Video-P2P: Video Editing with Cross-attention Control

Supplementary Material

6. Implementation Details

6.1. Attention Control

In this paper, attention maps are calculated as:

$$M = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right), \tag{12}$$

where Q and K are the query matrix and key matrix, and d is the latent projection dimension. To complete prompt refinement, following [13], we only apply attention injection over the shared tokens in both source and target prompt:

$$\left(\text{Edit}\left(M_t, M_t^*, t\right)\right)_{i,j} := \begin{cases} \left(M_t^*\right)_{i,j} & \text{if } A(j) = None \\ \left(M_t\right)_{i,A(j)} & \text{otherwise,} \end{cases} \tag{13}$$

where A is an alignment function that inputs a token index from the target prompt $\mathcal{P}^*$ and outputs the corresponding index in $\mathcal{P}$. It returns $None$ when the token cannot be found in the source prompt. To achieve attention re-weighting, we scale attention maps of a specific token with a parameter c:

$$\left(\text{Edit}\left(M_t, M_t^*, t\right)\right)_{i,j} := \begin{cases} c \cdot \left(M_t\right)_{i,j} & \text{if } j = j^* \\ \left(M_t\right)_{i,j} & \text{otherwise,} \end{cases} \tag{14}$$

with which we can manipulate the extent of a specific token j^*. Notice i corresponds to a pixel value and j corresponds to a text token.

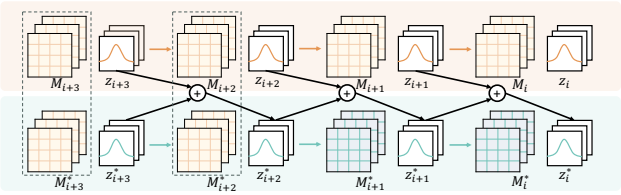

Figure 11. Relationship between two branches for "Word Swap".

6.2. Quantitative Evaluations

We comprehensively evaluate our proposed method with various carefully designed metrics, including CLIP Score, Perceptual Similarity, Masked PSNR, Object Semantic Variance, Temporal Consistency, and user study. In addition, we present a case-wise performance comparison in Fig. 12.

CLIP Score & LPIPS. We choose the official ViT-Base-Patch16 CLIP model and use the output logits as the CLIP Score output. In addition, we apply a standard VGG [36] extractor for LPIPS feature extraction.

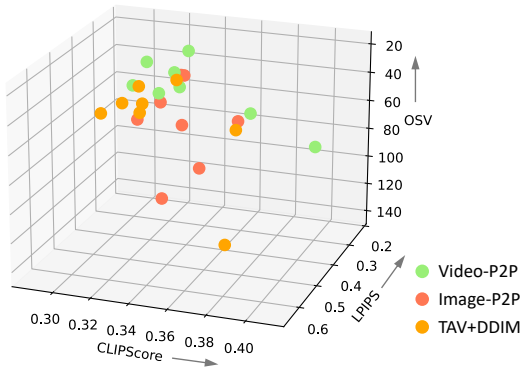

Figure 12. Evaluation metrics for each case. Compared with other editing methods, Video-P2P can perform more consistent and faithful editing with structure preservation.

Masked PSNR. To evaluate the preservation of local structures in the edited video, we propose Masked PSNR (M.PSNR), which complements LPIPS [51] by measuring the low-level pixel distance in unrelated regions. As depicted in Fig. 13, given the averaged attention mask sequence M of the changed object, we compute Masked PSNR by comparing the pixel distance in the unrelated regions of the edited video $\mathcal{V}^*$ and the input video $\mathcal{V}$,

$$\text{M.PSNR}(\mathcal{V}^*, \mathcal{V}) = \text{PSNR}(B(\mathcal{V}^*, M), B(\mathcal{V}, M)). \tag{15}$$

We define $B(\mathcal{V}, M) = \mathcal{V}_M$ as a reversed mask binary function with a threshold of 0.3, so only unrelated regions are considered in Masked PSNR calculation.

