MaskINT: Video Editing via Interpolative Non-autoregressive Masked Transformers

Supplementary Material

1. Dataset Details

ShutterStock [1] is a commercial dataset with over 30 million paired text-video examples. Our model is trained using a subset of 100k videos. The WebVid dataset [2] offers an accessible version of ShutterStock, but they include a central watermark due to copyright constraints.

2. Diverse Structural Conditions

Although we by default utilize HED edge map as our structure condition for both stages, our method can also employ other structural controls in both stages due to the disentanglement. In this study, we explore utilizing ControlNet with depth map to perform key frame editing, and using the depth map as the guidance to perform structure-aware frame interpolation. Additionally, we explore various combinations of these approaches. As summarized in Table 1, all of these combinations achieve the same level of performance. The performance of prompt consistency (P.C.) is determined by the specific key frame editing methods employed. For the second stage, depth control typically offers greater flexibility than HED edge control for frame interpolation. This could be the potential reason for the slightly worse performance in temporal consistency with edge-based key frame editing.

Stage1	Stage2	T.C.	P.C
HED edge	HED edge	0.9714	0.3038
depth map	HED edge	0.9713	0.3159
HED edge	depth map	0.9683	0.3035
depth map	depth map	0.9719	0.3171

Table 1. Quantitative comparisons of the combination of varied structural conditions in each stage on ShutterStock. "T.C." stands for "temporal consistency", and "P.C." stands for "prompt consistency".

3. Additional Examples for Comparisons

We add more comparisons with diffusion methods in Fig. 2. Our methods can maintain the temporal consistency among variant examples like other pure-diffusion methods.

4. Additional Editing Examples

We present more video editing examples in Figure 3 to show the generalization of our method.

5. Example of Failure Cases

Since we disentangle the video editing tasks into two separate stage, the final performance of the generated video depends on the key frame editing in the first stage. In certain challenging scenarios, the attention-based key frame editing stage struggles to produce consistent frames, primarily due to the complexity of the scene or the presence of exceptionally large motion. In this case, our MaskINT can still interpolate the intermediate frames, albeit with the potential for introducing artifacts. Figure 1 show some failure cases when the first stage fails.

Figure 1. Examples of failure cases.

