Supplementary Materials for MotionEditor: Editing Video Motion via Content-Aware Diffusion

Shuyuan Tu[1,2] Qi Dai[3] * Zhi-Qi Cheng[4] Han Hu[3] Xintong Han[5] Zuxuan Wu[1,2] * Yu-Gang Jiang[1,2]
[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing
[3]Microsoft Research Asia [4]Carnegie Mellon University [5]Huya Inc.
https://francis-rings.github.io/MotionEditor

1. Skeleton Signal Alignment

The details of our skeleton signal alignment are depicted in Algorithm 1. (x_s, y_s), h_s, w_s refer to the coordinates of the upper left corner, the height, and the width for the bounding box of source protagonist. (x_r, y_r), h_r, w_r refer to the counterparts in reference frame, *i.e.*, the coordinates of the upper left corner, the height, and the width for the bounding box of reference protagonist. w_r^* indicates the resized width of source protagonist. $\texttt{Rectangle\_Boundary}(\cdot)$ indicates $cv2.boundingRect(\cdot)$.

Algorithm 1 Skeleton Alignment

Input: Source Skeleton $\boldsymbol{S}_{sr}$, Source Mask $\boldsymbol{M}_{sr}$; Reference Skeleton $\boldsymbol{S}_{rf}$, Reference Mask $\boldsymbol{M}_{rf}$

▷ Resize Operation
$(x_s, y_s), h_s, w_s = \texttt{Rectangle\_Boundary}(\boldsymbol{M}_{sr})$
$(x_r, y_r), h_r, w_r = \texttt{Rectangle\_Boundary}(\boldsymbol{M}_{rf})$
$\boldsymbol{ratio} = w_r / \texttt{float}(h_r)$
$w_r^* = \texttt{Round}(\boldsymbol{ratio} \cdot h_s)$
$\boldsymbol{P}_S = \texttt{resize}(\boldsymbol{S}_{rf}[\, y_r : y_r + h_r, x_r : x_r + w_r \,], (h_s, w_r^*))$
$\boldsymbol{P}_M = \texttt{resize}(\boldsymbol{M}_{rf}[\, y_r : y_r + h_r, x_r : x_r + w_r \,], (h_s, w_r^*))$
$\boldsymbol{S}_{rf} = \texttt{zeros\_like}(\boldsymbol{S}_{sr})$
$\boldsymbol{M}_{rf} = \texttt{zeros\_like}(\boldsymbol{M}_{sr})$
if $w_r^* < w_s$:
  $\boldsymbol{S}_{rf}[\, y_s : y_s + h_s, x_s : x_s + w_r^* \,] = \boldsymbol{P}_S$
  $\boldsymbol{M}_{rf}[\, y_s : y_s + h_s, x_s : x_s + w_r^* \,] = \boldsymbol{P}_M$
else:
  $\boldsymbol{S}_{rf}[\, y_s : y_s + h_s, x_s - (w_r^* - w_s) : x_s + w_s \,] = \boldsymbol{P}_S$
  $\boldsymbol{M}_{rf}[\, y_s : y_s + h_s, x_s - (w_r^* - w_s) : x_s + w_s \,] = \boldsymbol{P}_M$

▷ Translation Operation
$\textbf{coordinates}_s = \texttt{where}(\boldsymbol{M}_{sr} == 1)$
$\textbf{coordinates}_r = \texttt{where}(\boldsymbol{M}_{rf} == 1)$
$\textbf{center}_s = \texttt{Mean}(\textbf{coordinates}_s, \texttt{axis} = 0)$
$\textbf{center}_r = \texttt{Mean}(\textbf{coordinates}_r, \texttt{axis} = 0)$
$\boldsymbol{v}_{trans} = \textbf{center}_s - \textbf{center}_r$
$dx = \boldsymbol{v}_{trans}[\, 0 \,], \;\; dy = \boldsymbol{v}_{trans}[\, 1 \,]$
$\boldsymbol{M}_{trans} = [[\, 1, 0, dx \,], \; [\, 0, 1, dy \,]]$
$\bar{\boldsymbol{S}}_{tg} = \texttt{WarpAffine}(\boldsymbol{S}_{rf}, \boldsymbol{M}_{trans}, \boldsymbol{S}_{rf}.shape)$
Output: Aligned Target Skeleton $\bar{\boldsymbol{S}}_{tg}$

