion models to reverse the process for image editing, while suffering from the loss of fidelity. Prompt-to-Prompt [12] and Plug-and-Play [45] perform semantic editing by mixing activations from original and target text prompts. InstructPix2Pix [2] applies semantic editing at test time and personalizes the model through finetuning and optimization to learn a special token describing the content. UniTune [46] and Imagic [21] finetune on a single image for better editability while maintaining good fidelity. There are also some works exploring the controllability [55, 18, 39, 22, 4, 3] and personalization [9, 38, 29] of diffusion-based generation. Our proposed video editing method leverages existing image editing methods that can preserve the structures, such as [55, 31, 45] and [28].

2.2. Diffusion for Video Editing

Compared to image editing, video editing is more challenging for diffusion-based methods for geometric and temporal consistency. Tune-a-Video [49] inflates a text-to-image model for video editing. However, since temporal correlations are not fully considered, the editing results suffer from inconsistency of geometry and motion. Dreamix [27] develops a text-to-video backbone for motion editing while maintaining temporal consistency. There are also some works based on video generation like [8, 40, 13, 54, 56] and [17]. Unlike these approaches, our purpose is to enable diffusion models to perform appearance editing with both geometric and temporal consistency.

2.3. Temporal Propagation in Video Editing

Temporal propagation plays an important role in natural video process, since it is the essential factor for temporal consistency. Some methods rely on key frames [19, 44, 50] or optical flow [37] to propagate contents between frames.

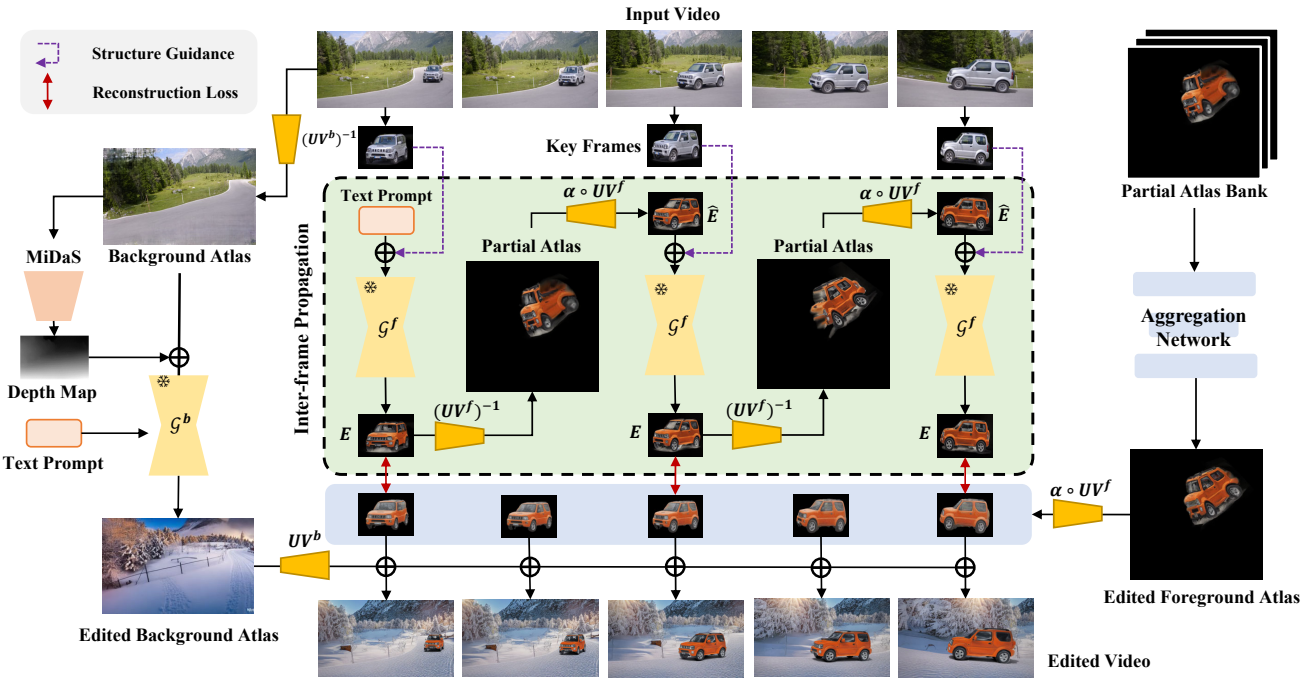

Figure 2: **Framework of the proposed StableVideo.** The input video is first fed into NLA [24] to generate foreground and background atlases using the pre-trained model. $\mathcal{G}_b$ is the diffusion model used to edit background atlas, and $\mathcal{G}_f$ is used to edit the foreground key frames. Note that $\mathcal{G}_b$ and $\mathcal{G}_f$ share the same weights, but accept different conditions. We employ depth information, extracted by MiDaS [35], for $\mathcal{G}_b$ to maintain the consistency between the foreground motion and the environment, while structure guidance is used for $\mathcal{G}_f$ to keep geometric consistency between the new generated foreground and the old one. After being edited, the foreground and background are blended together to reconstruct the edited frames.

