Ponimator: Unfolding Interactive Pose for Versatile Human-human Interaction Animation

Shaowei Liu* Chuan Guo[2]† Bing Zhou[2]† Jian Wang[2]†
[1]University of Illinois Urbana-Champaign [2]Snap Inc.
https://stevenlsw.github.io/ponimator/

Abstract

The supplementary material provides implementation details, limitation analysis, qualitative results and future work. In summary, we include

- *Appendix A. Implementation details and model architecture of the interactive pose animator and generator.*
- *Appendix B. Limitation analysis of our current approach.*
- *Appendix C. Additional qualitative results of long interactive motion generation, complex interaction synthesis, two-person image animation, single-person image interaction generation, interactive pose animation, text-to-interaction motion synthesis, and single-pose-to-interaction motion synthesis.*
- *Appendix D. Future direction and potential applications of our work.*

A. Implementation details

Interactive pose extraction. Given a two-person pose from a motion sequence, we determine close contact by measuring the minimum distance between their SMPL-X meshe vertices. Following [14], we downsample the mesh based on predefined contact regions and compute pairwise distances. If the smallest distance is below 1.3cm, we classify the pose as a proximity pose—indicating contact between the individuals. This interactive pose is then used to train human interaction dynamics.

Model architecture. Our pose animator and pose generator follow the DiT architecture [16], which consists of stacked Transformer blocks [21], each incorporating an attention mechanism and a feed-forward network (FFN). Both the animator and generator comprise 8 Transformer layers, with the animator utilizing both spatial- and temporal-attention blocks, while the generator employs only spatial attention. The model has a latent dimension of 1024, with 8-head multi-head attention, and uses the GELU activation function. The input motions are first encoded with positional encoding before being processed by Transformer blocks. The input has the shape (B, P, N, D), where B is batch size, $P = 2$ represents the number of individuals, and N corresponds to number of frames, and D is the dimension of diffusion target $\mathbf{z}_0$. Spatial attention operates along the P-dimension to model interactions between individuals, while temporal attention captures motion dynamics along the T-dimension. The model's output layer is a linear MLP, initialized with zero weights, which generates residual motion outputs. These residual motions are added to the interactive pose to produce the final output. Conditional information is incorporated into the model using Adaptive Instance Normalization [9].

Training. We apply training data augmentation to interactive poses in the interactive pose animator by adding random noise with a scale of 0.02 to account for real-world inaccuracies in pose estimation. This ensures that even if the interactive pose estimator introduces noise, the animator can still produce reasonable results. This augmentation is performed online during training. Following prior work [5, 11], we align one person's pose in the interactive pose to face the positive Z direction and center it at the origin. The interaction loss in the pose animator follows [11] and consists of a **contact loss**, which encourages contact between two individuals when their joints are close, and a **relative orientation loss**, which aligns their global orientations with the ground truth. The velocity loss $\mathcal{L}_{\text{vel}}$, following MDM [20], ensures motion coherence by minimizing the velocity difference between the generated motion and the ground truth. For diffusion training, we use a cosine scheduler with 1000 diffusion steps and DDIM sampling [18] for 50 steps during inference. The model is trained with a learning rate of $1e$-4 and weight decay of 0.00002 for 4000 epochs. The batch size is 256 for the interactive pose animator and 512 for the interactive pose generator. Training takes 2 days for the pose animator and 1 day for the pose generator on 4×A100 GPUs.

*Work done at an internship at Snap Research NYC, Snap Inc.
†Co-corresponding author