2. Task Comparison

We propose MotionEditor to tackle a higher-level and more challenging video editing—video motion editing. Given the source video, target prompt, and reference video, our model can directly edit the motion of the source video according to that of the reference video and the description of the target, while preserving the appearance information of the source video. The comparison details are depicted in Table 1.

3. Experiment Details

Since existing methods are not designed for motion editing, we make several modifications to their models. For pose-guided video generation models, we input DDIM inverted source video latent to Follow-Your-Pose [3] and ControlVideo [10], which enables them to perform controllable video editing. For video attribute editing models, Tune-A-Video [7], MasaCtrl [1], and FateZero [4] are equipped with ControlNet [9] which enables them to accept additional controllable signals inputs. In terms of human motion transfer models, we only feed the first frame of each video to LWG [2] and MRAA [5] to follow their original pipelines. The remaining settings are identical to those in our proposed MotionEditor. We illustrate the video motion editing results of 32 frames with a $4\times$ sampling ratio for better visual appeal.

4. Discussion on Video Motion Editing and Human Motion Transfer

Video Motion Editing requires directly performing motion transfer on video over temporal dimension, which considers the per-frame dynamic background information and camera movement. In contrast, the pipeline of human motion trans-

Table 1. Tasks comparison among pose-guided image generation, human motion transfer, pose-guided video generation, video attribute editing, and video motion editing.

Task	Input	Output
Pose-guided Image Generation (ControlNet [9])	Prompt + Pose	Image
Human Motion Transfer (LWG [2] and MRAA [5])	Image + Series of Poses	Video
Pose-guided Video Generation (Follow-Your-Pose [3] and ControlVideo [10])	Prompt + Series of Poses	Video
Video Attribute Editing (Tune-A-Video [7], MasaCtrl [1] and FateZero [4])	Video + Prompt	Video
Video Motion Editing (MotionEditor)	Video + Series of Poses + Prompt	Video

Table 2. Quantitative comparisons with our proposed MotionEditor and concurrent works.

Method	CLIP (↑)	L-S (↓)	L-N (↓)	L-T (↓)
DisCo [6]	27.75	0.355	0.177	0.150
MagicAnimate [8]	28.11	0.298	0.191	0.113
Ours	**28.86**	**0.273**	**0.124**	**0.082**

fer [2, 5] only demands one single image which ignores the additional dynamic information. We conduct two comparison experiments on a case with dynamic background information and a case with camera movement. The results are shown in Figure 7 and Figure 8. We can see that our MotionEditor can perform motion editing in higher quality while maintaining the additional dynamic information. The results of human motion transfer models are limited to a single given image, thereby resulting in static background information and camera movement. In addition, they are usually constrained to images with clean backgrounds. When the background is intricate with complex scenes, their models exhibit limited capabilities.

Human motion transfer methods (*e.g.*, LWG [2] and MRAA [5]) are also sensitive to the initial source pose. They demand simple initial source poses, for example, a person standing squarely. We conduct an experiment that selects frames with different and complex poses as the initial image for motion transfer. The result is illustrated in Figure 9. Note that only the first frame of each transfer is presented. It indicates that human motion transfer models fail to directly handle different and complex initial poses while our MotionEditor shows its superiority in video motion editing.

5. Additional Results

Figure 2 and 3 show additional video motion editing results of MotionEditor. Figure 4 shows the complete comparison results. Figure 5 and 6 provide additional comparison results and ablation study results.

Table 3. Quantitative validation of the impact of frame count.

