VideoHandles: Editing 3D Object Compositions in Videos Using Video Generative Priors—Supplementery Material

Juil Koo[1*] Paul Guerrero[2] Chun-Hao P. Huang[2] Duygu Ceylan[2] Minhyuk Sung[1]
[1]KAIST [2]Adobe Research

https://videohandles.github.io

In this supplementary material, we begin by presenting additional results, including experiments with a more recent advanced video generative model, CogVideoX [4], an alternative 3D reconstruction method [5], real and stylized video editing, and longer video editing (Section S.1). Next, we provide qualitative results demonstrating VideoHandles' superior temporal consistency using concatenated vertical slices across frames (Section S.2). We then compare inference time and computational cost (Section S.3), followed by implementation details (Section S.4) and a detailed setup of the user study (Section S.5), as presented in Section 5 of the main paper.

S.1. More Results

Other video generative models. As discussed in Section 6 of the main paper, while our method is constrained by the capabilities of video generative models, it is independent of the choice of video generative models. Results with a more recent advanced video generative model, CogVideoX [4], produce much sharper and more detailed outputs, as shown in Figure S1 (a). Notably, the results with CogVideoX [4] also demonstrate the adaptability of our method to various video generative models.

Real and stylized videos. We present additional editing results for real and stylized videos in Figure S1 (b) and (c).

Other 3D reconstructions. Figure S1 (b) shows results using MONST3R [5] instead of DUST3R [3] for 3D reconstruction, demonstrating the robustness of our method with other 3D reconstruction methods.

Longer videos. While we used 51-frame (2-second) videos due to GPU memory limits (80GB on the A100 we used), our method can be applied to longer videos if video models can generate them within limited memory, as 3D reconstruction is relatively lightweight. We present additional 102-frame results in Figure S1 (c), obtained by splitting inference across two GPUs.

S.2. Temporal Consistency Across Frames

Figure S2 also shows concatenated vertical slices across frames taken at the same position, highlighted by the blue line in the bottom-left image. These images clearly demonstrate that our output maintains temporal consistency compared to other per-frame-based editing methods.

S.3. Inference Time and Computational Cost

As no other method attempts to edit object compositions in videos, we compare ours with the most technically related image editing method, DiffusionHandles [2]. While DiffusionHandles requires editing each frame individually, our method processes all frames in a single feed-forward pass, reducing the editing time to 6 minutes compared to DiffusionHandles' 96 minutes for a 51-frame video on A6000 GPUs. Please note that additional time may be accounted for due to inter-GPU communication overhead, as we split inference across two GPUs to accommodate VRAM constraints.

[*]Work done during an internship at Adobe Research.