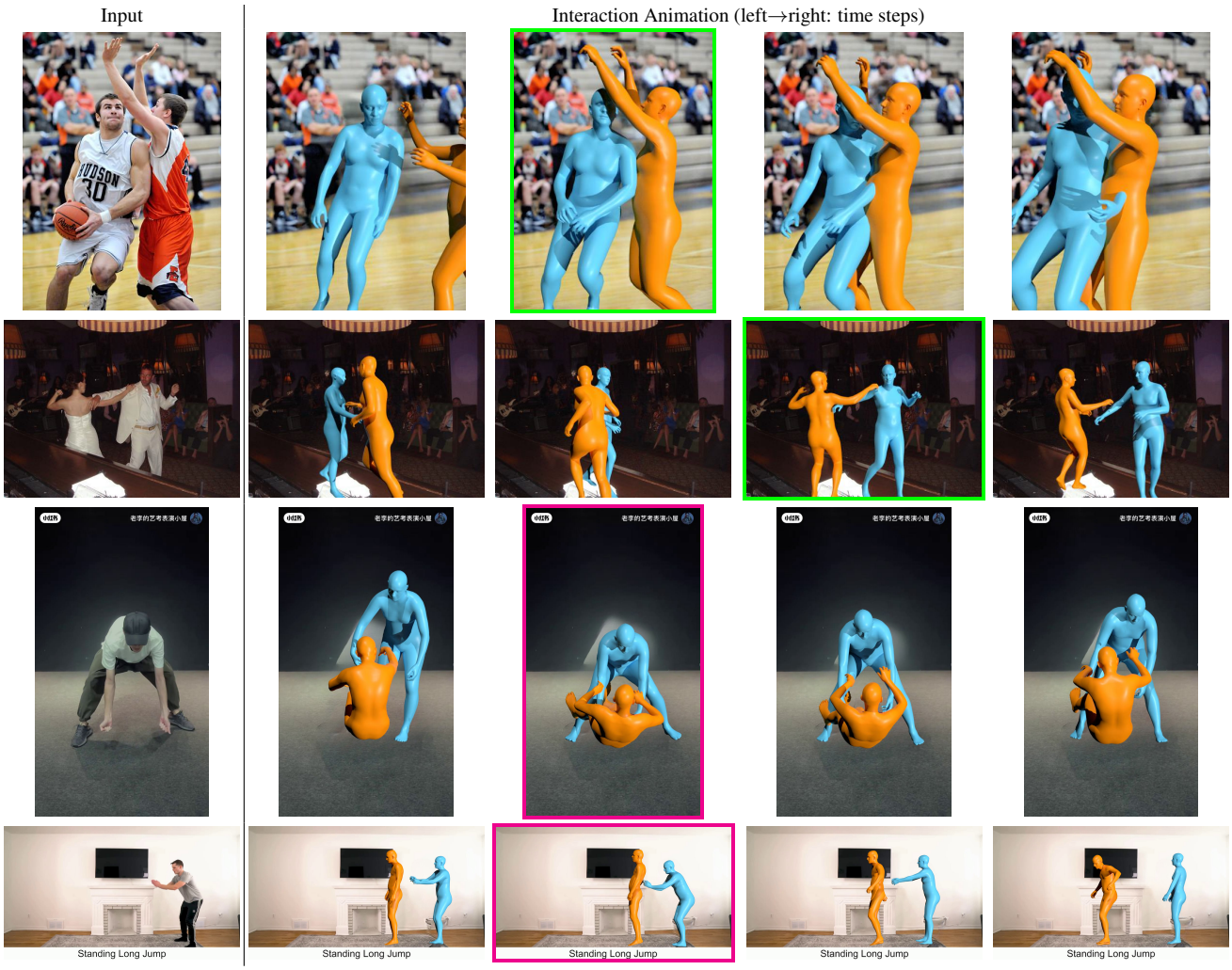


Figure 1. Method limitation analysis. The first two rows show in-the-wild interactive pose animation results. In the first sample, severe interpenetration occurs as our method does not explicitly model penetration between two individuals. In the second, the generated motion is physically implausible due to the lack of scene context awareness, leading to collisions with the environment. The bottom two rows illustrate interaction motion generation from a single pose input. Due to inaccuracies in interactive pose generation, our method fails to produce realistic contact, resulting in unnatural motion.

Inference speed comparison. Our interactive pose generation takes 0.21s on a single A100 on average, the interactive pose animator generates 3s motion at 10fps in 0.24s, comparable to InterGen [11] which requires 0.76s for the same motion length.

B. Limitation Analysis

Our method has the limitations below. The common failure modes are illustrated in Fig. 1.

Short motion modeling. Our method is mainly focus on short interactive motion segments. While our framework could support longer generation by interactive pose chaining as shown in Fig. 2, the benefit of interactive pose prior would diminish over time. In text-to-interaction synthesis, our framework prioritizes interactive motion-relevant information, which can result in partial rather than complete motion sequences when the input text describes extended human interactions. Moreover, our pose animator—taking only interactive poses as input—cannot fully capture the semantic context or temporal ordering in text (e.g., distinguishing "lifting up" from "putting down"). Incorporating text conditioning into the pose-to-interaction stage is a promising avenue for improving text-to-interaction–specific tasks. However, since our main focus is on pose-to-interaction animation without enforced text input, this ambiguity can be a strength, enabling multiple valid and physically plausible motion interpretations from the same interactive pose.

Inter-person penetrations. While our method enhances

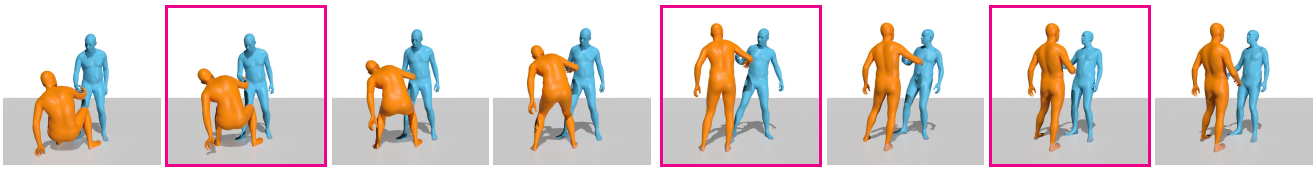


Figure 2. Longer motion generation by chaining interactive poses. We reuse the last generated pose as the next input, resetting interactive time to zero, enabling sliding-window synthesis of longer motions (key-frame in magenta box).