Input Setting	CLIP (↑)	L-S (↓)	L-N (↓)	L-T (↓)
Single Image	28.82	0.312	0.168	0.121
2 Frames	28.68	0.292	0.160	0.117
4 Frames	28.77	0.287	0.151	0.104
8 Frames	28.72	0.284	0.142	0.097
Intact Video	**28.86**	**0.273**	**0.124**	**0.082**

Table 4. Ablation results for attention injection.

Method	CLIP (↑)	L-S (↓)	L-N (↓)	L-T (↓)
w/o mask	28.42	0.295	0.133	0.113
Attention Map	26.34	0.344	0.158	0.179
[Cur]	28.56	0.292	0.138	0.111
[First, Cur]	28.77	0.285	0.128	0.092
[Prev]	28.47	0.310	0.136	0.126
w/o temporal	28.60	0.298	0.134	0.104
w/o target	27.32	0.322	0.153	0.141
Ours	**28.86**	**0.273**	**0.124**	**0.082**

6. Comparison with Concurrent Works

We further compare with the only two open-sourced **concurrent works** (DisCo [6] and MagicAnimate [8]) as shown at Table 2 and Figure 10. We observe ours outperforms theirs.

7. Impact of frame count

MotionEditor addresses motion-level editing that demands maintaining original camera movement and dynamic background. Table 3 and Figure 11 demonstrate the impact of the frame count. MotionEditor performs better as more frames are used. Moreover, we maintain the camera movement in original videos, while DisCo has a static background as it only takes single-image input.

8. Ablation for attention injection

We have conducted an ablation for potential designs. In Table 4, w/o mask, w/o temporal, and w/o target refer to our proposed attention mechanism without mask separation, temporal attention injection, and concatenation with

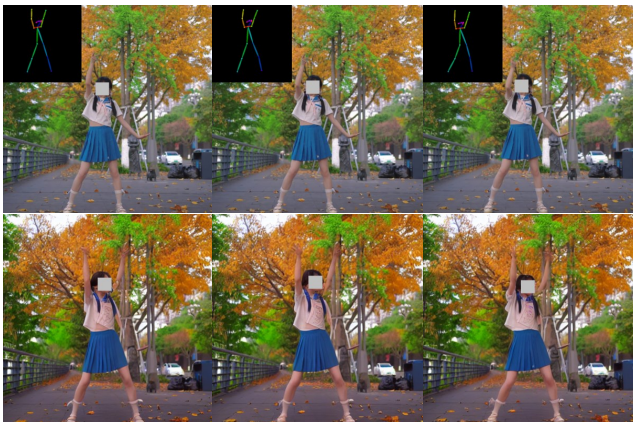

Figure 1. The failure case of our MotionEditor.

Table 5. Similarity of the CLIP feature in terms of the foreground and the background.

Method	Foreground-Sim (↑)	Background-Sim (↑)
Tune-A-Video	0.818	0.885
DisCo	0.741	0.705
MagicAnimate	0.903	0.728
MotionEditor	**0.947**	**0.972**

target key and value respectively. Attention Map indicates the direct attention map injection. [Cur], [First, Cur], and [Prev] refer to the modifications to Eq.8 in the main paper. Our proposed attention injection mechanism has the best performance.

9. Similarity metrics

We measure the similarity of the CLIP feature between edited results and ground truth separately for foreground and background. According to Table 5, the results highlight our effectiveness in both cases.

10. Limitations and Future Work

Figure 1 shows one failure case of our MotionEditor. The hands of the girl are confused with the surrounding background. The plausible reason is that the foreground latents are confused with background latents, thereby introducing additional bias to self-attention in the U-Net during the denoising process. The probable solution is to explicitly decouple the foreground and background before the denoising process and perform the corresponding editing operation on them. Moreover, a learnable dedicated mixture adapter can be designed to blend foreground with background naturally. This part is left as future work.