Another bunch of methods are to achieve consistent inter-frame editing by forming a compressed representation of a video. Omnimattes [24, 25] estimate RGBA layers for target subject and scene effects for each frame independently, but cannot achieve consistent propagation of contents along temporal direction. Atlas [1, 20] tackles this problem by decomposing the video into unified 2D atlas layers for each target. This approach allows contents to be applied to the global summarized 2D atlases and mapped back to the video, achieving temporal consistency with minimal effort. Inspired by the concept of atlas approach, we employ the pre-trained neural layered atlas model to solve the inconsistency problem in diffusion video editing, thereby achieving high-quality editing results with temporal coherence.

3. Method

Fig. 2 shows the pipeline of our proposed StableVideo. We utilize NLA [20] as the propagator for consistent video editing. Specifically, we conduct foreground and background editing separately. For foreground editing, we adopt key frame editing to generate atlas layers with high quality and the inter-frame propagation module to ensure better geometric and temporal consistency. The edited key frames are then mapped to partial atlases and aggregated by the aggregation network to produce the edited foreground atlas. It is noteworthy that our approach can also handle more than one foreground layers.

3.1. Problem Formulation

We employ the pre-trained NLA model [24] to propagate the edited contents to ensure that the target objects and scenes can maintain homogeneous appearances and motions across the entire video. The concept of NLA is to decompose the input video into layered representations, namely foreground atlas and background atlas, which globally summarize the correlated pixels for the foreground and the background, respectively. Three mapping networks, *i.e.*, $\mathcal{M}^b(\cdot)$, $\mathcal{M}^f(\cdot)$ and $\mathcal{M}^\alpha(\cdot)$, are provided for this purpose. Given an input video I, for each frame I_i, we obtain the mapping relationships of the atlas in the background and the foreground with respect to the pixel coordinate system, named as $\mathrm{UV}^b(\cdot)$ and $\mathrm{UV}^f(\cdot)$, as well as the foreground opacity α_i on the pixel coordinate system, formulated as:

$$\mathrm{UV}_i^b(\cdot) = \mathcal{M}^b(I_i), \mathrm{UV}_i^f(\cdot) = \mathcal{M}^f(I_i), \alpha_i = \mathcal{M}^\alpha(I_i). \tag{1}$$

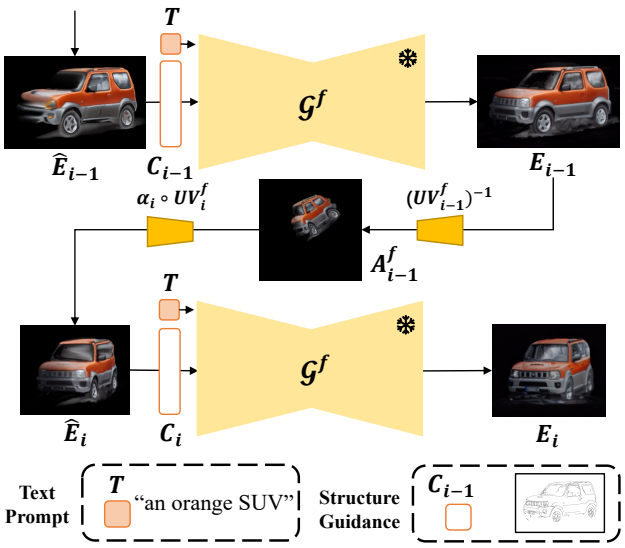

Figure 3: **Inter-frame propagation for foreground editing.** We use two edited key frames, E_{i-1} and E_i, to illustrate the process more clearly. The structure guidance and the text prompt is added into the denoising UNet via the concatenation and cross-attention mechanism respectively.

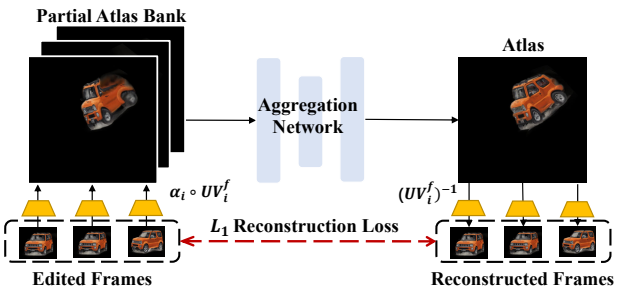

Figure 4: **Training process of aggregation network.** We employ a simple 3D network to aggregate the partial atlases generated from the edited key frames. For each training iteration, the sampled key frames are edited and mapped to partial atlases by UV_i^f. The partial atlases are then fed into the Aggregation Network to generate the edit atlas, which is mapped back by $(UV_i^f)^{-1}$ to generate the reconstructed frames. A L_1 loss is used to guarantee the aggregation consistency between the edited key frames and the reconstructed frames.