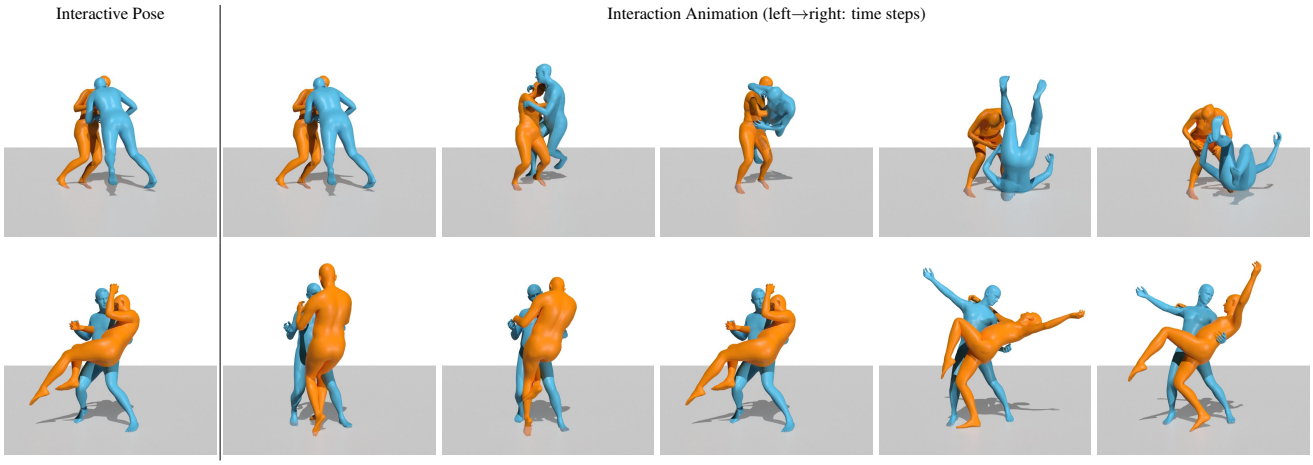


Figure 3. Complex interactive pose animation. Given an interactive pose, our pose animator can synthesize high-dynamics (1st row) and close-contact (2nd row) human-human motions, leveraging the strong interactive prior learned from high-quality mocap data.

contact in two-person interactions, it does not explicitly model interpenetration between individuals. Consequently, in close-contact scenarios—such as the first row in Fig. 1—some interpenetration may occur in the generated motion sequences. Achieving a balance between realistic contact and preventing interpenetration remains a challenging problem, as enforcing strict physical constraints could compromise natural motion quality. Addressing interpenetration modeling and ensuring physically plausible two-person interaction motion generation is an important direction for future work.

Lack of scene awareness. When applied to in-the-wild two-person pose animation or motion generation, our method relies solely on human pose information and ignores the surrounding environment. As a result, generated motions may appear physically implausible in certain cases, such as the 2nd row of Fig. 1, where collisions occur. Moreover, interactive poses can sometimes be ambiguous, causing noticeable motion errors when used as the sole input. A more robust approach would integrate additional scene information (e.g. image features) to improve motion prediction and dynamics forecasting.

Inaccurate contact. The interactive pose estimator or our interactive pose generator may occasionally produce inaccurate interactive poses, resulting in poor human–human contact in the generated motions, as seen in the 3rd and 4th rows of Fig. 1. These inaccuracies result in unrealistic motion due to the lack of precise interactive pose inputs. Since the pose animator primarily models temporal dynamics and depends on the interactive pose for spatial information, it often cannot correct errors arising from inaccurate interactive poses. Additionally, our generated interaction motions may exhibit artifacts such as foot sliding, a common issue in human motion synthesis. While such artifacts can often be mitigated through post-processing, we do not apply any post-processing in our examples.

C. Qualitative results

Longer interactive motion generation. Our framework is designed for short-term interaction generation but naturally extends to longer sequences. The pose animator takes an interactive pose together with an interactive time to synthesize both past and future motions centered on that pose. Longer sequences are produced by chaining segments in a sliding-window manner: the last generated pose of one segment is reused as the starting pose for the next, the interactive time index is reset to zero (beginning of the new segment), and generation continues. Repeating this process yields coherent long-term interactions, as shown in Fig. 2, where key-frames are labeled in magenta box.

Complex interactive pose animation. As shown in Fig. 3, beyond daily motions, our pose animator can synthesize complex interactive motions involving high dynamics (1st row) and close contact (2nd row) between two people, benefiting from the strong interaction dynamics learned from

Figure 4. Interactive pose image animation on FlickrCI3D dataset [4]. Left shows the input image, right shows the animated interaction motions. Interactive-pose frame is labeled in green box. Our model generalizes well to in-the-wild interactive poses, producing realistic human-human interaction dynamics.

high-quality mocap data.

