

# Supplementary Materials: HeroMaker: Human-centric Video Editing with Motion Priors

Anonymous Authors

This supplementary materials consist of a PDF file and a video to provide more details of our methods and additional results, organized as follows:

- The pipeline of the whole model as in the main paper. (video.mp4 00:00 ~ 00:42)
- Implementation details of the motion warping acquisition (Sec. 1).
- Performance comparison (Sec. 2). (video.mp4 00:42 ~ 02:37)
- Multiple persons editing results (Sec. 3). (video.mp4 02:37 ~ 03:20)
- User study details (Sec. 4). (video.mp4 03:20 ~ 03:35)

We also give explanations aligned with the video and list below.

## 1 IMPLEMENTATION DETAILS OF THE HUMAN MOTION WARPING

In the human motion warping module, we can integrate information from both frontal and back canonical images into each frame.

For each image, we possess the camera parameters, the weight index map of the mesh  $W$ , and  $C$  as the correspondence map of the mesh, and the value in each pixel indicates the face index of the mesh. For canonical images and the  $i_{th}$  frame’s image, we first project their mesh to image coordinate using the corresponding camera parameters and determine the barycentric coordinates of each mesh face  $f_f$ ,  $f_b$ , and  $f_{s_i}$ .

Subsequently, the matching correspondence is established between correspondence map  $C$  and the coordinates of the mesh face  $f$  and get transformation matrix  $T_{s \rightarrow t} \in \mathbb{R}^{H \times W \times 2}$ . So we can also get transformation matrix  $T_{f_i \rightarrow t}$  or  $T_{b_i \rightarrow t}$  through replace  $f_{s_i}$  with  $f_f$  or  $f_b$ .

Because both the front and back view images need to be warped into each frame, we need to specify the sources of information in different locations.

Initially, considering the transformation matrices from a source image to a target image, there are two types. The first is to transform the visible part, and the second is to complete the original mesh texture before the transformation.

The mask corresponding to the human mesh is  $S_c$  for a frame. First, we choose to use the transformation involving only the visible parts as  $M_{f_i}$ , converting the texture of the front view canonical mesh into the mesh of the video frame, and get  $T_{f \rightarrow i}$ . Subsequently, we employ the transformation involving only the visible parts  $M_{b_i}$  from the back to convert the texture of the back view canonical mesh into the video frame’s mesh in the remaining part and get  $T_{b \rightarrow i}$ . Finally, we use the front view image information for completion the remaining small regions  $M_{f_2}$  that are not covered. Overall, the motion warping as:

$$M_b = M_{b_i} \cap M'_{f_i} \quad (1)$$

$$M_{f_2} = S_c \cap M'_{f_i} \cap M'_b \quad (2)$$

$$M_f = M_{f_i} \cup M_{f_2} \quad (3)$$

$$T_{front}^{tr} = T_{f \rightarrow i} \times M_f \quad (4)$$

$$T_{back}^{tr} = T_{b \rightarrow i} \times M_b \quad (5)$$

## 2 PERFORMANCE COMPARISON

We present the comparison results between our approach and all baselines.

On the prompt-based video editing task in Fig. 1, videos edited by Text2Video-Zero [5] maintain textual fidelity but encounter flickering and inconsistent shapes. Rerender-A-Video [7] has achieved some improvement in terms of flickering via hierarchical cross-frame consistency constraints. However, it still lacks accurate correspondences and leads to issues of rotation cases. Rerender-A-Video use canny as a condition for ControlNet [3] resulted in the inability to capture rotational correspondences. Therefore, in Text2Video-Zero, we use openpose [1] as a condition for ControlNet [8] but found it still needs to work on addressing issues effectively. StableVideo [2] relies on the NLA [4] method, which faces challenges in effectively separating foreground and background atlases, leading to a weaker textual fidelity. CoDeF [6] utilizes optical flow but encounters inaccuracies in detecting large-scale motions, resulting in semantic-less canonical images, yielding results differing from natural video. Our approach demonstrates superior temporal consistency and accurate correspondence, yielding commendable results in two editing cases.

In Fig. 2, we attempted to edit videos using editing frames and editing layers in NLA [4] model. The results show that this method needs help distinguishing between foreground and background atlas for human motion videos. Establishing correspondences in complex human motion is challenging for CoDeF [6], which leads to visual flaws. Additionally, our method has achieved better user interactive video editing performance with motion priors.

## 3 MULTIPLE PERSONS EDITING RESULTS

Compared to other methods, our approach, along with CoDeF [6], enables flexible editing of multiple persons according to specific requirements. In Fig. 3, we show the results of all baselines and our methods for editing the entire video. Our and CoDeF [6] method have high textual fidelity, outperforming other baseline methods slightly. CoDeF [6] learns the deformation field and the canonical image without structure information, causing semantic-less canonical images. Additionally, we achieve better visual effects with motion priors.