After that, we formulate the mappings from the atlas representation of the background A^b and the foreground A^f, to the pixel coordinate systems of B_i and F_i:

$$B_i = \mathrm{UV}_i^b(A^b), F_i = \mathrm{UV}_i^f(A^f). \tag{2}$$

Our method achieves geometrical consistent editing by fixing the mappings of UV^b and UV^f, and generating the edited atlases of A^b and A^f. We adopt a pre-trained latent diffusion model [36] with guided conditions as our generator, namely $\mathcal{G}^b(\cdot)$ and $\mathcal{G}^f(\cdot)$. Note that we are not simply editing the foreground and the background atlases directly. We apply inter-frame propagation mechanism on the editing process of foreground atlas. More details are explained in Sec. 3.2. After that, the entire video I can be reconstructed frame by frame as the following equation:

$$I_i = \alpha_i \circ \mathrm{UV}_i^f(\mathcal{G}^f(A^f)) + (1-\alpha_i) \circ \mathrm{UV}_i^b(\mathcal{G}^b(A^b)), \tag{3}$$

where $\circ$ denotes pixel-wise product.

3.2. Inter-frame Propagation

In this section, we further elaborate on how inter-frame propagation mechanism helps consistent foreground editing. One of the major challenges for diffusion models is to generate video contents with temporal consistency. Existing state-of-the-art text-driven diffusion methods [55, 28] can maintain the similar geometry between the target objects and the generated ones for image editing by adding structure conditions. However, the situation is different for videos. Generating temporally consistent geometry needs to handle some uncertain changes across time, *e.g.*, motion and deformation, which can not be supported by them. We tackle this problem by introducing a conditional denoising process to enable the diffusion models to consider both the structure of the current frame and the appearance information from previous frame, thereby sequentially generating new objects with geometric consistency across time. In specific, we employ canny edge as structure guidance, which is also adopted by existing diffusion methods [55]. Another important question is how to propagate the information of one object across frames to achieve consistent appearance. With the help of NLA, we can transfer the appearance features of the overlapping parts of previous frame to the next frame. Inspired by SDEdit [26] and ILVR [7], we further use a process of adding noise and denoising to obtain a more complete output. We illustrate this generation process in Fig. 3. The inter-frame propagation method is only applied on the foreground objects.

Specifically, we first select N foreground key frames from the original video I, ensuring that 1) there is significant overlap between the adjacent frames, and 2) these key frames capture the appearance of all faces of the object. Given a generator $\mathcal{G}^f(\cdot)$ and a text prompt T, we edit the first frame F_0 in pixel coordinate system with its structure condition C_0 as the extra guidance:

$$E_0 = \mathcal{G}^f(T, C_0), \tag{4}$$

where E_0 represents the editing result. Then for the remaining key frames, we propagate the editing result from the previous key frame E_{i-1} to obtain the one of the current key frame E_i. To be specific, we map E_{i-1} from the pixel

Figure 5: **Compositing editing.** We demonstrate two editing examples with non-rigid and rigid foreground objects. Our approach can well preserve the geometry of "bear" and "car" across frames.

coordinate to atlas A^E_{i-1} and then map it back to the pixel coordinate in the current frame i as $\hat{E}_i$ with multiplying the opacity α_i, formulated by:

$$\begin{aligned} A^f_{i-1} &= (\mathrm{UV}^f_{i-1})^{-1}(E_{i-1}), \\ \hat{E}_i &= \alpha_i \circ \mathrm{UV}^f_i(A^f_{i-1}). \end{aligned} \tag{5}$$

It is noteworthy that the entire video shares the same foreground atlas A^f for each target object, where the subscript i represents an incomplete partial atlas in frame i. We then use the atlas to propagate the pixel values from previous frame to their corresponding positions in the current frame to obtain an incomplete partial appearance $\hat{E}_i$.

Given the partial appearance $\hat{E}_i$, we first encode it with VQ-VAE [47] to get the latent representation $\hat{Z}_i$, and then add noise to it with Variance Preserving Stochastic Differential Equation (VP-SDE), formulated as:

$$\hat{Z}_i(t_0) = \alpha(t_0)\hat{Z}_t(0) + \sigma(t_0)\mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}) \tag{6}$$

where $\sigma(t_0)$ and $\alpha(t_0)$ are two scalar functions that satisfy $\alpha^2(t) + \sigma^2(t) = 1, \forall t \in (0, 1]$, and $t_0 \in [0, 1]$ is a hyper parameter of the noise strength. Then we apply denoising process $\hat{Z}_i(t_0)$ under the condition guidance from both text prompt T and structure guidance C_i to get the latent representation Z_i. Finally, we decode the latent representation to Z_i propagate editing result E_i. Our experiments further demonstrate that this mechanism can achieve good propagation results without training or fine-tuning the model.

