VIDMP3: Video Editing by Representing Motion with Pose and Position Priors

Supplementary Material

8. Adaptability of VIDMP3

Personalization. VIDMP3 supports customized or personalized concepts through additive methods such as ED-LoRA [13]. In Fig. 8, we present edited videos generated using VIDMP3 with pre-trained customizations provided by [14]. During model optimization on the source video, the LoRA layers were not attached; they were utilized only during inference on the saved model. The degree of customization can be adjusted using the LoRA blend weight parameter.

In Fig. 9, we present editing results generated using VIDMP3 with the Anything-v4.0 personalized model as the foundation model. This model specializes in producing anime-style images.

Background and Style Editing We also explore background and style edits, with the results shown in Fig. 10, illustrating background edit in the second row and style edits in the fourth row.

Latent Blending. Our proposed method is robust enough to incorporate plug-and-play features such as latent blending as used in VideoSwap. Latent blending facilitates subject swapping while preserving the background region in the edited video to remain identical to the source video. The core concept relies on latents maintaining spatial correlations within pixels. During the diffusion process, at each step, the spatial values representing the background in the predicted latents are replaced with the corresponding spatial values from the source video latents, obtained during the inversion process. Results in Fig. 8, Fig. 11, and Fig. 12 utilize latent blending to preserve the background.

9. Additional results

We provide additional results of 1) Cross-Domain Editing, 2) Structure Editing, and 3) scaling to SDXL in Figs. 11, 12, and 13, respectively. We also provide some video results in the supplementary zip file.

Figure 8. **Personalization.** Using pre-trained ED-LoRA concepts during inference, we generated the illustrated frames featuring personalized subjects: a concept car (top) and the character Thanos (bottom).

Figure 9. **Theme Personalization** Edited video frames rendered in anime style using Anything-v4.0 as the foundation model in VIDMP3.

Figure 10. **Background and Style Edit.** Results of background modification (top) and style modification (bottom) using SD-v1.5 as the foundation model in VIDMP3.

10. First Frame Editing method

We show results of the first frame edit propagation method AnyV2V for the case of swapping "silver jeep" to a novel car whose image has been provided. We use AnyDoor to edit the first frame of the video (as AnyV2V suggests), and then provide this to AnyV2V for video editing. Firstly, we note that the editing quality of AnyDoor is subpar and also requires human effort in masking regions. However, it should be noted that the quality of AnyDoor is better for editing than the prompt-based method InstructPix2Pix which is another option employed by AnyV2V for first frame editing. Secondly, we notice that the identity of the car changes drastically over frames finally becoming gray, which shows that AnyV2V cannot handle pose changes of objects. For comparison with VIDMP3 for this case, please refer to the first row of Fig. 8.

11. MOTIONGUIDE Architecture

We utilize the correspondence map C_n and segmented depth map D_n as inputs to the MOTIONGUIDE module. To reduce computational complexity, these inputs are scaled down using the same scaling factor applied by the VAE encoder in most T2I models. The first convolutional layer then expands the input from three channels to 64 channels. The second convolutional layer operates on a 128-channel