Object Semantic Variance. Evaluating the cross-frame content consistency is challenging when object structures are changed after editing. Inspired by recent works in 3D rendering [48], we introduce Object Semantic Variance (OSV), which measures the semantic consistency in a video sequence by calculating the frame-wise feature variance in the edited region. We use DINO-ViT [4] F_θ to extract feature maps as perceptual guidance of objects. As shown in Fig. 13, OSV is computed as

$$\text{OSV}(\mathcal{V}^*, M) = \sum^{c} \text{Var}(\sum^{h,w}(M \cdot F_\theta(\mathcal{V}^*))), \tag{16}$$

where $\sum^{h,w}(M \cdot F_\theta(V^*))$ is a feature map spatial pooling weighted by normalized masks M of size $[n, h, w]$. Then a variance is calculated across frames, and we accumulate all c dimensions to get OSV. Fig. 14 compares OSV case-wise among video editing methods. Notably, besides surpassing other methods, edited videos by Video-P2P have a comparable semantic consistency with real-world input videos.

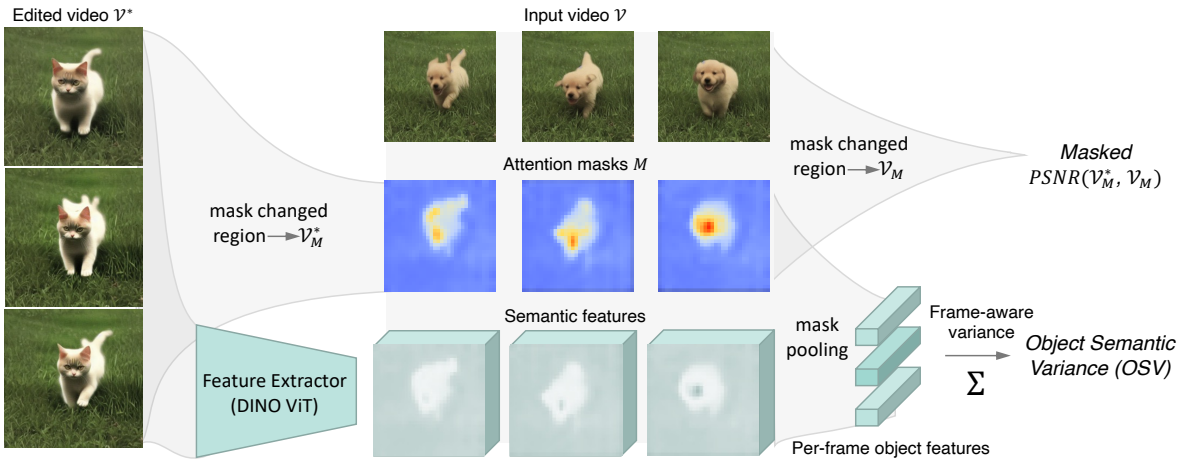

Figure 13. Illustration of Masked PSNR (M.PSNR) and Object Semantic Variance (OSV). Equipped by attention mask M for the edited object, we can evaluate the reconstruction quality in the background, and the semantic consistency in the edited region.

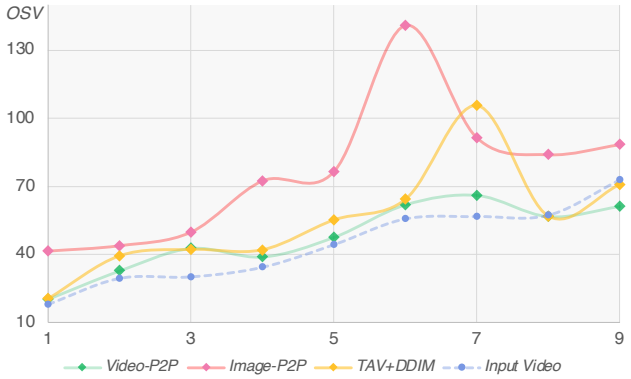

Figure 14. A case-wise comparison of OSV. Video-P2P can generate highly consistent sequences with OSV 47.57 even compared to the input real-world videos (OSV 44.32).

Figure 15. Smooth Results. Observing the difference by enlarging the point-wise tracking result.

User Study. We conduct a user study by distributing a Google Form to users to compare our method with other state-of-the-art methods. The user study interface is shown in Fig. 16. Following [13], we ask users to rank video editing results based on structure preservation, text alignment, and overall quality. We collect pairwise comparison results on eight videos and 32 cases, totaling 414 pairwise comparison responses.

7. Additional Results

Most videos in this paper are from YouTube[1] and GIPHY[2]. Some samples are from TAV [47] and text2live [2]. More results are shown in this section. With prompt refinement, Video-P2P can also complete global editing, such as transferring styles as illustrated in Fig. 18. Videos are transferred into children drawings and oil paintings without changing the structure and content of the original video.