3.3. Aggregation Network

Different from [1] and [24], our approach edits video frames rather than atlases, which have the chance to achieve more information of different viewpoints. This brings two advantages. Firstly, the geometries and pixels from different viewpoints provide more details of the target objects, allowing the diffusion model to generate the edited content with higher fidelity. Secondly, this alleviates the risk of failure editing due to the potential wrong mapping from the atlas to the video frames. We then aggregate the edited key frames by using a simple yet effective two-layer 2D convolution network with skip connection as shown in Fig. 4. Our goal is to guarantee that the aggregated atlas is highly aligned with the original one, in terms of locations, so that appearance edit will not affect the geometric consistency and the temporal continuity. Reconstruction loss, $\mathcal{L}_{rec}$, between the edited and reconstructed key frames is employed in the training process as:

$$\mathcal{L}_{rec} = \sum_{i=1}^{N} ||E_i - \mathrm{UV}^f_i(A^f)||_1. \tag{7}$$

where N is the number of key frames.

4. Experiments

4.1. Experimental Settings

In practice, we implement our approach over Stable Diffusion [36]. Despite there are several image-based meth-

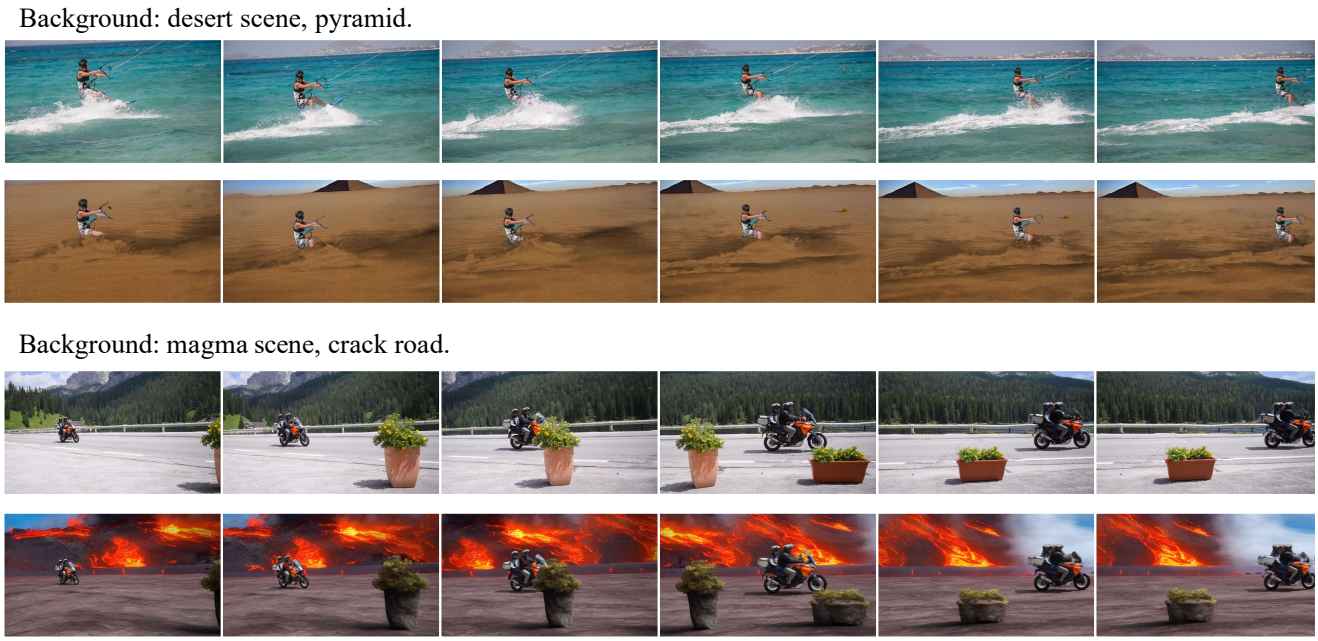

Figure 6: **Background replacement.** Since our approach can effectively maintain the geometry of the foreground, it can perform background replacement while maintaining geometric consistency of depth and temporal continuity of perspective.

ods [31, 45, 28] can perform structure-preserving editing, we choose the canny condition branch from [55] as the structure guidance for the proposed inter-frame propagation in our method. We apply our method on several videos from DAVIS [33], with each video containing a *moving* object in 50 $\sim$ 70 frames. The image resolution is set to 768 $\times$ 432, and the resolution of foreground atlas is set to 2000 $\times$ 2000. We employ DDIM [43] sampler with 20 steps. In this case, our method requires only $\sim$10 GB GPU memory and takes $\sim$30 seconds for each video in a single NVIDIA A40 GPU.

4.2. Editing Results

We tested our method for various editing types among several videos. Here we demonstrate several scenarios:

Compositing editing. Our method can edit foreground and background separately to achieve high-quality and semantically matched editing results as shown in Fig. 5.

