# Supplementary Material: 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


## 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 [2] 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.

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

| Method | Video Training | Edit Training | Edit Inference |
|---|---|---|---|
| Text2LIVE [1] | $\sim$ 10 hr | $\sim$ 1 hours | $\sim$ 10 sec |
| Tune-A-Video [4] | $\sim$ - | 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.

### 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 [4] 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 [5, 3]. In the case of partial editing, *e.g.*, changing the color of the skirt, the diffusion models may generate the whole person instead.

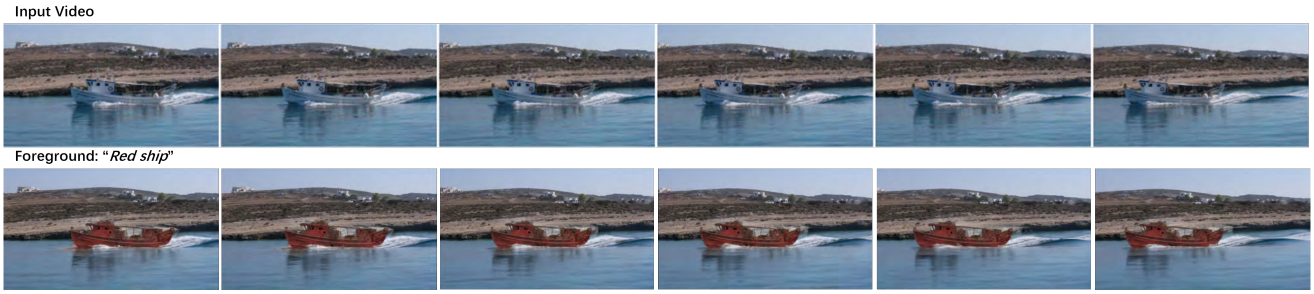


Figure B: The editing results of foreground. The ship in this video has relatively complex geometry. Our approach can well preserve the temporal consistency.

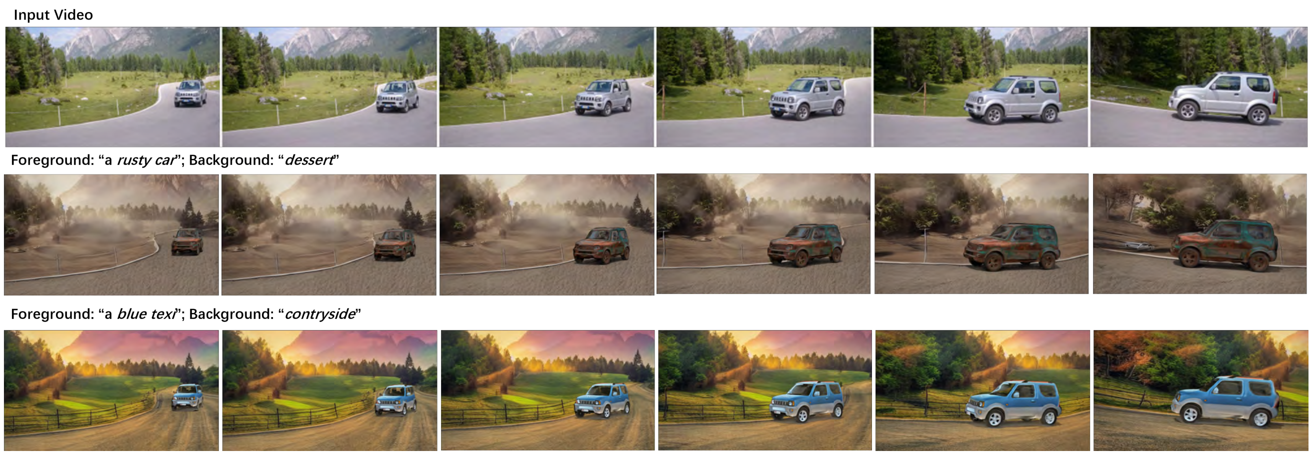


Figure C: The results of composite editing. We separately edit the foreground and the background with semantically correlated prompts.

## 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.

## 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. 1

[2] 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] 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

[4] 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

[5] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1