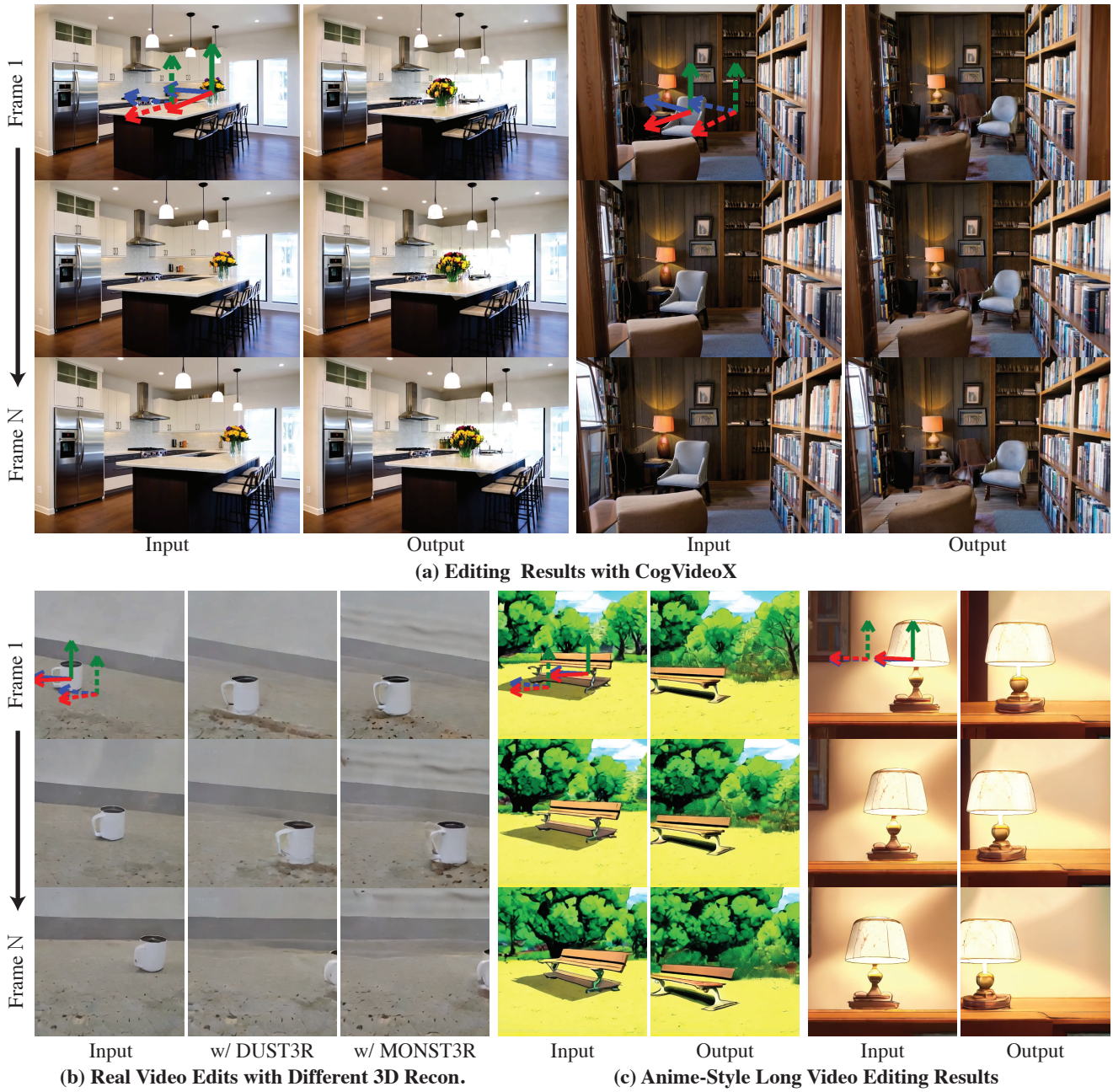

Figure S1. **(a) Using CogVideoX [4] instead of OpenSora [6] as the video generative model; (b) additional real video result with different 3D reconstructions (DUST3R [3] and MONST3R [5]); (c) longer and stylized videos.** Solid axes represent the original 3D position, dotted axes the user-provided target position. Zoom in for the best view.

S.4. Implementation Details

To define the object binary mask in the first frame, $\mathcal{B}^{(1)}$, we leverage SAM [1]. For the object transformation energy function, $\mathcal{G}_o$, we use features from all 24 spatial self-attention layers in OpenSora [6], while for the background preservation energy function, $\mathcal{G}_b$, we only use features from the first 14 self-attention layers. For both the input video generation and the guided generative process for editing, we use 30 steps with a classifier-free-guidance scale of 7. For the gradient step sizes of the energy functions discussed in Section 4.4, we set ρ_o and ρ_b to 650 and 100, respectively, for all editing examples in the comparisons. To better preserve the background details of the input video and smoothly adjust it to the new object

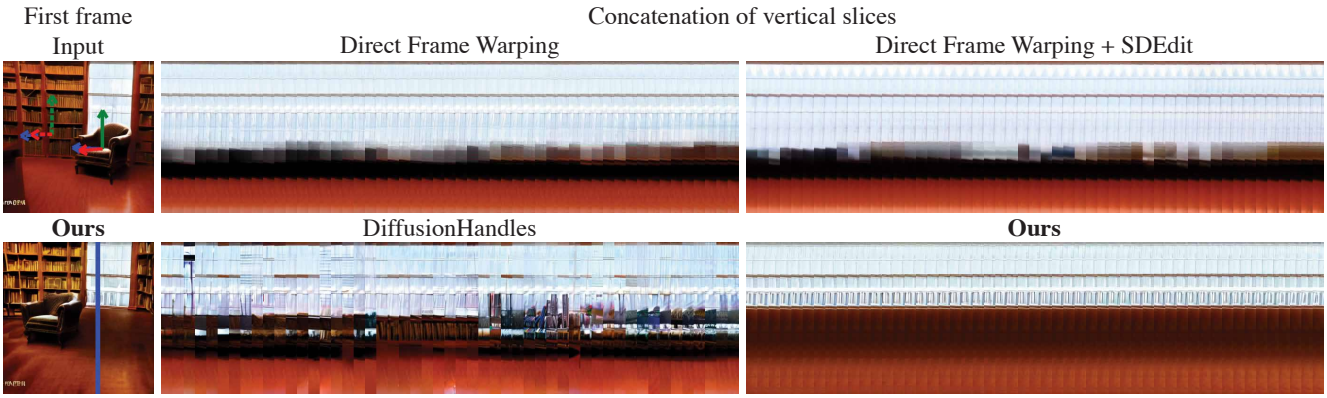

Figure S2. **Concatenation of vertical slices across frames.** We show the same spatial slice across time. Our results have higher temporal consistency due to using a video prior.

composition, we initially compute the background preservation energy function without the averaging operator, measuring feature discrepancy element-wise. We then transition to the average loss to allow the background to adapt smoothly to the new object composition. This switch occurs at the ninth step out of 30 steps in the generative process.