It is possible to control the editing degree with different attention-replacing ratios. As shown in Fig. 19, replacing 90% attention maps (the 4th row) can completely change the "man" into an "Iron Man", while replacing 40% or 60% (the 2nd and 3rd rows) can only change his clothes. Here replacing K% attention maps means using maps from the source prompt during the first K% denoising steps.

More attention re-weighting results are shown in Fig. 20. By changing the scale parameter c in Eq. 14, Video-P2P can control the extent of a semantic token in a video, like "night" and "snowy".

Video-P2P is also able to edit longer videos. Examples of editing 24-frame videos are demonstrated in Fig. 21.

There is room for improvement in the generated videos to achieve smoother results. The original results are limited by the image diffusion models. It has practical significance because video generation models are not always available. They are expensive to utilize and have weaker spatial diversity. To generate smoother results with our method, one can simply replace the image diffusion model with a video generation model as shown in 15.

[1]https://youtube.com/
[2]https://giphy.com/

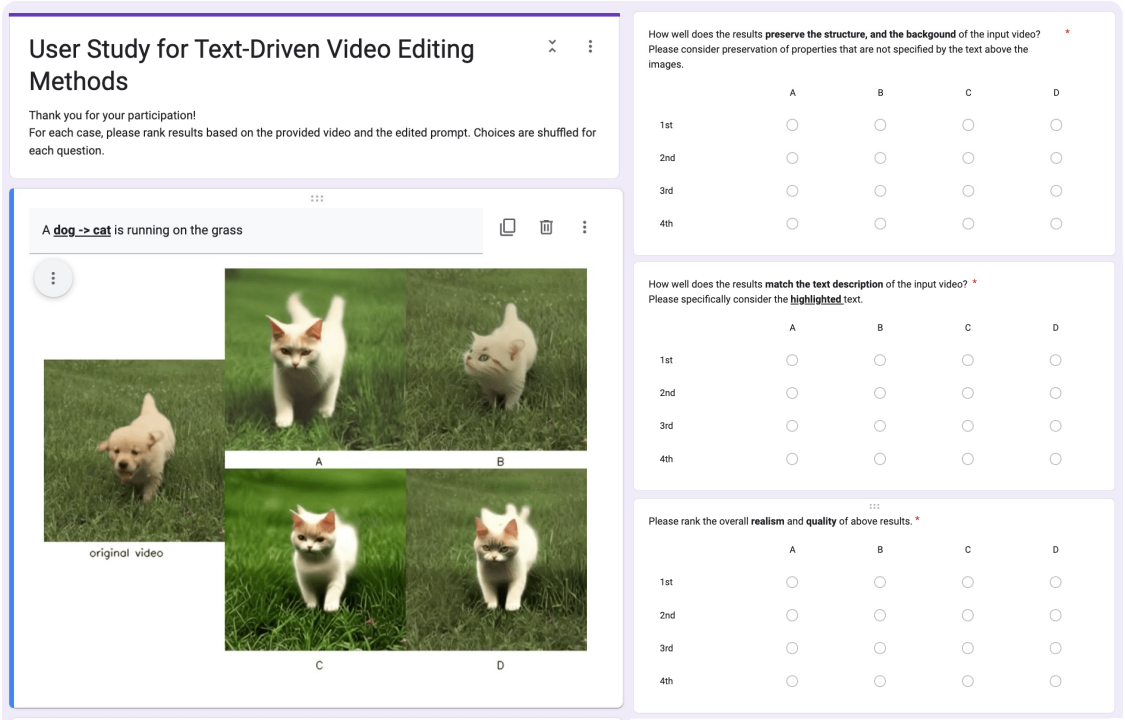

Figure 16. Screenshots of the user study interface. We shuffle four methods randomly for each question and combine them into a .GIF file with the original video input.

	Self-Attn	ST-Attn	Frame-Attn
PSNR↑	12.43	22.73	22.75

Table 4. Comparison among different attention structures.

	test set	CLIP↑	M.PSNR↑	LPIPS↓	OSV↓	Temp↑
Dreamix	Dreamix	0.3355	17.88	0.4329	49.56	**0.9718**
Ours	Dreamix	**0.3354**	**20.67**	**0.3210**	**47.51**	0.9706

Table 5. Results on Dreamix's demos.