Background replacement. Due to the separate editing capabilities enabled by NLA, as well as the diverse and high-quality editing brought by the diffusion model, our method achieves high-quality video background replacement while maintaining geometric consistency of depth and temporal continuity of perspective, as shown in Fig. 6.

Style transfer. Our method accomplishes a wide range of style transfers, while simultaneously ensuring temporal consistency, as shown in Fig. 7.

4.3. Comparison to Prior Arts

In this section, we compare our editing results with state-of-the-art atlas-based method Text2LIVE [1] and diffusion-based method Tune-A-Video [49].

Comparison to Text2LIVE [1]. As shown in Fig. 8, given the text prompt *"an orange SUV"*, Text2LIVE shows incomplete editing, while our method demonstrates a much more holistic result. It is because our method employs key frame editing, while Text2LIVE creates a new layer to edit the atlas somehow directly. The potential failure editing in atlas will lead to the error propagation to the entire video. In addition, compared to Text2LIVE, we can achieve higher quality and richer content with faster inference speed.

Comparison to Tune-A-Video [49]. As shown in Fig. 9, vanilla video diffusion models like Tune-A-Video often fail in video editing. While it adeptly captures the semantic information of text prompts, it struggles to preserve consistency in video layout and object geometry.

Consistency analysis. To the best of our knowledge, there is currently no widely accepted metric for evaluating the geometric and temporal consistency of videos. In this paper, we employ motion consistency of dense optical flow and deviation consistency of the edited video frames for this metric. For motion consistency, we employ the Farneback algorithm in OpenCV to calculate the average L2 distance of dense optical flow between the edited and original videos. We select the "car-turn" video from DAVIS and apply the same prompt to all methods. The experiment is repeated several times to obtain the average number. Our method shows better stability than Tune-A-Video as shown in Tab. 1. For deviation consistency, we conduct experiments on 1) CLIP score: target text faithfulness 2) LPIPS-P: deviation from the original video frames and 3) LPIPS-T:

Figure 7: **Style transfer.** Our approach achieves diverse video style transfer while ensuring high temporal consistency.

Figure 8: **Comparison to Text2LIVE.** Foreground prompt: an orange SUV. Our method achieves more holistic editing results on the foreground.

deviation between adjacent frames, as shown in Tab. 2. Our method achieves comparable CLIP score and much lower deviations, which shows the effectiveness and stability.

4.4. Ablation Study

To verify the necessity of the key frame editing, we apply editing in atlas layer directly for the foreground as a simple baseline. The atlas might not be so deformed for human perception, but it significantly affects the diffusion models. Fig. 10 shows an example, where obvious deformation and inconsistency exist.

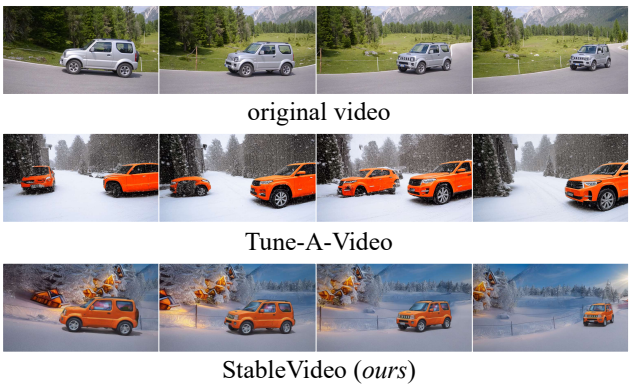

Figure 9: **Comparison to Tune-A-Video.** Prompt: an orange SUV in sunny snowy winter, cabins. Our method achieves much more consistent editing results.

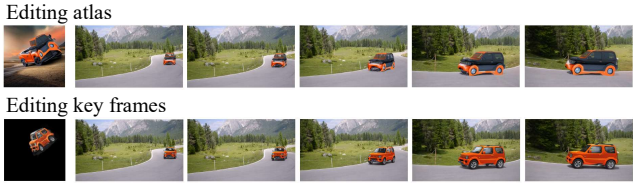

Figure 10: **Ablation study on directly editing the atlas.** The deformation in atlas affects the diffusion models.

We also conduct extensive ablation study on inter-frame propagation module. The objective of this module is to

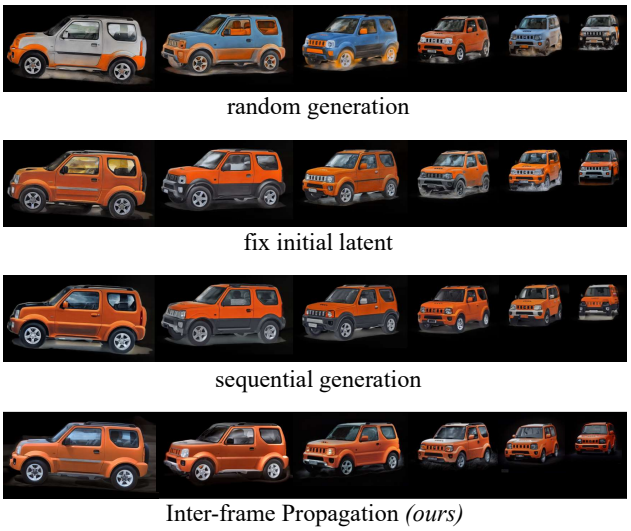

Figure 11: **Ablation study on inter-frame propagation module.** Foreground prompt: an orange SUV. Our method achieves excellent consistency in key frame appearance.