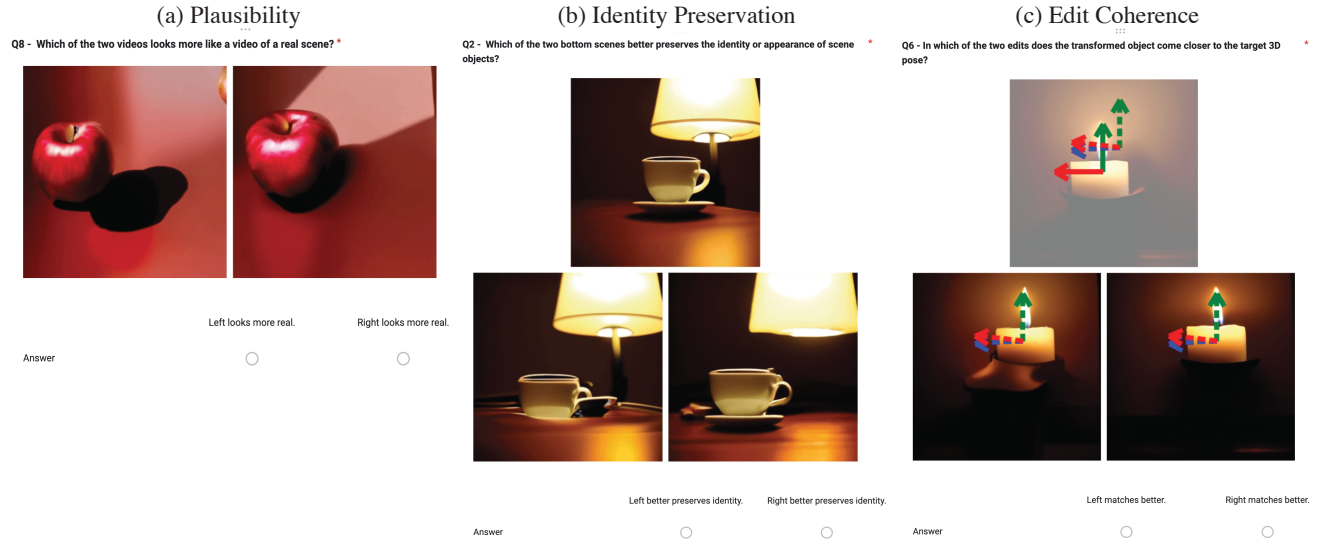

Figure S3. **Screenshots of user study questions.** We asked three types of questions in the user study to assess plausibility (a), identity preservation (b), and edit coherence(c). In each question, the videos were shown, and users selected their preferred option.

S.5. Details of User Study Setup

Figure S3 presents screenshots of our user study questions, which evaluate plausibility, identity preservation, and edit coherence of the edited videos.

For plausibility, participants were shown two videos—one generated by our method and either a competing method or the input video (to represent an upper bound of plausibility)—and asked: `"Which of the two videos looks more like a video of a real scene?"`

For identity preservation, we showed an input video along with two edited videos, and asked: `"Which of the two bottom scenes better preserves the identity or appearance of the scene objects?"`

For edit coherence, we visualized the 3D axes before and after transformation in the input video to help users understand how the selected object should be transformed, then asked: `"In which of the two edits does the transformed object come closer to the target 3D pose?"`

We conducted the user study separately for comparisons with other baselines and ablation cases. The number of participants was 21 and 7, respectively. In Figure 5 of the main paper, we present 95% confidence intervals to reflect the certainty of the results. Each participant answered 20 randomly selected questions on plausibility, 15 on identity preservation, and 15 on edit coherence.

References

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 2

[2] Karran Pandey, Paul Guerrero, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J Mitra. Diffusion handles enabling 3d edits for diffusion models by lifting activations to 3d. In *CVPR*, 2024. 1

[3] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 1, 2

[4] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 2

[5] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 1, 2

[6] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 2