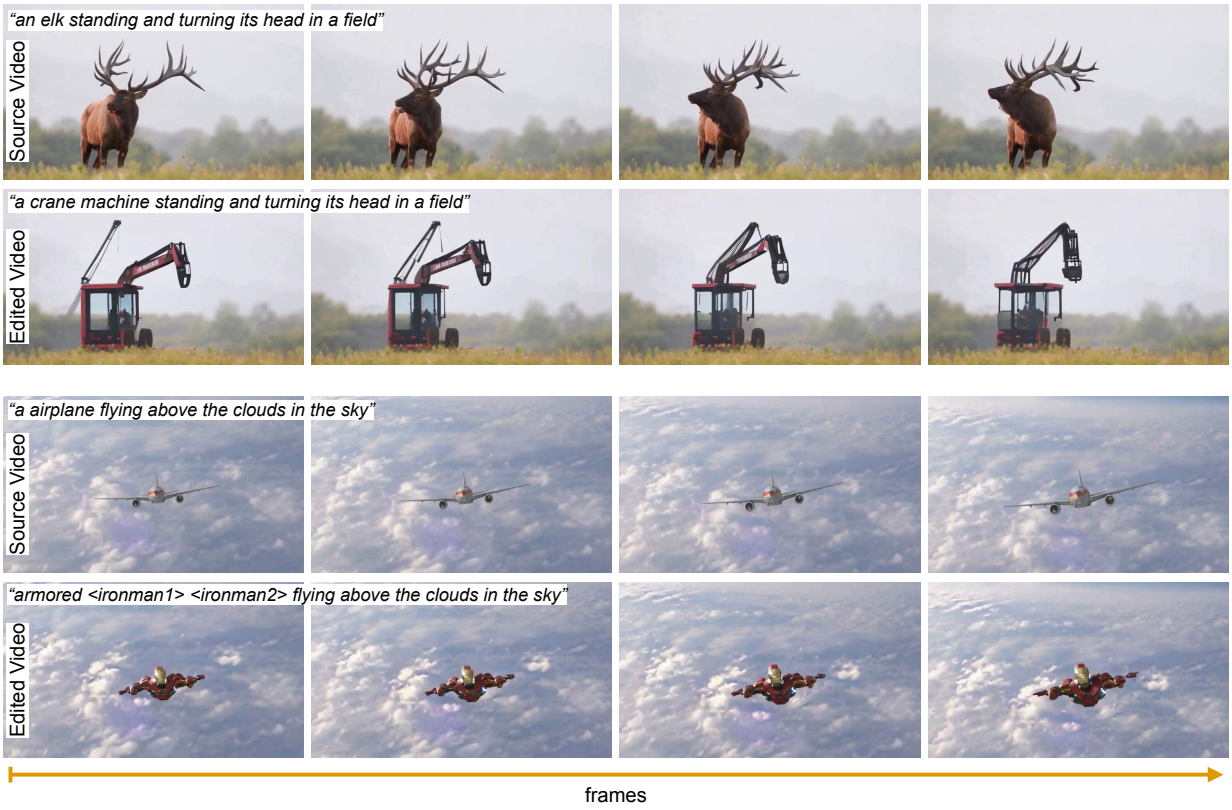

Figure 11. **Cross-Domain Edit.** Examples of cross-domain edits where an animate object is replaced with an inanimate object (top: "elk" → "crane machine") and an inanimate object is replaced with an animate object (bottom: "airplane" → "Ironman").

Method	Structure		Cross-Domain	
	Image-Text	Image-Image	Image-Text	Image-Image
Tune-A-Video	25.64	**97.74**	25.57	95.01
FateZero	25.55	97.39	24.25	94.60
VideoSwap	26.70	97.71	27.19	95.13
VidMP3	**26.74**	97.58	**30.75**	**97.94**