Frame Attention is similar to ST-Attn. It is not considered as an important contribution. We just point out that the former frame has little influence during the video editing. The attention is for keeping the id. Temporal information can be preserved from the video inversion and tuning. Frame attention is a simpler choice for us.

Dreamix was not included in the baselines because it is not available for inference. We compare the quantitative results on Dreamix's demos and show them in 5.

8. Failure Case

Video-P2P may experience issues in cases where the attention maps generated by the pre-trained image diffusion model are ambiguous. For example, as shown in Fig. 17, when we want to change the property of the motorbike, both Video-P2P and Image-P2P change the man into a Lego toy (the 3rd and 4th row). This issue stems from incorrect attention maps in the 2nd row, where the man is erroneously included in the motorbike's attention maps. As Video-P2P relies on the quality of the image diffusion model's attention maps, its performance is limited by the model's pre-training.

9. Societal Impact

This paper proposes a new framework for video editing which can edit the content on real videos. This application may be utilized by malicious parties to spread fake information. However, malicious editing detection also has great progress. We believe our work would also contribute to this region. Our experiments and analysis will help people to better understand text-to-video methods and prevent the abuse of this kind of technique.

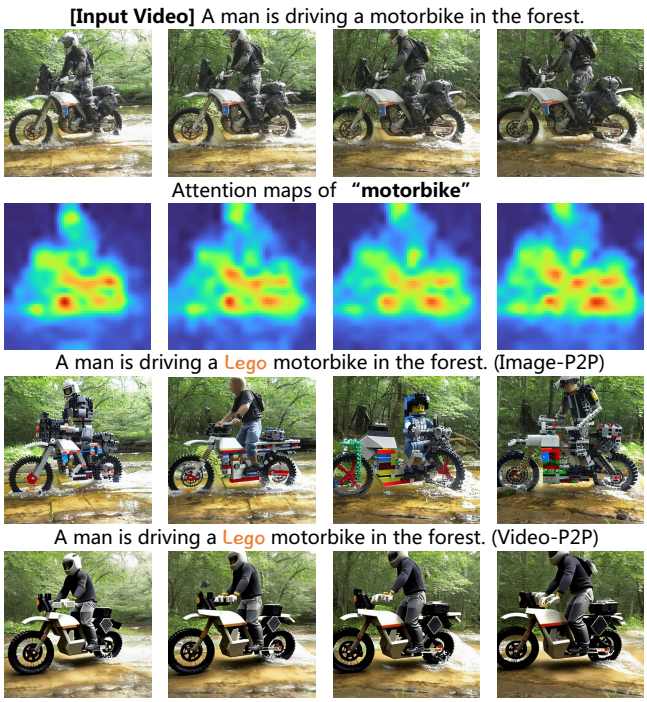

Figure 17. Failure Case. Attention maps of the pre-trained stable diffusion model are not accurate. Both Image-P2P and Video-P2P modify the "man" when editing the "motor".