Two person image human motion animation. We provide additional in-the-wild interactive pose animation results in Fig. 4. Given an interactive frame, we extract two-person poses using an off-the-shelf model [14], and animate the them with our interactive pose animator. To render the interaction, we use an off-the-shelf inpainting model [19] to remove the original individuals and overlay the generated

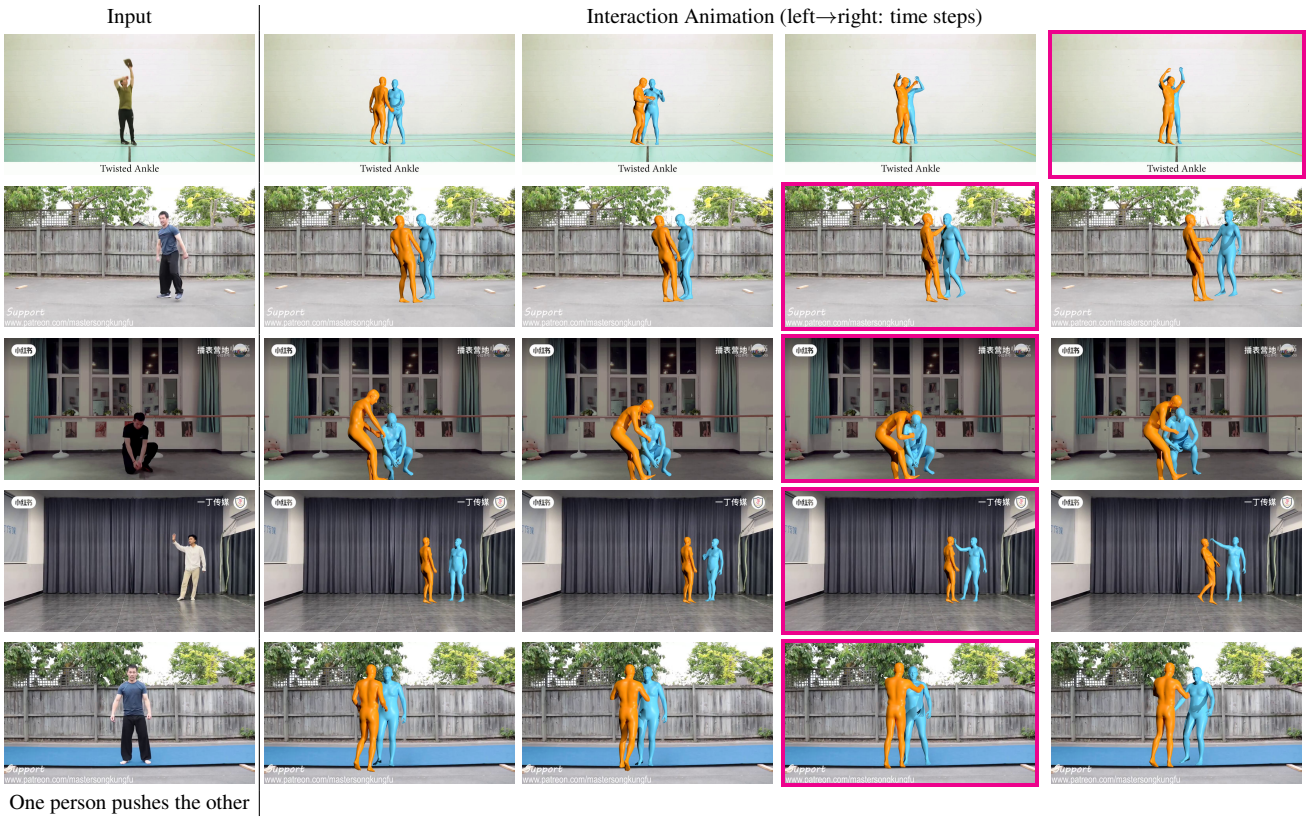


Figure 5. Single-person pose interaction generation on Motion-X dataset [12]. Left shows the single person image input, right shows the generated two-person interaction dynamics. The generated interactive pose frame is labeled in magenta box. The bottom row show the single-pose input with accompanying text input. Given different single-person poses, our interactive pose generator produces plausible interactive poses under flexible conditions, while our interactive pose animator synthesizes realistic human-human motions. Our model demonstrates strong performance in challenging in-the-wild settings.

motion. The results demonstrate that our model generalizes well to in-the-wild interactive poses, producing realistic human-human interactions.

Single-person image human motion interaction generation. We present additional single-person image interaction motion generation results on the Motion-X dataset [12] in Fig. 5. Given a single-person image, we first extract the pose using an off-the-shelf pose estimator [2] and then generate interactive poses with our interactive pose generator. As shown, our model synthesizes plausible interactions from diverse single-person inputs. Finally, we apply our interactive pose animator to generate two-person dynamics, demonstrating its effectiveness in challenging in-the-wild scenarios.

Interactive pose animation. We provide additional visualizations of interactive pose animation on the Inter-X dataset [22], Dual-Human dataset [3], and Duolando dataset [17] in Fig. 6. Our model could successfully synthesize realistic dancing motions from out-of-domain interactive poses on the unseen Duolando dataset.

We further evaluate our method on the InterHuman dataset [11], a more challenging out-of-distribution setting, with results presented in Fig. 7. The InterHuman dataset provides SMPLH [13] annotations for two-person interactions, primarily for text-to-motion synthesis. However, the annotated motions exhibit less accurate contact compared to other datasets. To align with our framework, we convert the provided SMPLH [13] representation to SMPLX [15] and extract interactive poses from the test motion sequences. Despite inherent contact inaccuracies due to the dataset's annotation conventions and diverse pose distributions, our model successfully synthesizes realistic interaction motions, demonstrating the strong generalization capability of the interactive pose prior for human interaction animation.