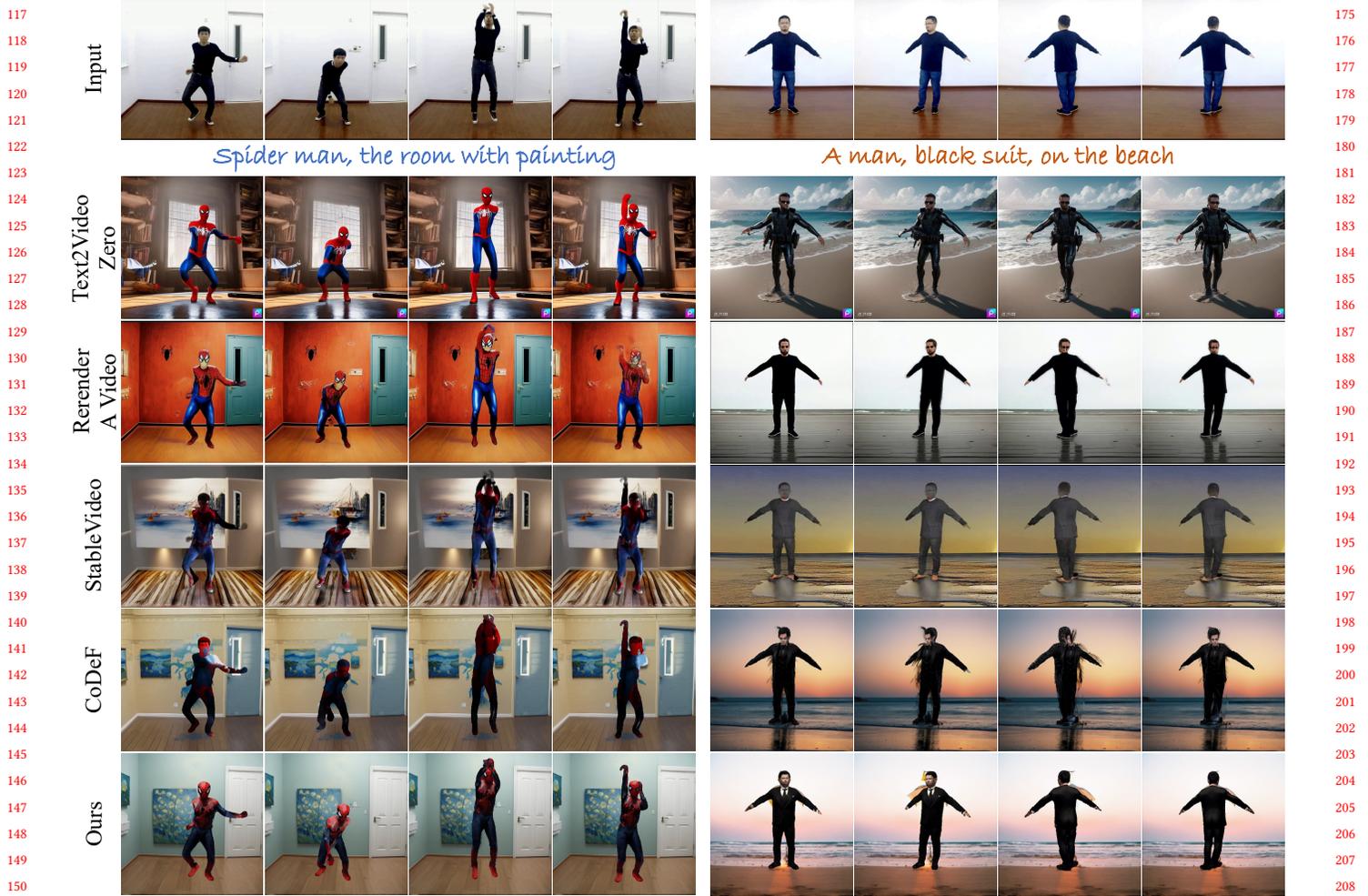
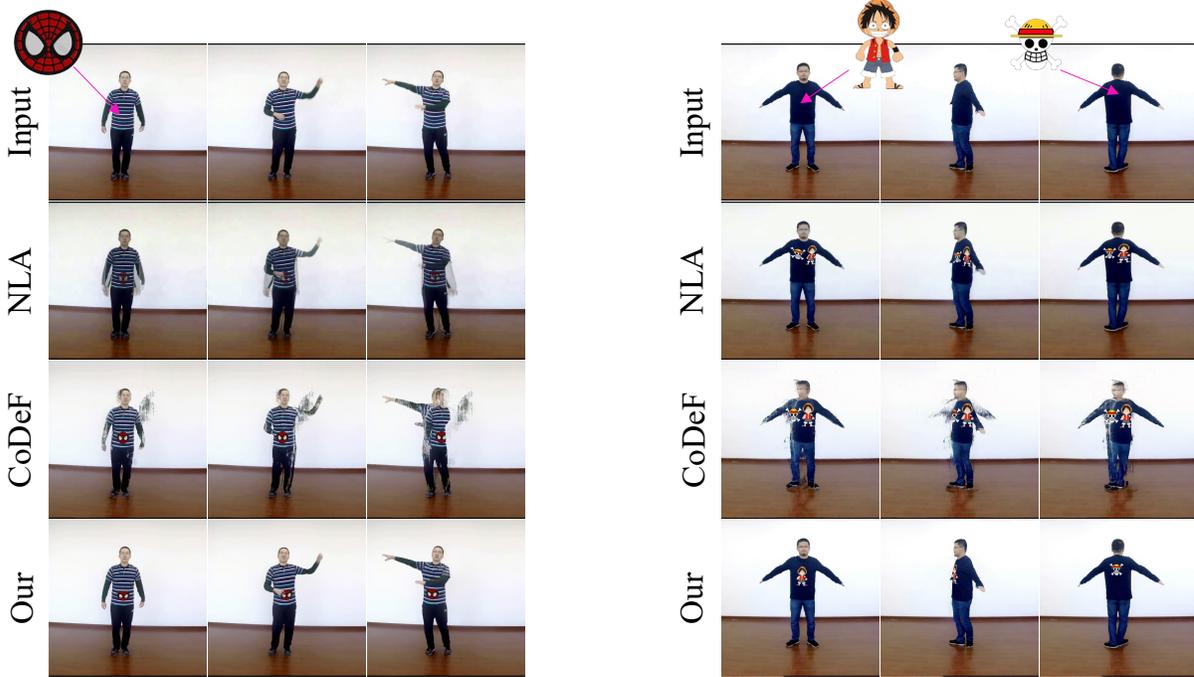