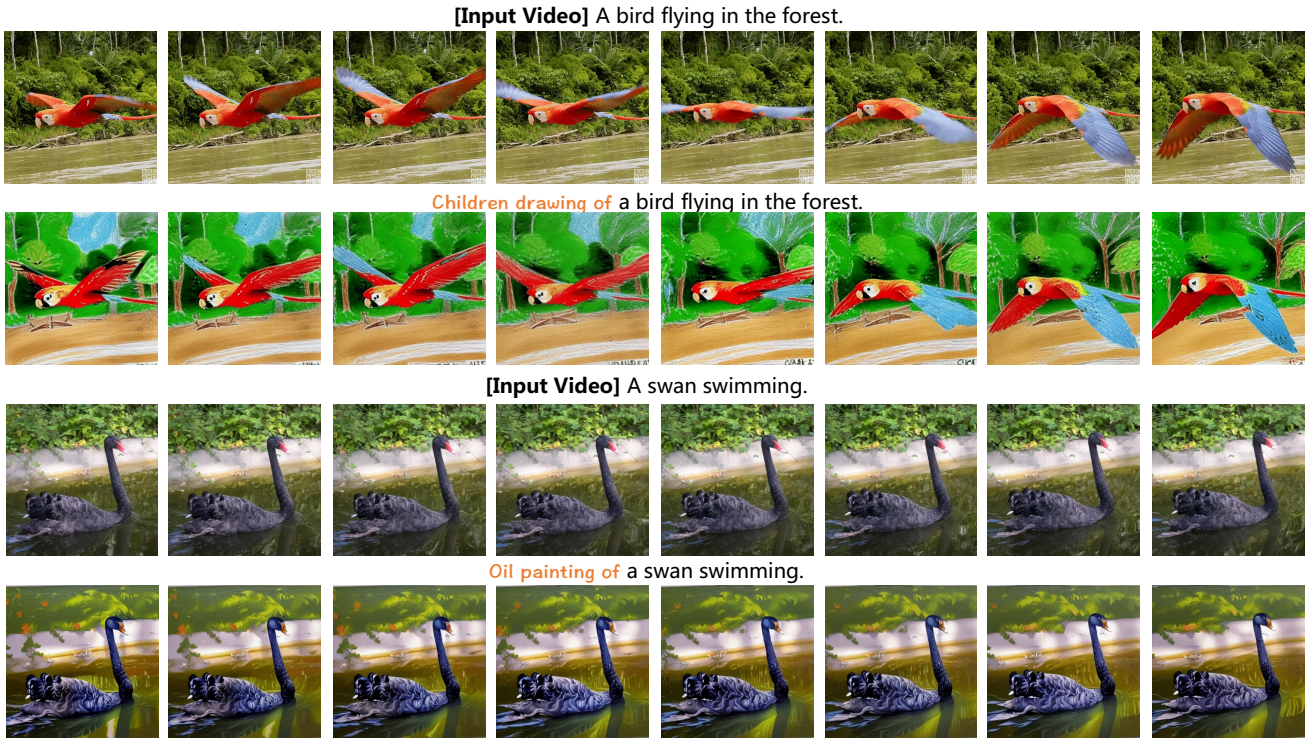

Figure 18. Style Transfer. Video-P2P also enables transferring global styles with prompt refinement.