Table 1. **Quantitative evaluation with CLIP ViT-L/14@336px.**

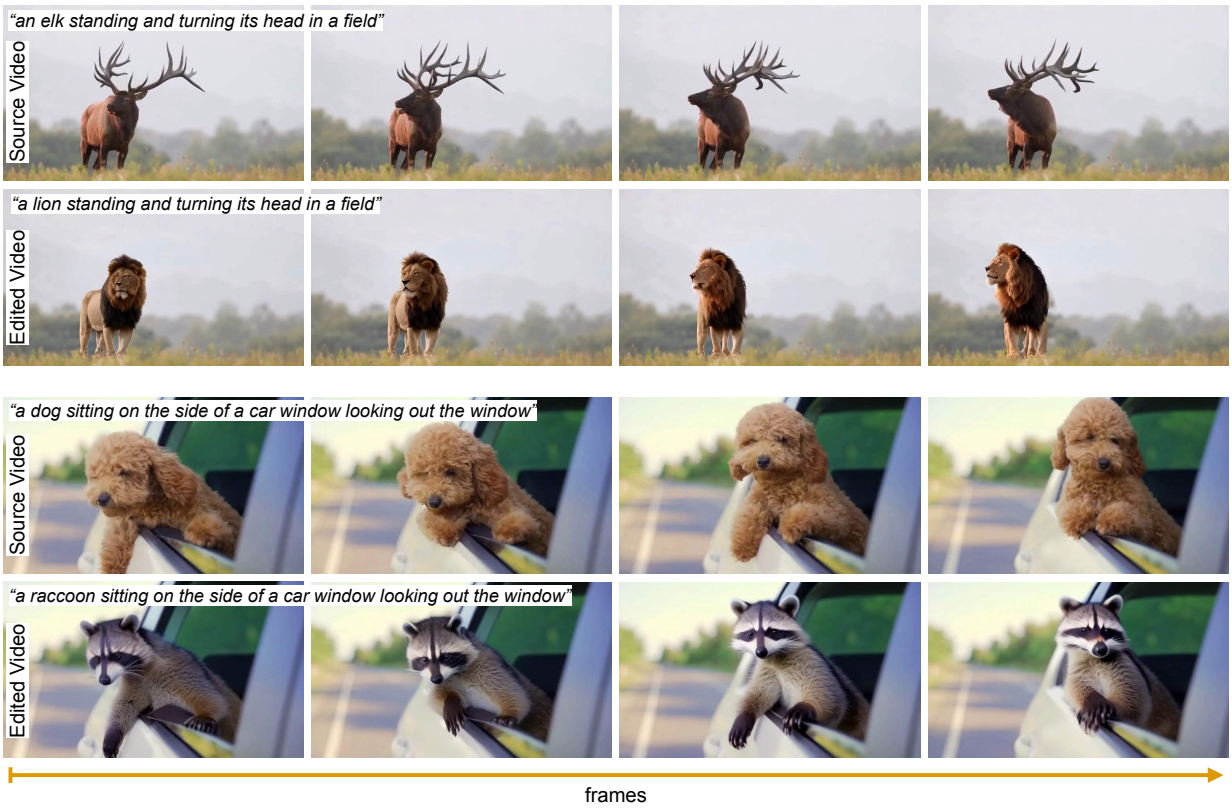

Figure 12. **Structure Edit.** Examples of structural editing, while keeping the edited subject in the same domain.

input, formed by concatenating the positional encoding P with the output of the previous convolutional layer, producing an output of 256 channels. Finally, a linear layer transforms this 256-channel activation into the desired number of channels, making it suitable for integration with the attention layer's values. Figure 15 illustrates the architecture of the MOTIONGUIDE module.

12. Evaluation

We evaluate VIDMP3 and previous methods using the same videos as described in Sec.4 **Datasets**. 180 edited results from each video editing method are compared in both automatic and human evaluation settings as described below.

Automatic Evaluation.

We utilized CLIP-Score [16] as an automatic evaluation metric to quantitatively assess all video editing methods. To compute the video-text alignment score for a test video, we averaged the image-text alignment scores across all its frames. Subsequently, the video-text alignment scores of all test videos were averaged to derive the overall video-text alignment score for each method.

As a preliminary analysis of temporal consistency in a test video, we calculate the image-to-image alignment score for every alternate frame pair and average these scores across all frames to determine the video's temporal consistency. The temporal consistency scores of all test videos are then averaged to compute the overall temporal consistency score for each method.

The results of the automatic evaluation, categorized into (1) Structure Editing and (2) Cross-Domain Editing, are summarized in Table 1. For structure editing, VIDMP3 achieves performance comparable to previous methods. However, for cross-domain editing, VIDMP3 demonstrates significantly superior performance.

Human Evaluation. We conducted a controlled laboratory study to evaluate different methods based on the following criteria: (1) Subject Identity, (2) Motion Alignment, (3) Temporal Consistency, and (4) Overall Preference. Preference-based feedback was collected for all 180 edits from 10 participants, with each participant providing ratings for all edits, resulting in a total of 180 ratings per participant. While a larger sample size of feedback per edit is generally preferred, the task of identifying issues in the edited results is relatively straightforward, making 10 participants a reasonably sufficient number for this study. The human evaluation results shown in Fig. 7 of the main paper