References

[1] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023. 1, 2

[2] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *ICCV*, 2019. 1, 2

[3] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 1, 2

[4] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *ICCV*, 2023. 1, 2

[5] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 1, 2

[6] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2023. 2

[7] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023. 1, 2

[8] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498*, 2023. 2

[9] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1, 2

[10] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 1, 2

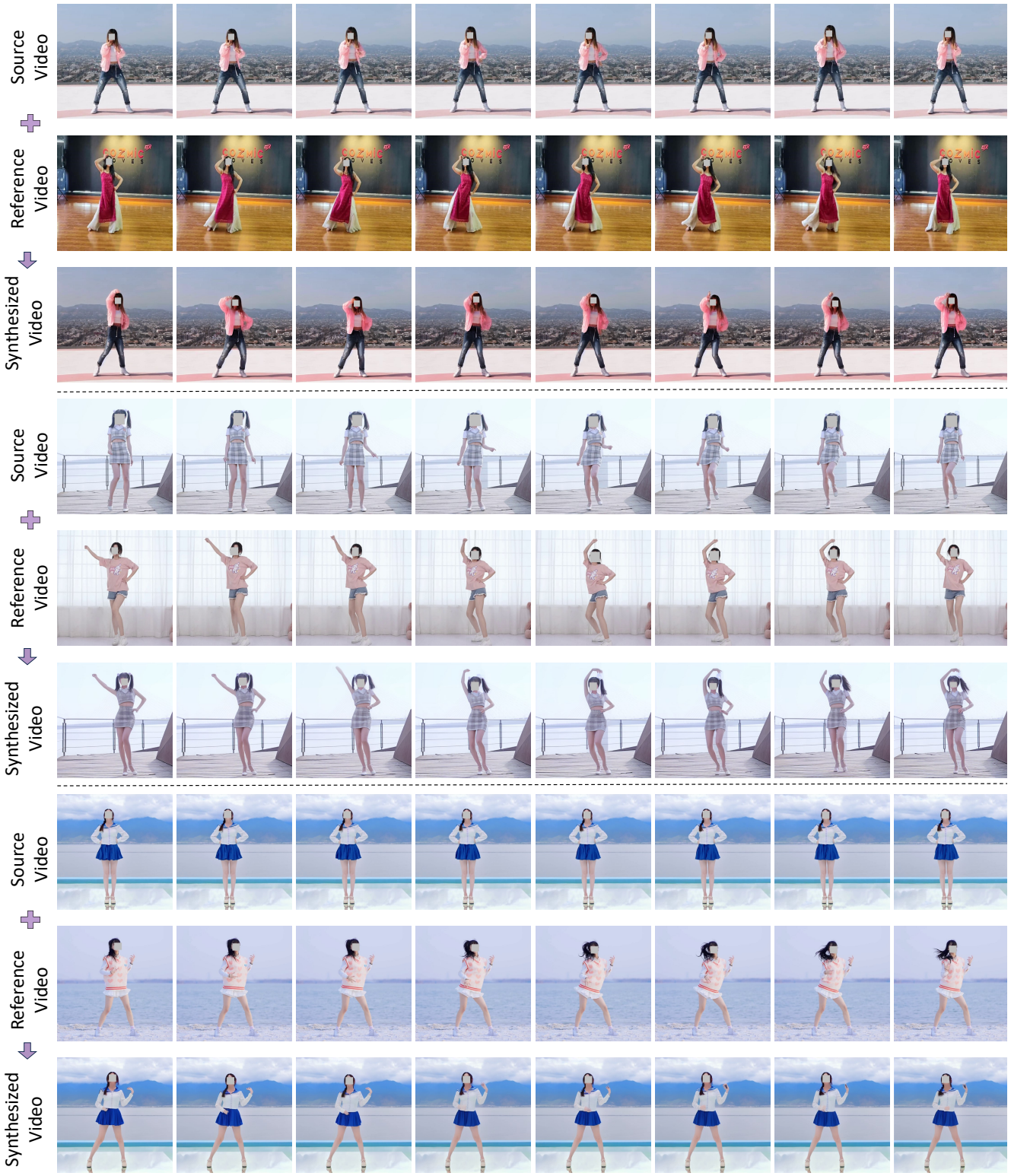

Figure 2. Additional video motion editing results (1/2).

Figure 3. Additional video motion editing results (2/2).