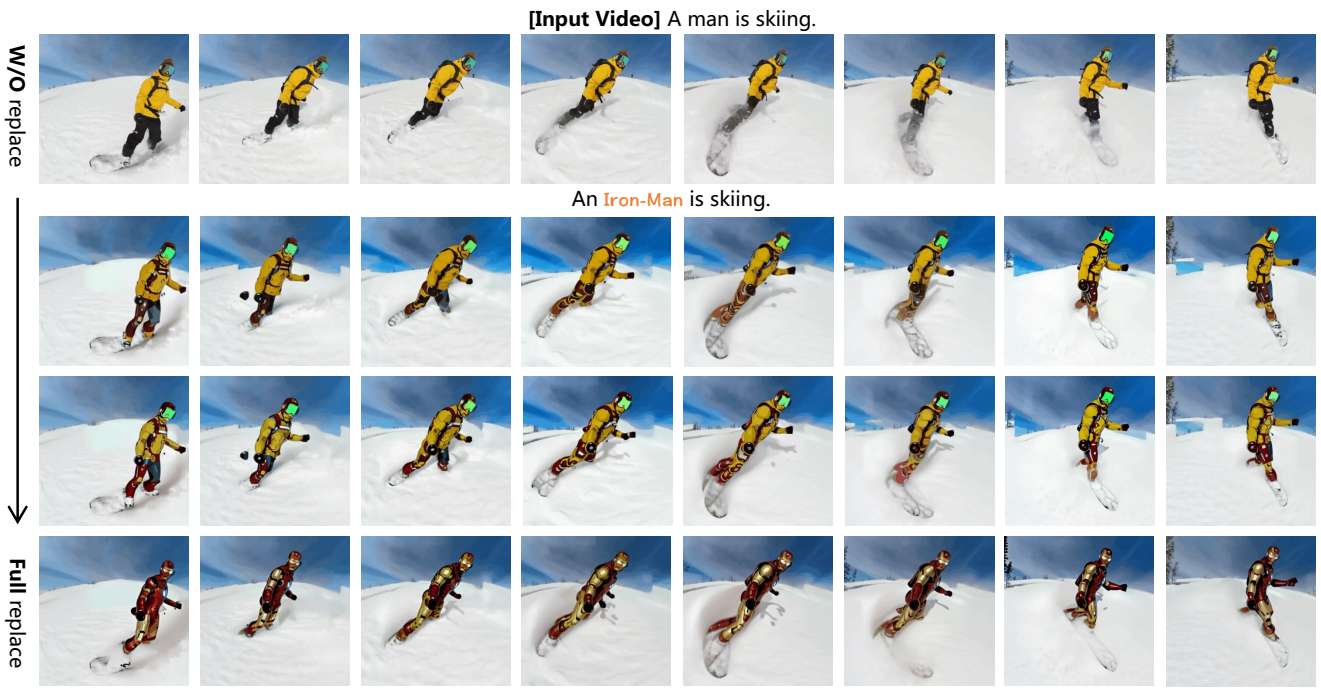

Figure 19. Ablation Study on attention replacing ratio. Replacing the cross-attention maps in more denoising steps enables larger semantic editing.

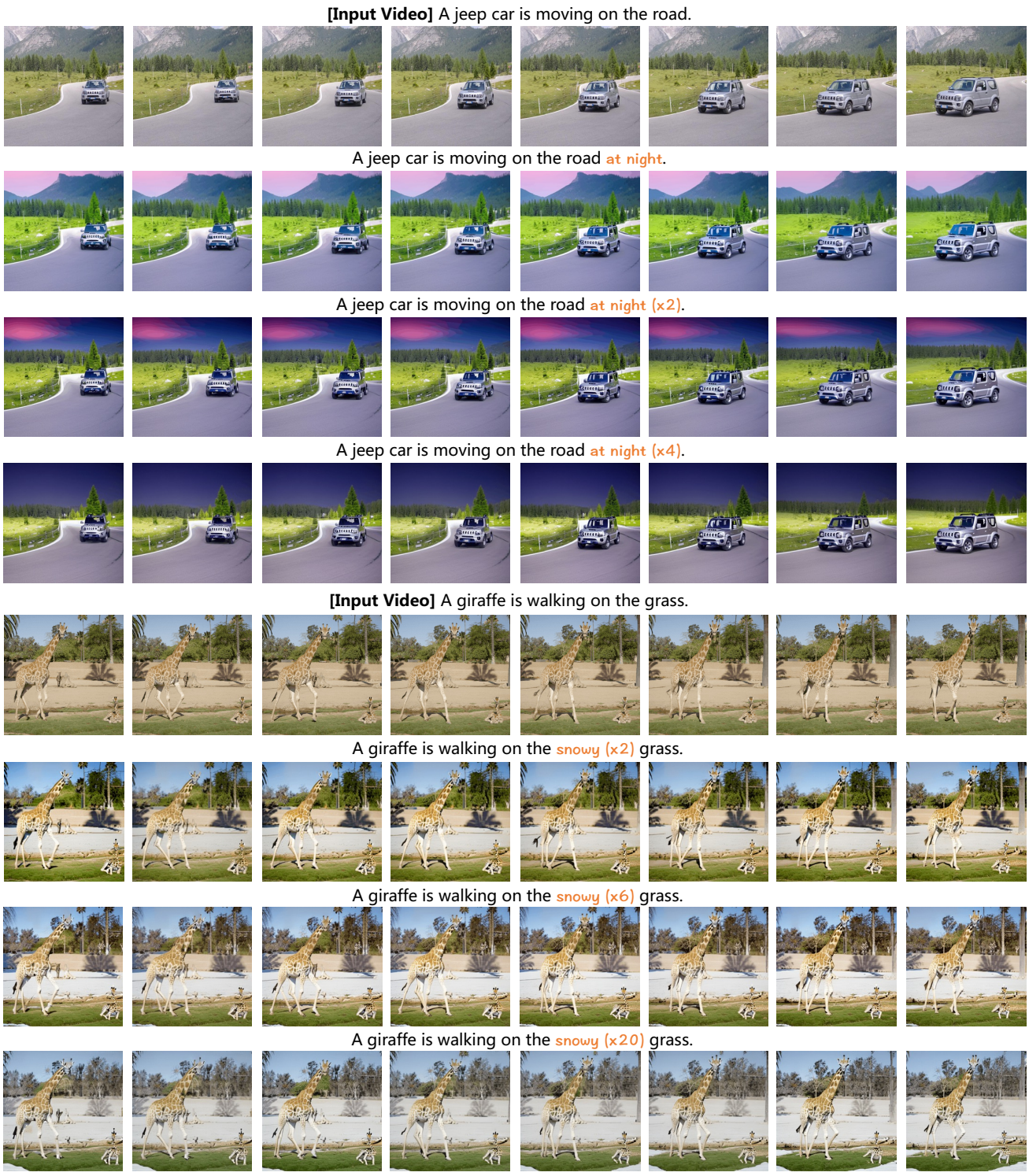

Figure 20. Prompt Refinement and Attention Re-weighting. Video-P2P can manipulate the extent of a specific token.