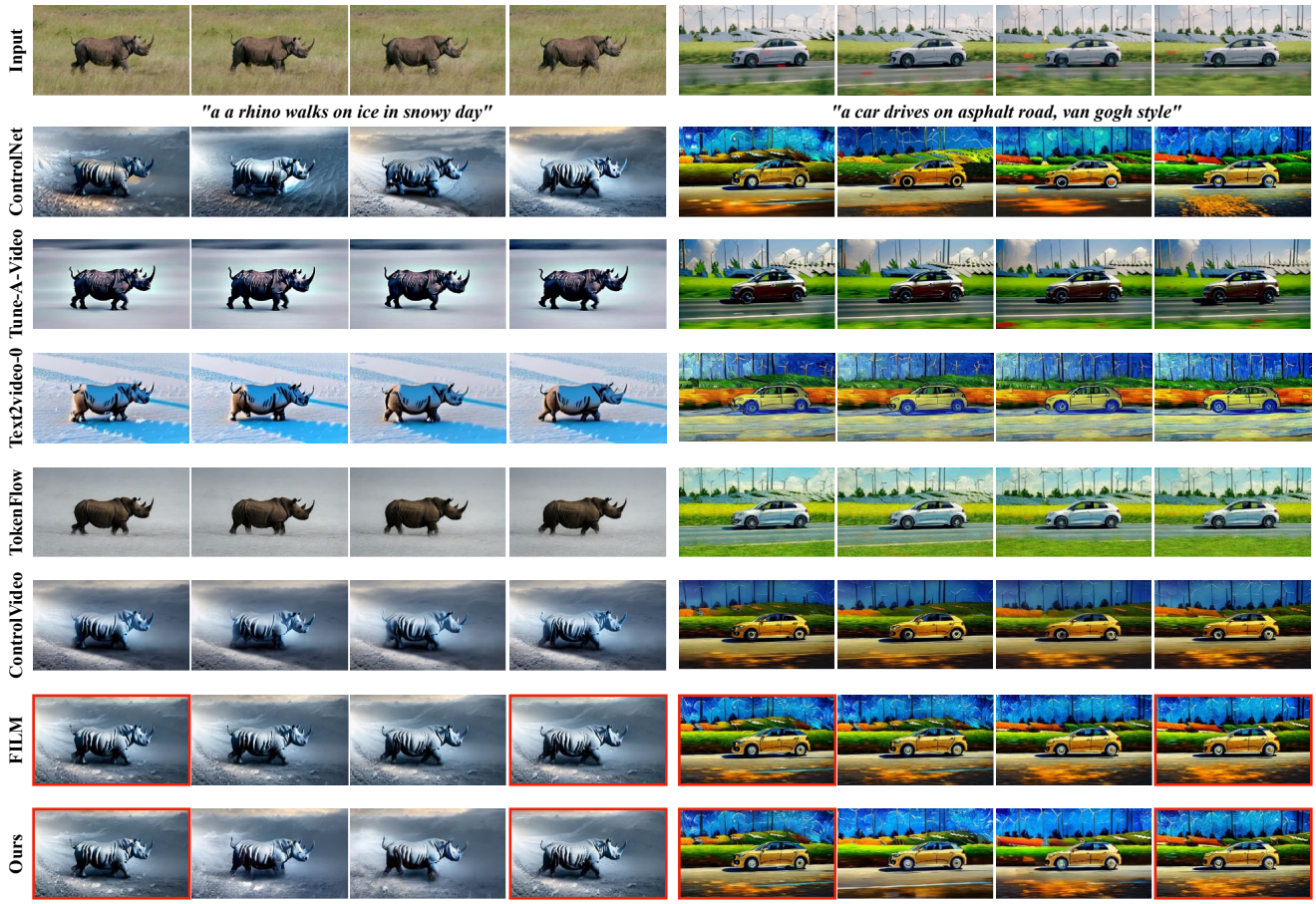

Figure 2. Additional Qualitative comparisons with diffusion-based methods. Frames with red bounding box are jointly edited keyeframes.

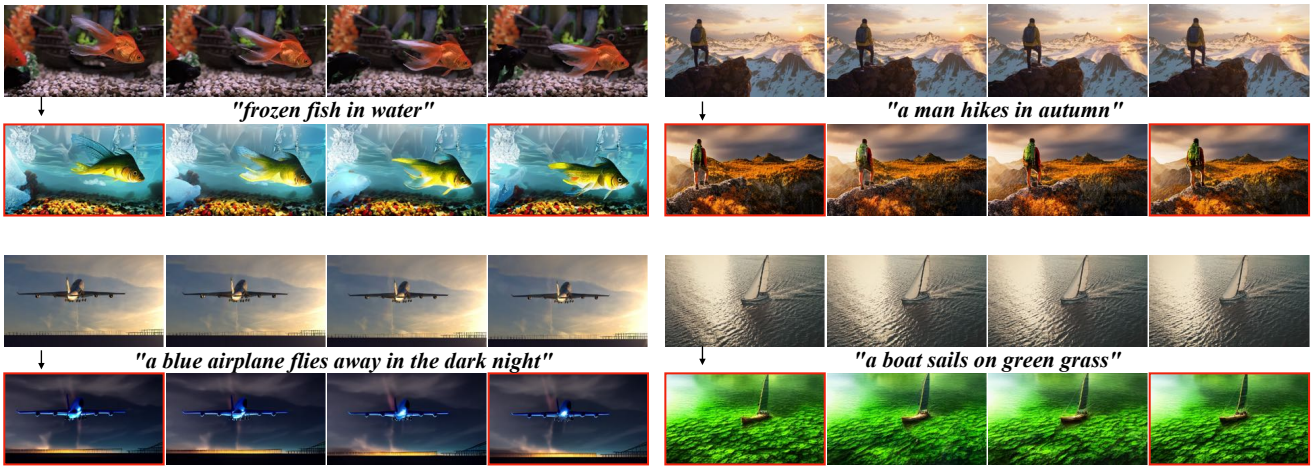

Figure 3. Additional Editing examples with MaskINT. Frames with red bounding box are jointly edited keyeframes.