Figure 4. Complete video motion editing comparison results.

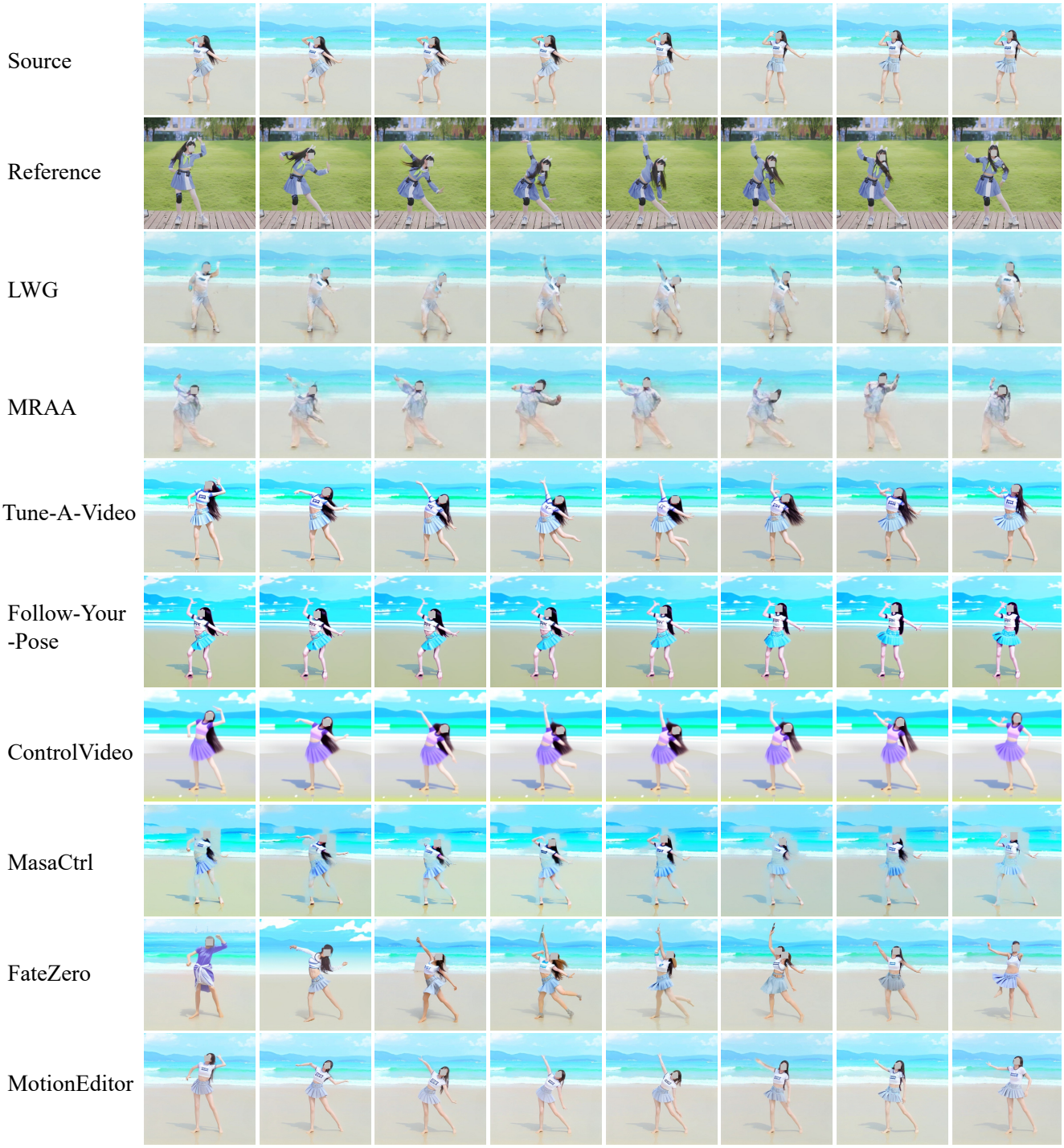

Figure 5. Additional video motion editing comparison results.