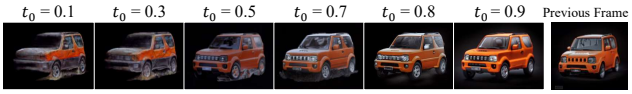

Figure 12: **Binary search on hyper parameter t_0 in inter-frame propagation module.** The appearance of previous frame is shown on the right side. The sequence on the left shows the generated results of the current frame as a function of t_0. The value of t_0 around 0.8 is a good sweet spot.

maintain the geometry of the foreground when editing key frames. Firstly, we consider four different settings for editing key frames as shown in Fig. 11.

Random generation. Each key frame only shares the same text prompt with the others. In this case, there are significant differences among the generated key frames.

Fix initial latent. Unlike starting from random noise every time we edit, we start generating each key frame from the same latent noise and share the text prompt. In this case, there is higher similarity in the content generated for each frame, but the consistency is still not satisfactory.

Sequential generation. Furthermore, we concatenate the latent noise between frames. Specifically, we apply image-to-image translation between frames. This method still cannot guarantee consistency since the appearances of the objects between the two frames do not match.

Inter-frame propagation (ours). Our final approach is to employ partial atlas to geometrically align the appearances between two frames, followed by a process of adding noise and then apply denoising process.

In addition, We also conducted experiments on the selection of the hyper parameter t_0 in this module as shown in Fig. 12. We observe that as t_0 gradually increases, the generated results become more realistic but gradually lose their match with the appearance of the previous frame. Using binary search, we find out that a reasonable trade-off between fidelity and realism for t_0 lies around 0.8.

Method	Foreground Editing	Composite Editing
Text2LIVE [1]	1.99	7.39
Tune-A-Video [49]	4.63	12.74
StableVideo (*ours*)	3.34	11.48

Table 1: **Consistency of dense optical flow.** Original video: car-turn. Foreground prompt (column 1): an orange suv. Composite prompt (column 2): a car driving in winter.

Method	CLIP (↑)	LPIPS-P (↓)	LPIPS-T (↓)
Tune-A-Video [49]	0.2787	0.6346	0.1851
StableVideo (*ours*)	0.2713	0.1613	0.0386

Table 2: **Quantitative results.** Original video: car-turn. Prompt: an orange suv in the winter.

Figure 13: **Failure case.** Videos with non-rigid deformation may lead to failure editing, since the movement of the object is more difficult to be well captured in this case.

5. Limitations and Future Works

Firstly, our method is constrained by NLA. Learning atlas layers may fail for non-rigid objects with significant structural deformation as shown in Fig. 13. While we can mitigate this by dividing long videos into short clips where the objects can be considered to be rigid, it is still not feasible to address every single case. Secondly, our method is constrained by the capabilities of the diffusion models, which may struggle with specific scenarios such as human or animals. Besides, it may be better to optimize the diffusion model with the objective of aligning the generated contents to the reconstructed ones.

6. Conclusion

We have proposed a text-driven diffusion video editing approach. To solve the consistency problem for diffusion models in foreground object editing, we propose a inter-frame propagation mechanism and an atlas aggregation network. We conducted extensive experiments and demonstrated the superior qualitative and quantitative results of our method compared to state-of-the-art approaches.

References

[1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 707–723. Springer, 2022. 2, 3, 5, 6, 8, 12

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 1, 2

[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] Shidong Cao, Wenhao Chai, Shengyu Hao, Yanting Zhang, Hangyue Chen, and Gaoang Wang. Difffashion: Reference-based fashion design with structure-aware transfer by diffusion models. *arXiv preprint arXiv:2302.06826*, 2023. 2

[5] Wenhao Chai and Gaoang Wang. Deep vision multimodal learning: Methodology, benchmark, and trend. *Applied Sciences*, 12(13):6588, 2022. 1

[6] Sixiang Chen, Tian Ye, Yun Liu, Erkang Chen, Jun Shi, and Jingchun Zhou. Snowformer: Scale-aware transformer via context interaction for single image desnowing. *arXiv preprint arXiv:2208.09703*, 2022. 2

[7] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 4

[8] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 2

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[10] 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

[11] 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

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 2

[13] 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. 2

[14] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:210.02303*, 2022. 1

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[16] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 1