Furthermore, we present a qualitative comparison against two baselines (InterGen*, random-pose in Tab. 2 of the main paper) in Fig. 8. As shown, InterGen [11] and the model trained with random-pose shows less accurate contact and more body penetration than ours, underscoring the importance of interactive pose priors for realistic contact modeling and interaction generation.

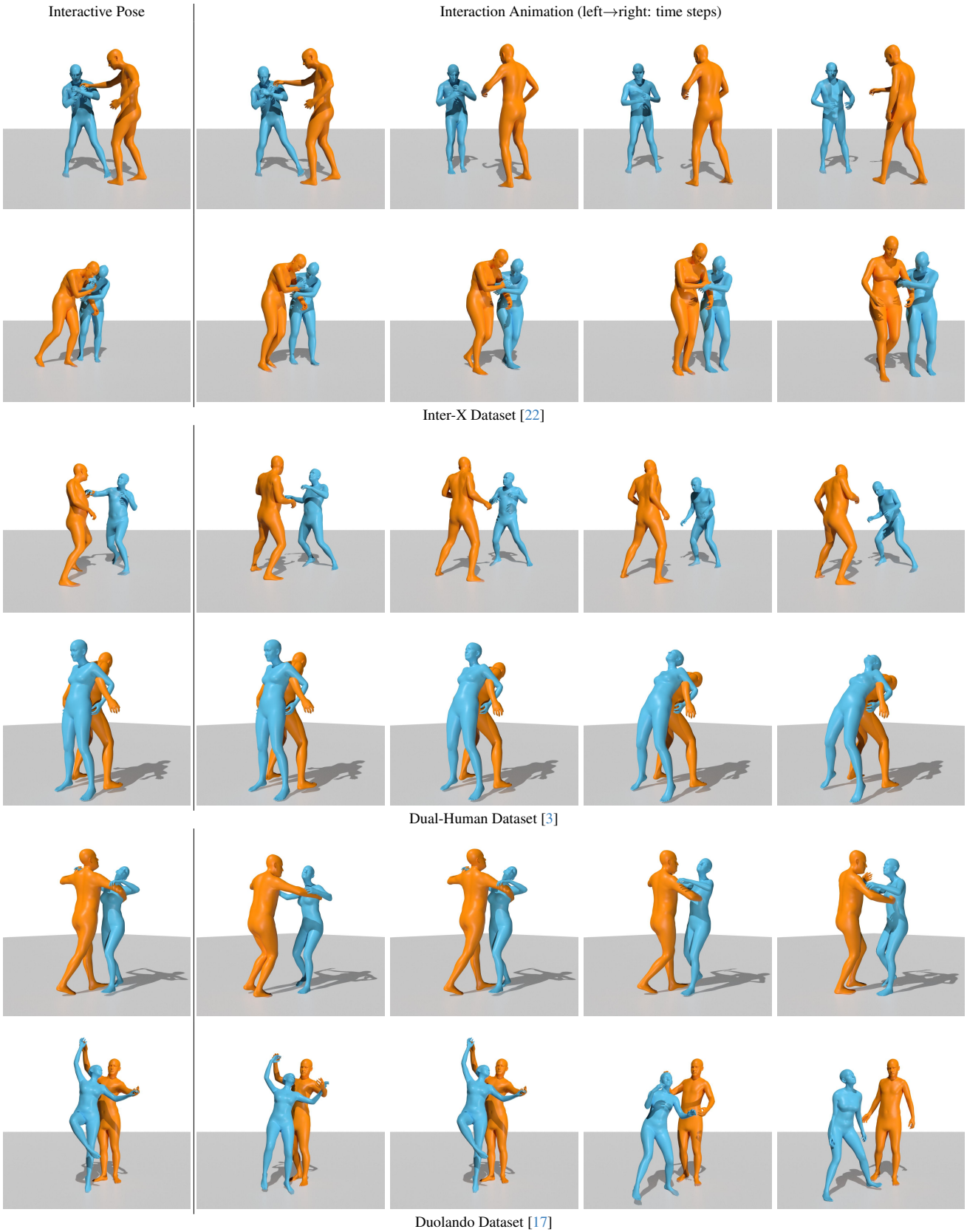


Figure 6. More interactive pose animation visualization on Inter-X dataset [22], Dual-Human dataset [3], Duolando dataset [17]. Our pose animator generalizes well to out-of-domain interactive poses and synthesizes realistic dancing motions on the unseen Duolando two-person dancing motion dataset.