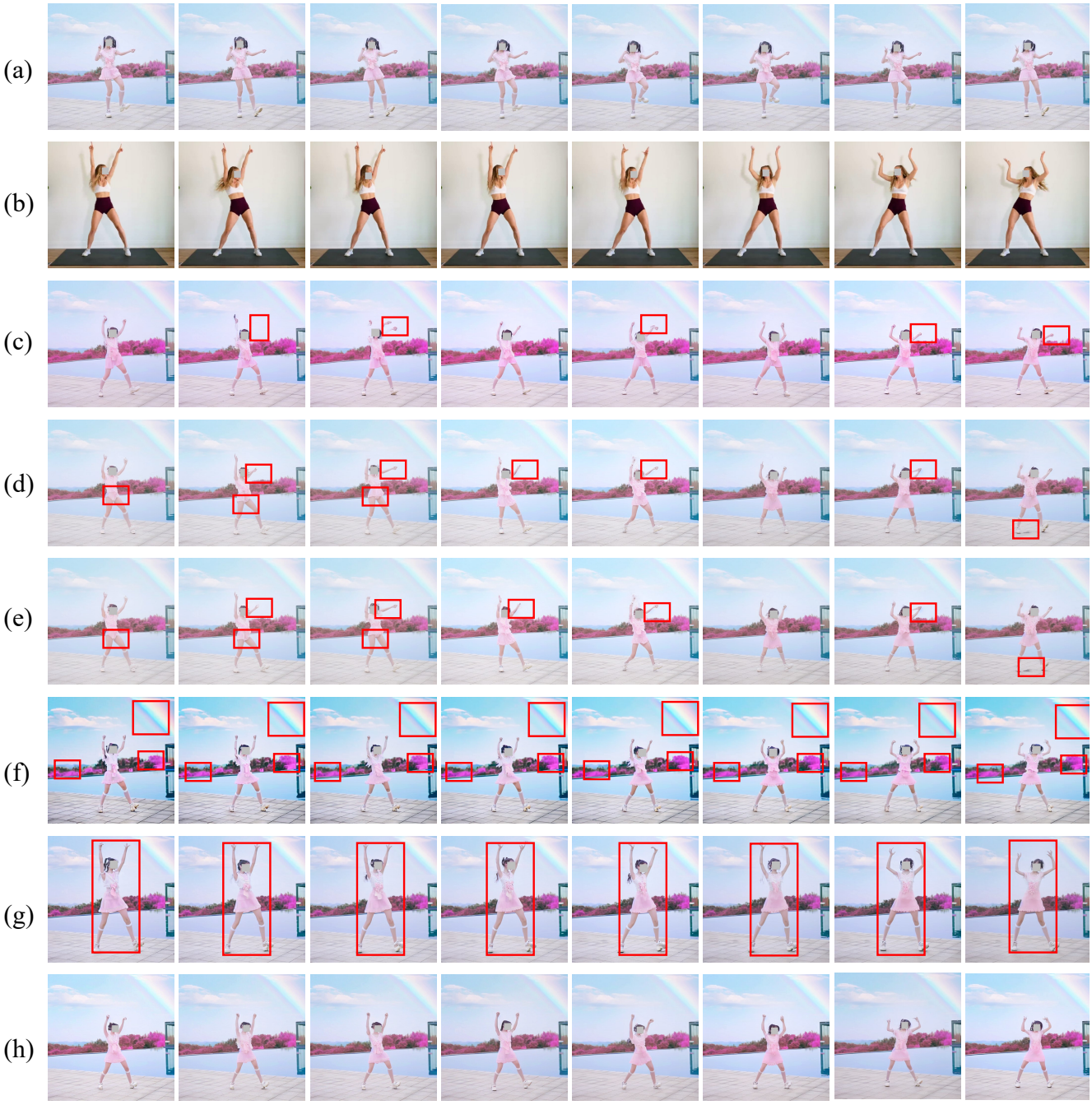

Figure 6. Additional ablation study results. Rows in the figure are: (a) source, (b) reference, (c) *w/o* CS Attention, (d) *w/o* cross attention in motion adapter, (e) *w/o* motion adapter, (f) *w/o* high-fidelity attention injection, (g) *w/o* skeleton alignment, and (h) MotionEditor.