Figure 1: More qualitative results of the prompt-based video editing. We present two more visual comparisons of our approach against all baselines regarding prompt-based video editing. Compared to other methods, our method is temporally coherent and plausible. (video.mp4 00:42 ~ 01:35)

#### 4 USER STUDY DETAILS

Since the edited video is very subjective, we conducted a user study to show the effectiveness of the proposed method over all baselines. Specifically, we ask 17 subjects on five different methods (*i.e.*, Text2Video-Zero [5], Rerender-A-Video [7], StableVideo [2], CoDeF [6], NLA [4] and ours). We provide eight samples of the results and let them score videos based on three aspects: textual fidelity & consistency, shape preservation, and visual effect for the prompt-based video editing and visual effects for the user interactive video editing. Fig. 4 and Fig. 5 show the example in our questionnaires.



(a) Comparison of our method with NLA [4](frame editing) and CoDeF [6] in local editing. (video/mp4 02:14 ~ 02:27)

(b) Comparison of our method with NLA [4](layer editing) and CoDeF [6] in local editing. (video.mp4 02:27 ~ 02:37)

Figure 2: More qualitative results of the user interactive video editing. We present two more visual comparisons of our approach against all baselines regarding local editing. (video.mp4 01:35 ~ 02:37)

Prompt: Iron man, Spider man, super man, on the beach



Figure 3: Performance comparison of multiple persons editing. We present a visual comparison of our approach against all baselines. Text2Video-Zero [5], Rerender-A-Video [7], and StableVideo [2] have insufficient textual fidelity, and CoDeF [1] lack motion priors. Our method has a high textual fidelity and visually natural result. (video.mp4 02:37 ~ 03:20)

Prompt: spider man, the room with painting



A

B

C

D

E

**Textual fidelity & consistency:** The video has a high matching degree with the prompt, and there is a smooth transition between video frames. The texture of the same position on the human body does not change over time.

**Shape preservation:** The structure of the human body is intact, and there are no phenomena such as missing limbs or distortion.

**Visual effect:** After comprehensive evaluation, the video having an overall better effect.

Please watch the above videos and score each video based on three aspects: **textual fidelity & consistency**, **shape preservation**, and **visual effect**. The scores should **range from 0 to 1**.

Figure 4: An example of our user study. For prompt-based video editing, we provide a prompt for users to score videos based on three aspects: textual fidelity & consistency, shape preservation, and visual effect.

Edited Icon:



A

B

C

**Visual effect:** Pay attention to whether the position of the pattern during person movement is correct, and whether the masking of relationship is correct.

Please watch the above videos and score each video based on **visual effect**. The scores should **range from 0 to 1**.

Figure 5: An example of our user study. For the user interactive video editing, we provide an icon for users to score videos based on visual effects.

## REFERENCES

- [1] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).
- [2] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. 2023. StableVideo: Text-driven Consistency-aware Diffusion Video Editing. *arXiv preprint arXiv:2308.09592* (2023).
- [3] Zhihao Hu and Dong Xu. 2023. VideoControlNet: A Motion-Guided Video-to-Video Translation Framework by Using Diffusion Model with ControlNet. *arXiv:2307.14073 [cs.CV]*
- [4] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. 2021. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–12.
- [5] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. *arXiv preprint arXiv:2303.13439* (2023).
- [6] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. 2023. CoDeF: Content Deformation Fields for Temporally Consistent Video Processing. *arXiv preprint arXiv:2308.07926* (2023).
- [7] Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. In *ACM SIGGRAPH Asia Conference Proceedings*.
- [8] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. [n. d.]. Adding Conditional Control to Text-to-Image Diffusion Models.