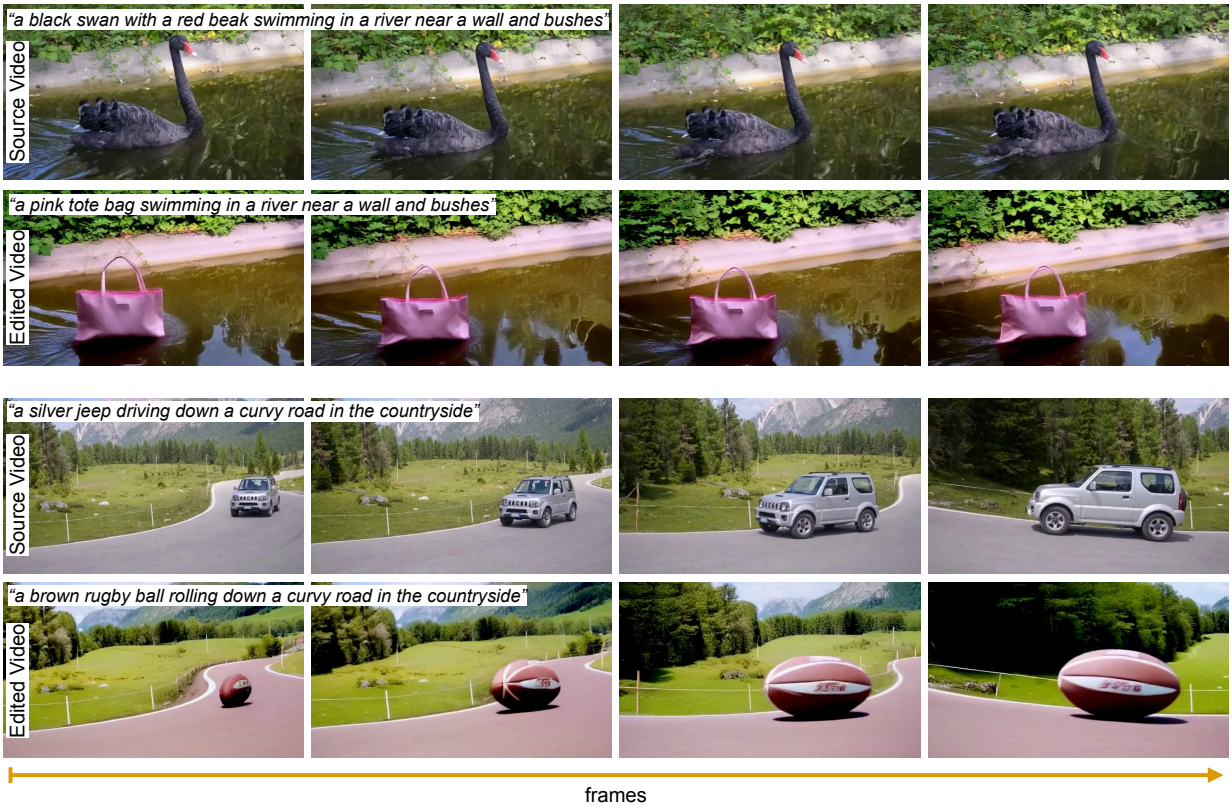

Figure 13. **More SDXL Results.** Additional examples of novel concepts generated using SDXL as the foundation model in VIDMP3.