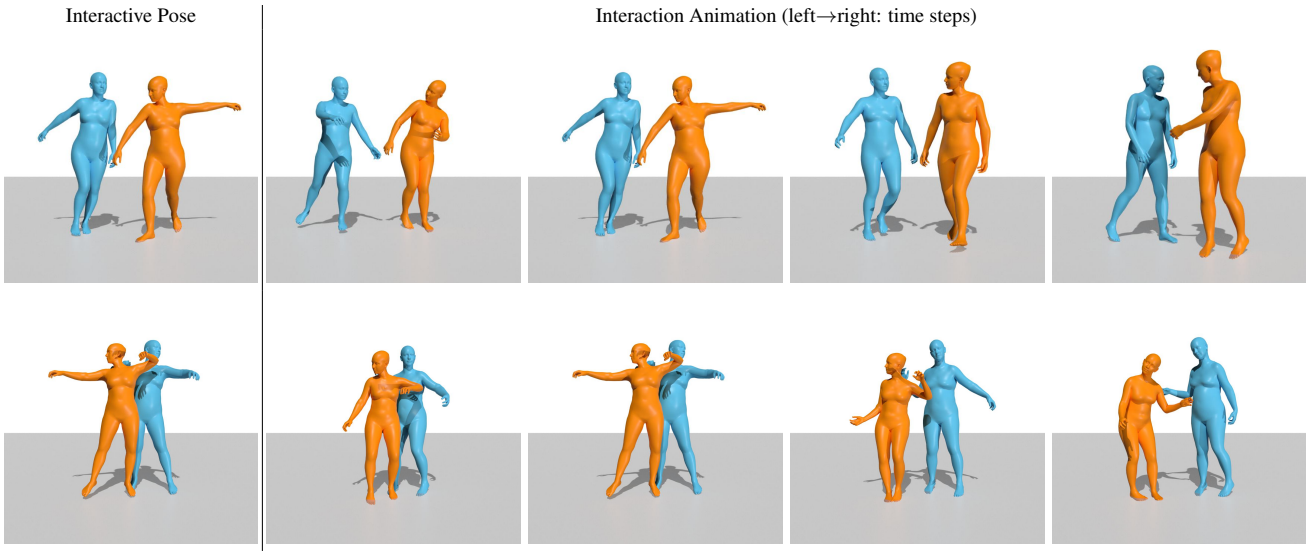


Figure 7. Interhuman dataset [11] interactive pose animation results. We convert dataset provided SMPLH [13] to SMPLX [15] representation and select interactive poses from test motion sequences. Despite contact inaccuracies due to dataset conventions and pose variations, our model synthesizes reasonable motions, demonstrating the strong generalization capability of interactive poses for guiding human interaction animation.

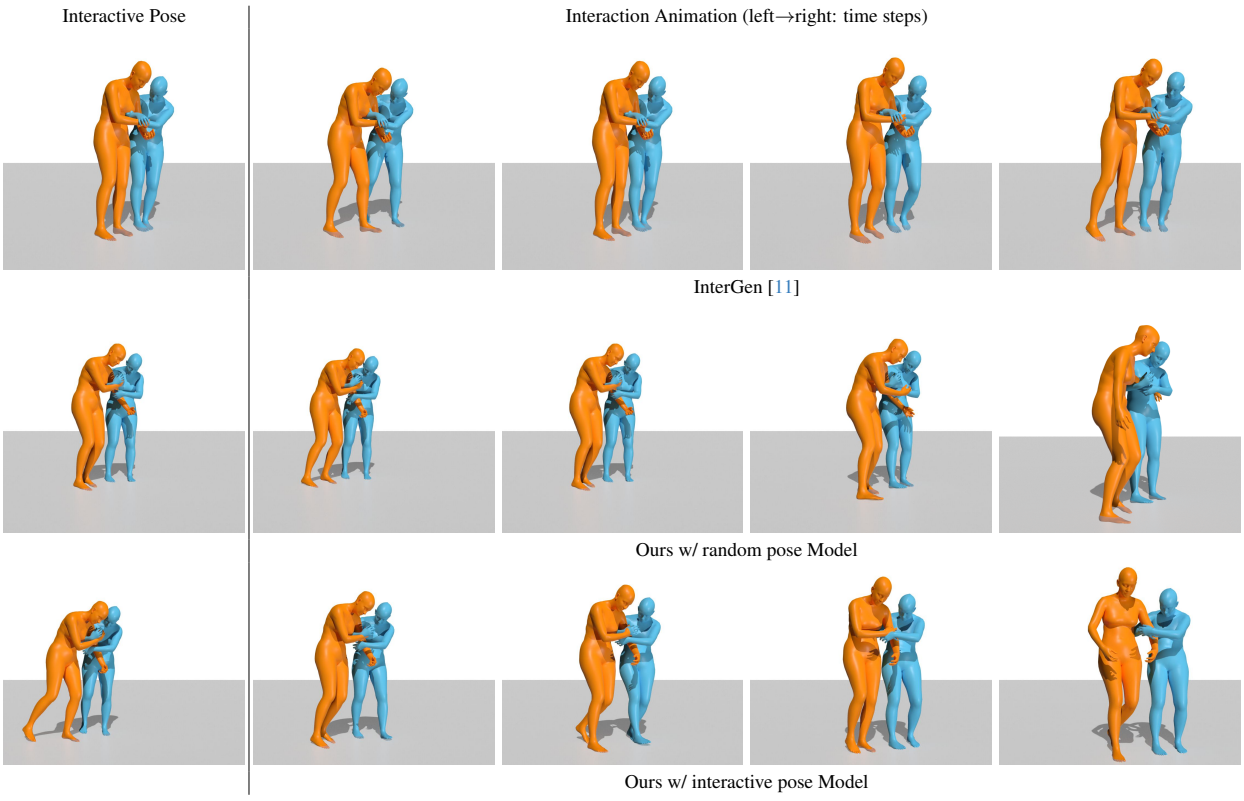
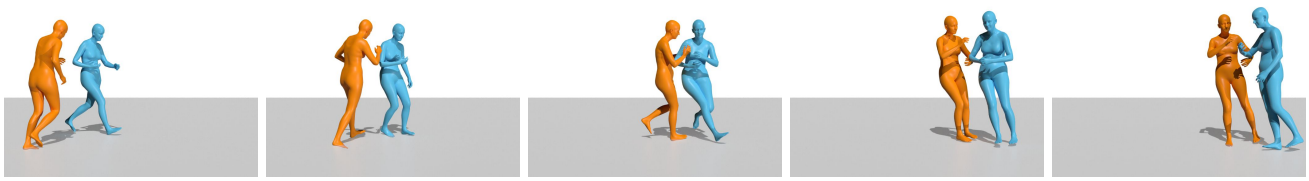
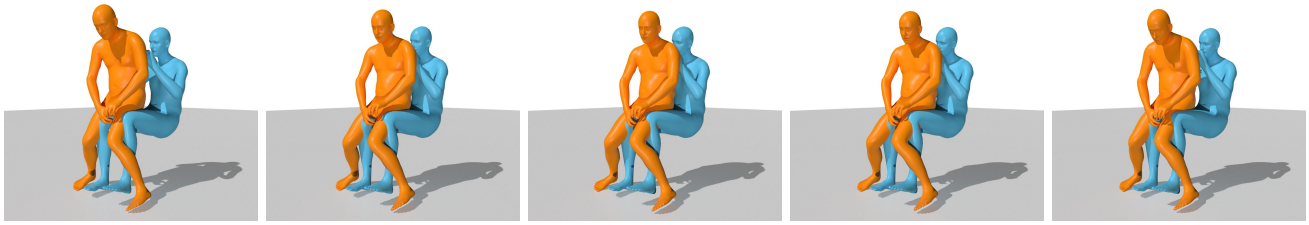


