Supplementary Materials for "Ada-VE: Training-Free Consistent Video Editing Using Adaptive Motion Prior"

Tanvir Mahmud, Mustafa Munir, Radu Marculescu, and Diana Marculescu
The University of Texas at Austin

A. Limitations: Qualitative visualizations

We note that our method primarily built on top of the PnP [1] image editing framework. PnP leverages structural guidance from the given video through extensive feature injection, which often fails to significantly alter the shape of subjects. We present a sample visualization of such failure cases in Fig. 1. It can be seen that the error mostly occurs when attempting to modify a Jeep car into a Porsche car due to their significant shape differences. These issues are largely inherited from the baseline PnP image editing method. Despite this, Ada-VE maintains consistent character across all frames, demonstrating its robustness in extending image models for video editing applications. Additionally, Ada-VE can be easily adapted to any image editing method due to its simple and general architecture.

B. Various extensions of Self-Attention: Qualitative visualizations

We present sample qualitative visualizations of various self-attention extension mechanisms in Fig. 2. We observe consistent character generation when using KVs from a fixed set of frames across the videos. For instance, even when using only the first frame's KVs, we observe consistent character generation, but the visual quality deteriorates significantly, resulting in structural deformations. Integrating KVs from the immediate previous frames along with the first frame shows some improvement in structural deformations. However, using different KVs across the video leads to significant flickering and inconsistent character generation. For example, the woman's face is altered, and the background grass is no longer visible.

Using fully extended self-attention significantly improves visual quality and character consistency but increases latency by approximately 13 times due to the computational burden. We observe repeated features across frames, such as the background scenarios of the running woman, which consume a large portion of redundant computation. By integrating Ada-VE, we leverage the motion prior of the guidance video to drastically reduce computational overhead. Ada-VE achieves around a $4\times$ speed-up for joint editing across these frames. Therefore, Ada-VE can potentially integrate a significantly larger number of frames in joint editing while using similar computational resources.

C. Extensive Qualitative Visualizations

The supplementary HTML page includes extensive qualitative visualizations on challenging examples, comparing Ada-VE with six state-of-the-art baseline methods. We also present results on ablation studies highlighting different self-attention extension mechanisms.

References

[1] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 1, 2

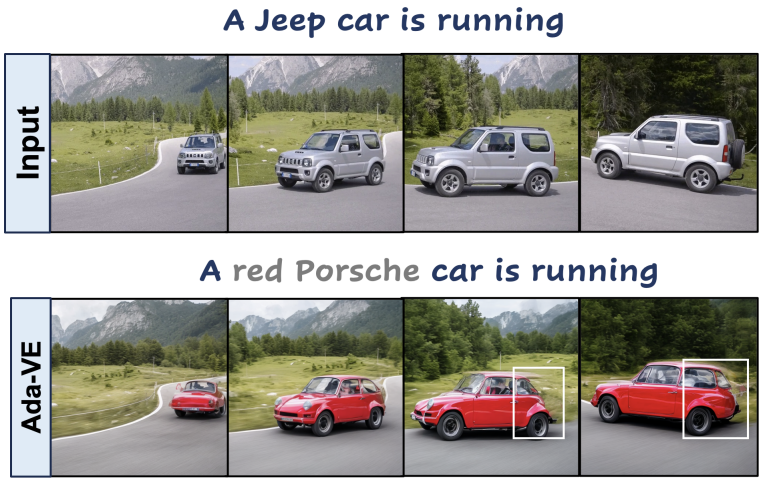

Figure 1. Visualizing failure cases: Since our method is built on top of Plug-and-Play [1] diffusion, it cannot inherently follow the modified prompts to change the structure of the subject. Nevertheless, our key contributions are easy to adapt in most existing video editing baselines.

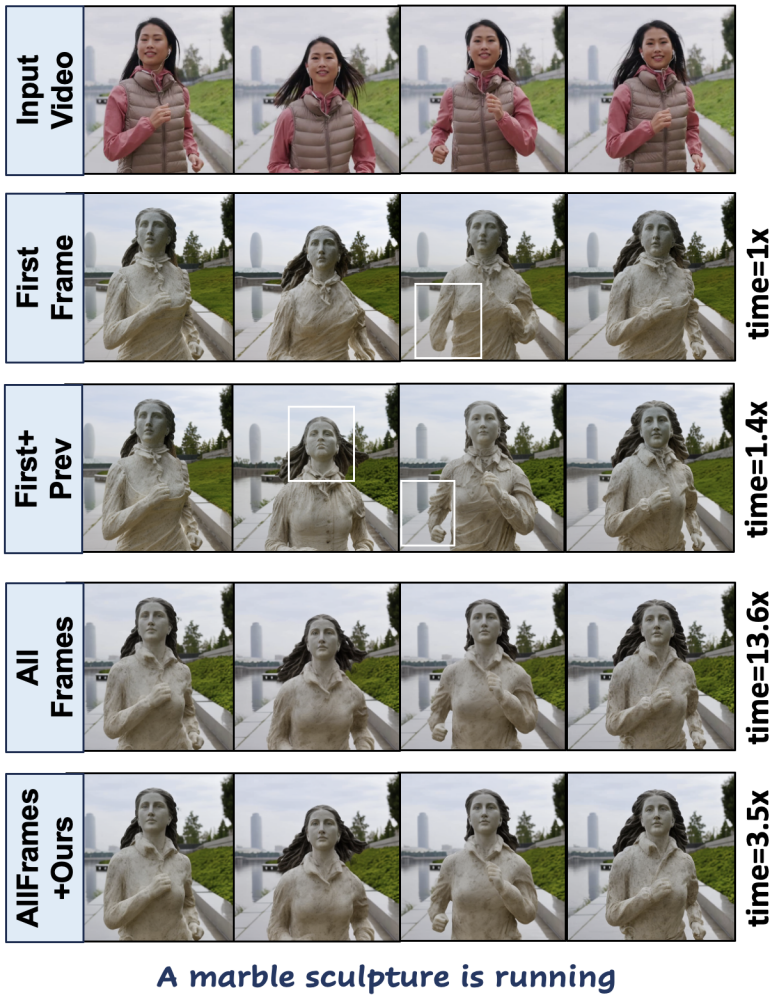

Figure 2. Qualitative visualizations of various self-attention extensions: Increasing the number of frames in the extended self-attention significantly improves performance but comes with an extensive computational burden. Ada-VE significantly reduces the operational latency of full extensions while preserving visual quality by leveraging the motion prior of the guidance video. This allows for a substantial increase in the number of frames in joint editing, enabling efficient operation on longer duration videos. For this study, joint editing of a total of 40 frames was used, and we present a portion of these.