

Supplementary Material for VIDEOSHOP: Localized Semantic Video Editing with Noise-Extrapolated Diffusion Inversion

Xiang Fan¹, Anand Bhattad^{*2}, and Ranjay Krishna^{*1}

¹ University of Washington

² Toyota Technological Institute at Chicago

<https://videoshop-editing.github.io/>

1 Implementation Details

Image-to-video model. We use Stable Video Diffusion [1] as our base model, which supports generating 14 frames of video from a single image at a resolution of 576×1024 (height \times width). The model takes an image as a conditioning signal and denoises a latent space noise into a video with the conditioning image as the video’s first frame.

Diffusion inversion with noise extrapolation. We summarize our algorithm for diffusion inversion in Algorithm 1.

Algorithm 1 Algorithm for diffusion inversion with noise extrapolation.

Require: F_θ : an image-to-video model. E : a VAE encoder. D : the corresponding VAE encoder. \mathcal{V} : input video. \mathcal{I} : edited first frame.

```
 $x_{\text{in}} \leftarrow E(\mathcal{V}).$  ▷ Encode the input video.  
 $\hat{x}_0 \leftarrow \frac{x_{\text{in}}}{\sigma_{\text{in}}}.$  ▷ Normalize the input video.  
for  $i$  in  $0..(\text{diffusion steps} - 1)$  do  
    Invert  $x_i$  into  $x_{i+1}$  using Eq. (7)  
end for  
for  $i$  in  $(\text{diffusion steps} - 1)..0$  do  
    Denoise from  $x_{i+1}$  to  $x_i$  using Eq. (3), conditioning the model on  $\mathcal{I}$   
end for  
 $x_{\text{out}} \leftarrow \frac{\sigma_{\text{img}}}{\sigma_0} (x_0 - \mu_0 + \mu_{\text{img}}).$  ▷ Rescale the output latents.  
 $\mathcal{J} \leftarrow D(x_{\text{out}}).$  ▷ Decode to the output video.  
return  $\mathcal{J}.$ 
```

Hyperparameters. We run the Stable Video Diffusion [1] with 13 denoising steps, fps=7, frames=14, guidance scale=4, and motion bucket id=127.

* equal advising

2 Video Sample and Additional Results

We provide video samples of our method in the Supplementary Material. Please refer to `./supplementary.html` for these results.

3 Human Evaluation

Task details. We ask human evaluators to rate the quality of the edited videos generated by our method and the baseline methods. We use a A/B testing framework to compare the quality of the edited videos. We debrief human evaluators on this task and show the evaluators a pair of videos, one generated by our method and the other generated by a baseline method, and ask them to select the video that they think is of higher quality, or choose that both videos are of equal quality. We give evaluators questions for both edit quality and video quality and aggregate the results. Our evaluation interface is shown in Fig. 1. The order of output is randomized for each input video to avoid bias.

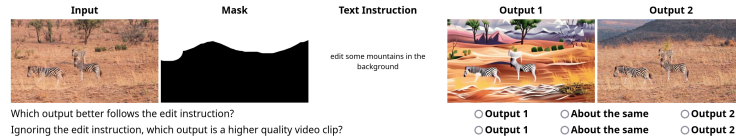


Fig. 1: Example of our human evaluation interface.

References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets (2023)