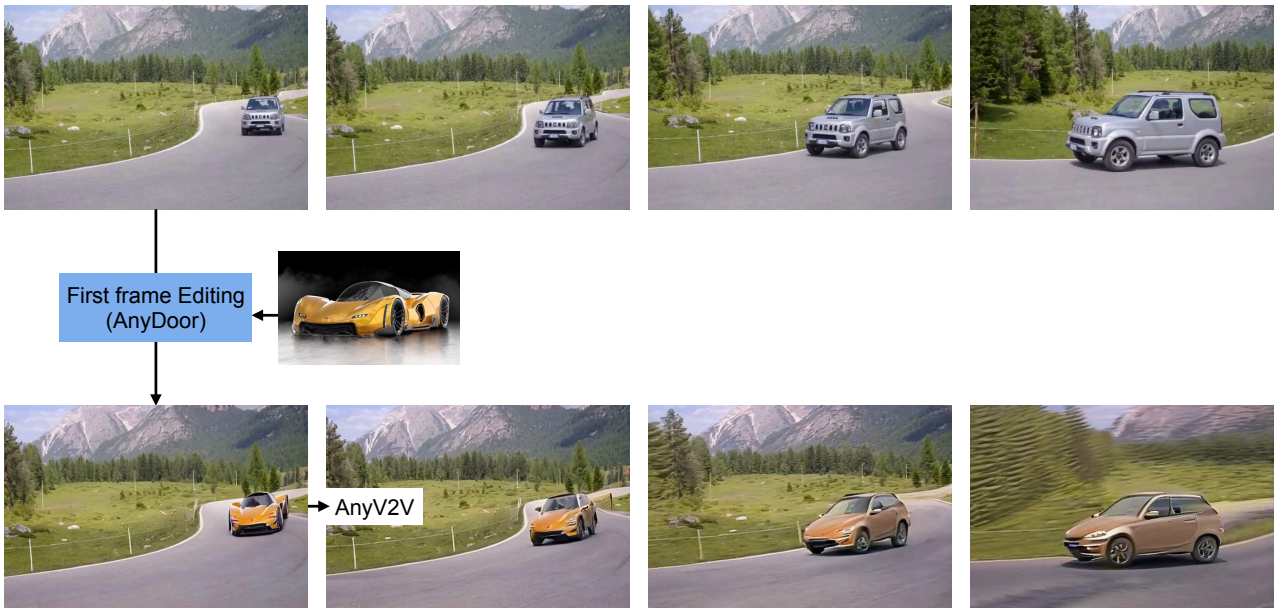

Figure 14. **Results of AnyV2V on subject swapping.** We observe that AnyV2V cannot handle pose changes in the subject. Additionally, it relies on the image editing quality of the first frame which is of poor quality and also requires human effort. Compare to Fig.8 which shows our results for the same edit.

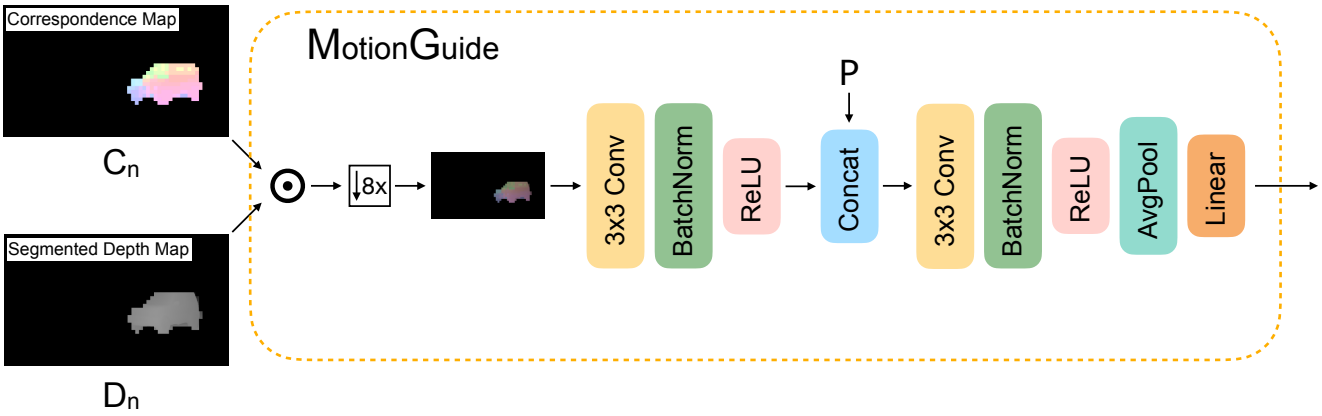

Figure 15. **MOTIONGUIDE Architecture.**

clearly indicate a strong preference for our method.

For each editing concept, the rating interface displays the source video, source prompt, edit prompt, and the edited videos generated by all methods under comparison. For personalized video editing, the interface also includes the reference images utilized for ED-LORA-based personalization.