Figure 8. Interactive pose animation comparison on Inter-X dataset [22]. Compared to InterGen [11] and model trained with random poses, our method achieves better contact and human dynamics. Both baselines exhibit severe body penetration and less accurate contact, while our approach, guided by interactive poses, ensures more realistic interactions.

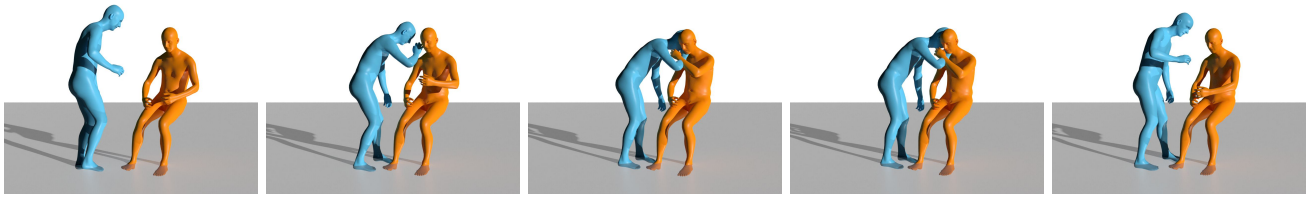
Text-to-interaction synthesis. We present additional text-to-interaction motion synthesis results in Fig. 9. Our method effectively generates realistic two-person interactions from short phrases or simple words. By leveraging an