[17] Jonathan Ho, Tim Salimans, Alexey A Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*. 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 Sỳkora. Stylizing video by example. *ACM Transactions on Graphics (TOG)*, 38(4):1–11, 2019. 2

[20] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 3

[21] 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. *arXiv preprint arXiv:2210.09276*, 2022. 2

[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. *arXiv preprint arXiv:2301.07093*, 2023. 2

[23] Weng Fei Low and Gim Hee Lee. Minimal neural atlas: Parameterizing complex surfaces with minimal charts and distortion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 465–481. Springer, 2022. 1

[24] Erika Lu, Forrester Cole, Tali Dekel, Weidi Xie, Andrew Zisserman, David Salesin, William T Freeman, and Michael Rubinstein. Layered neural rendering for retiming people in video. *arXiv preprint arXiv:2009.07833*, 2020. 1, 3, 5, 12

[25] Erika Lu, Forrester Cole, Weidi Xie, Tali Dekel, William T Freeman, Andrew Zisserman, and Michael Rubinstein. Associating objects and their effects in video through coordination games. In *Advances in Neural Information Processing Systems*. 3

[26] 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, 4

[27] 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

[28] 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, 4, 6

[29] Yaniv Nikankin, Niv Haim, and Michal Irani. Sinfusion: Training diffusion models on a single image or video. *arXiv preprint arXiv:2211.11743*, 2022. 2

[30] 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

[31] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023. 1, 2, 6

[32] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2

[33] 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

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[35] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 3

[36] 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. 4, 5, 12

[37] 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. 2

[38] 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. *arXiv preprint arXiv:2208.12242*, 2022. 2

[39] 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

[40] 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. *arXiv preprint arXiv:2209.14792*, 2022. 2

[41] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*. 2

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6

[44] Ondřej Texler, David Futschik, Michal Kučera, Ondřej Jamriška, Šárka Sochorová, Menclei Chai, Sergey Tulyakov, and Daniel Sỳkora. Interactive video stylization using few-shot patch-based training. *ACM Transactions on Graphics (TOG)*, 39(4):73–1, 2020. 2

[45] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. 2, 6

[46] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2022. 2

[47] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 5

[48] 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

[49] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, 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. *arXiv preprint arXiv:2212.11565*, 2022. 1, 2, 6, 8, 12

[50] Yiran Xu, Badour AlBahar, and Jia-Bin Huang. Temporally consistent semantic video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 357–374. Springer, 2022. 2

[51] Ren Yang, Radu Timofte, Xin Li, Qi Zhang, Lin Zhang, Fanglong Liu, Dongliang He, Fu Li, He Zheng, Weihang Yuan, et al. Aim 2022 challenge on super-resolution of compressed image and video: Dataset, methods and results. In *European Conference on Computer Vision*, pages 174–202. Springer, 2022. 2

[52] Tian Ye, Yunchen Zhang, Mingchao Jiang, Liang Chen, Yun Liu, Sixiang Chen, and Erkang Chen. Perceiving and modeling density for image dehazing. In *European Conference on Computer Vision*, pages 130–145. Springer, 2022. 2

[53] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. Magvit:

Masked generative video transformer. *arXiv preprint arXiv:2212.05199*, 2022. 1

[54] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. *arXiv preprint arXiv:2302.07685*, 2023. 2

[55] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2, 4, 6, 12

[56] 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

Supplement Material

A. Implementation Details

In our experiments, we choose key frames for foreground editing by evenly sampling the input frames, *i.e.*, every 20 frames. We train the aggregation network for 500 epochs with initial learning rate of 0.003 and momentum of 0.9. The network consists of two convolution layers with a ReLU in between, for which the training process is very fast. At inference stage, we conduct the training once for each edit. We set the lower and upper thresholds of Canny edges as 100 and 200 respectively, which can make the edges better represent the structure of the foreground. The numbers in Tab. 1 are the optical flow differences between the videos before and after editing (lower is better). We use *cv2.calcOpticalFlowFarneback* with default parameters. More detailed setting could be found in our code that will be released soon.

B. Failure Cases

Since our approach edits the key frames by using existing pre-trained diffusion models, some failure cases will occur due to the ineffective diffusion control. For example, our inter-frame propagation can well preserve the structure of the target objects across time, but cannot guarantee the quality of partial editing, as shown in Fig. A. This problem could be handled by using the masks provided by the users in practical applications, which would be our future work. As we discussed in the manuscript, NLA [24] may fail to build the foreground atlas due to the complex motion or occlusion. In this case, our editing will also fail. However, since our approach edits directly on key frames and generates corresponding partial atlases, such failure can be alleviated.

C. Complexity Analysis

Since inference is also an essential factor for video editing, we provide the comparison of our approach to existing state-of-the-art methods, *i.e.*, Tune-A-Video [49] and Text2LIVE [1] as shown in Tab. A. Our approach only needs to perform lightweight training for atlas aggregation at inference stage, thereby being more efficient in pratical application compared to Text2LIVE and Tune-A-Video.