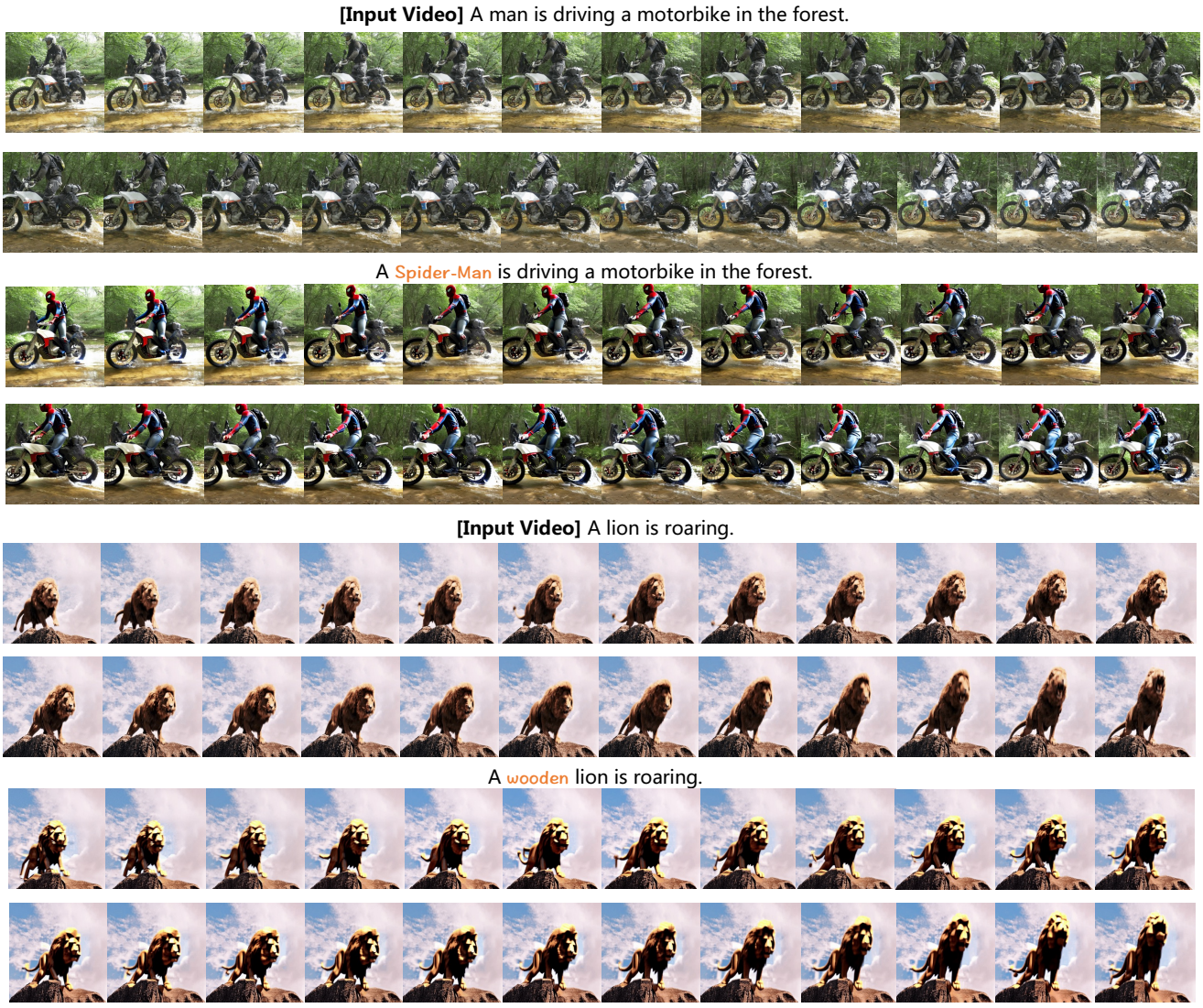

Figure 21. Editing longer videos. Video-P2P also works on videos with 24 frames.