Figure 7. Comparison between video motion editing and human motion transfer with dynamic backgrounds.

Figure 8. Comparison between video motion editing and human motion transfer with camera movement.

Figure 9. Comparison between video motion editing and human motion transfer with complex initial poses.

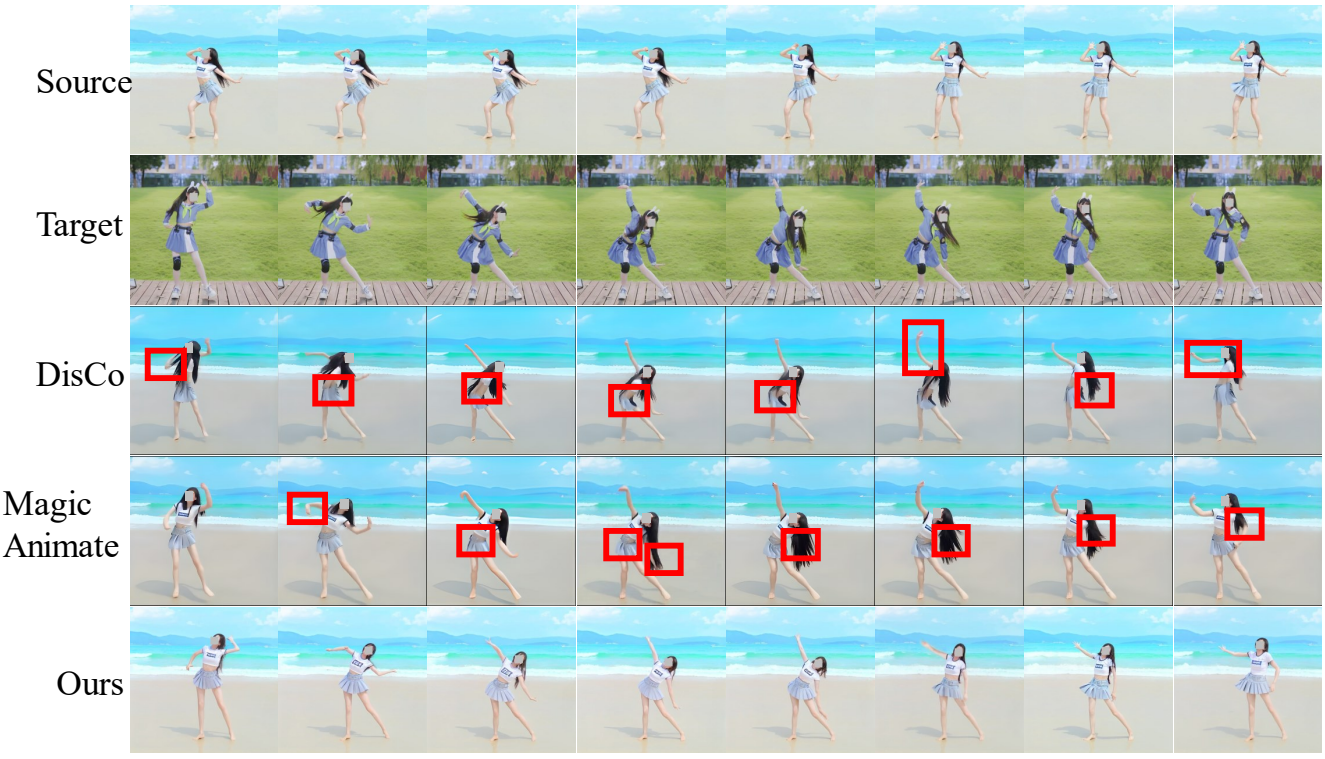

Figure 10. Qualitative comparisons with our proposed MotionEditor and concurrent works.

Figure 11. Qualitative validation of the impact of frame count.