Figure A: An example of failure editing. Our method generates the edited contents by leveraging existing diffusion models [55, 36]. In the case of partial editing, *e.g.*, changing the color of the skirt, the diffusion models may generate the whole person instead.

Method	Video Training	Edit Training	Edit Inference
Text2LIVE [1]	$\sim$ 10 hr	$\sim$ 1 hours	$\sim$ 10 sec
Tune-A-Video [49]	$\sim$ -	30 min	$\sim$ 4 min
StableVideo (*ours*)	$\sim$ 10 hr	-	$\sim$ 30 sec

Table A: The inference speed of three methods. Video Training: training once for each video. Edit Training: training once for each edit. Edit Inference: inference time. The approximated cost time is tested under the video with 768 $\times$ 432 resolution and 70 frames in a single NVIDIA A40. For StableVideo, we pick three key frames for foreground editing.

D. More Editing Results

We provide more editing results to demonstrate the effectiveness of our approach. Fig. B shows the foreground editing for the video of "boat". We can see that the temporal consistency is well preserved. Fig. C shows the composite edit of our approach. Since the foreground and background are generated by the same diffusion model, they are highly semantically consistent. Besides, the geometry is also be well preserved across time.

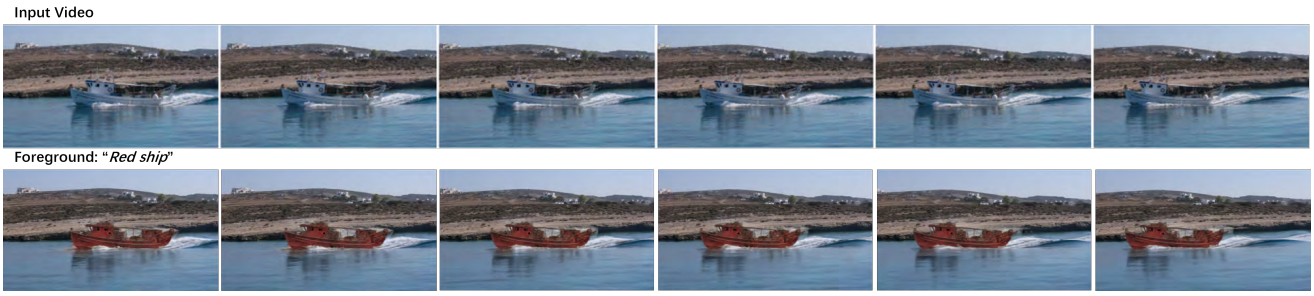

Figure B: The editing results of foreground. The ship in this video has relatively complex geometry. Our approach can well preserve the temporal consistency.

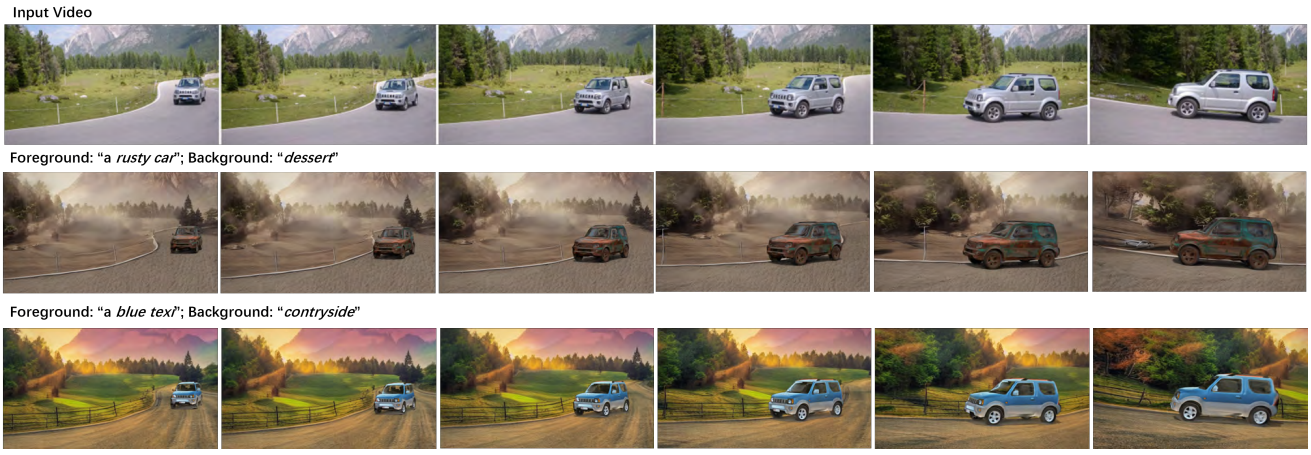

Figure C: The results of composite editing. We separately edit the foreground and the background with semantically correlated prompts.