Input Text: One person chases the other person

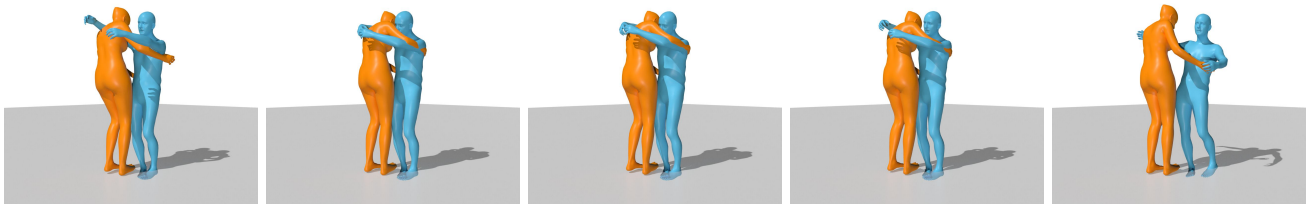


Input Text: One person sits down first, another sits on his/her lap

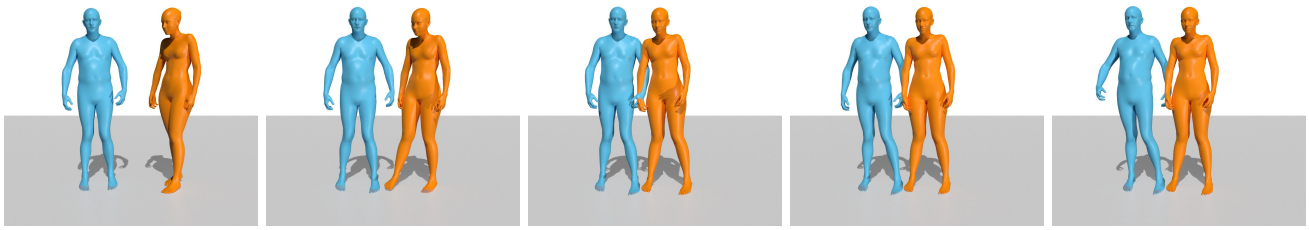


Input Text: One person goes to the other person's ear and whispers to him/her

Input Text: hand shake



Input Text: hug



Input Text: posing

Figure 9. More text-to-interaction motion synthesis results. Our method synthesizes realistic two-person interactions from short phrases or single words.

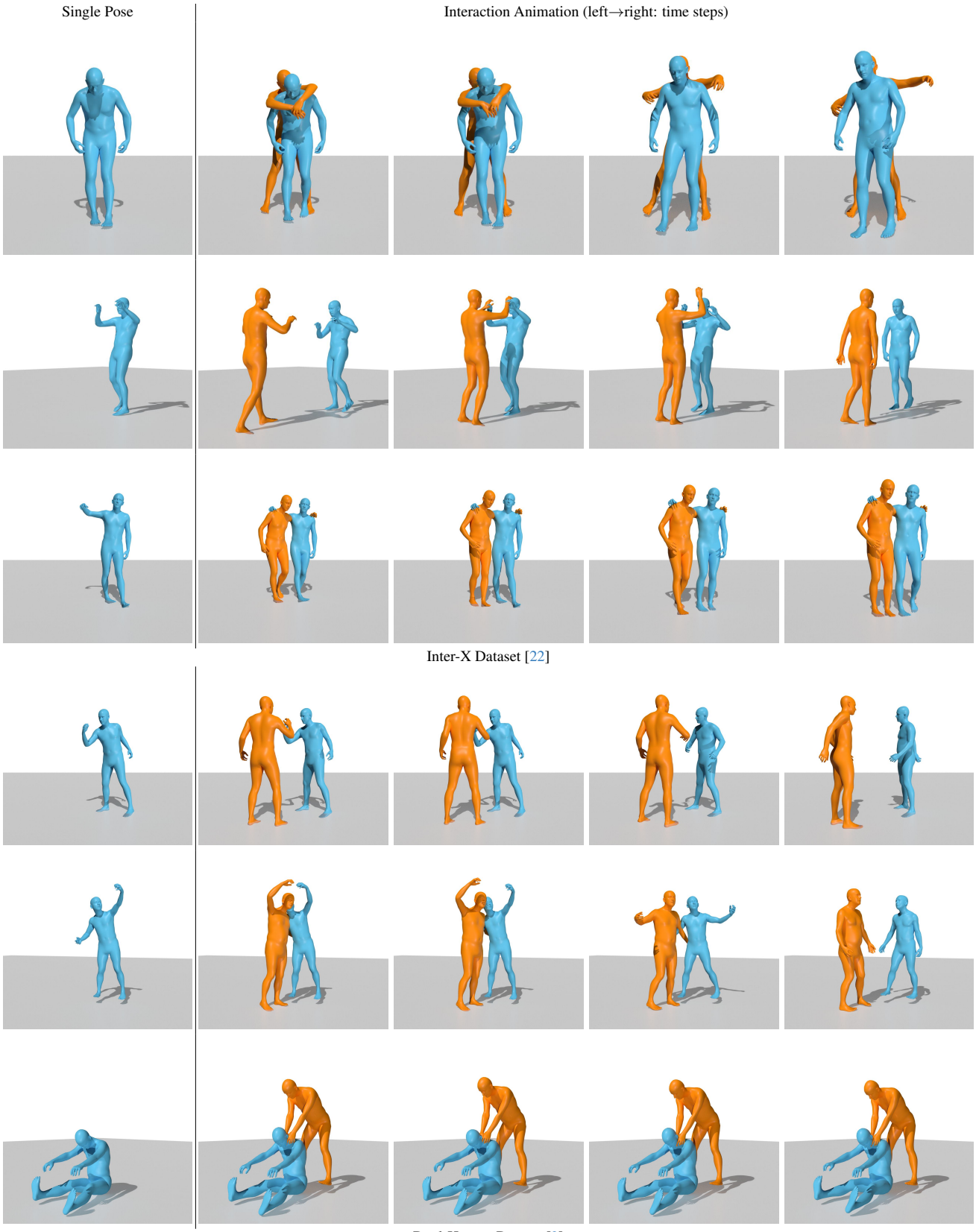


Figure 10. Single-pose guided interaction motion synthesis result on Inter-X [22] and Dual-Human [3] datasets. The input single-person pose is shown on the left. Our method generates appropriate interactive poses from various inputs, capturing vivid underlying human dynamics.

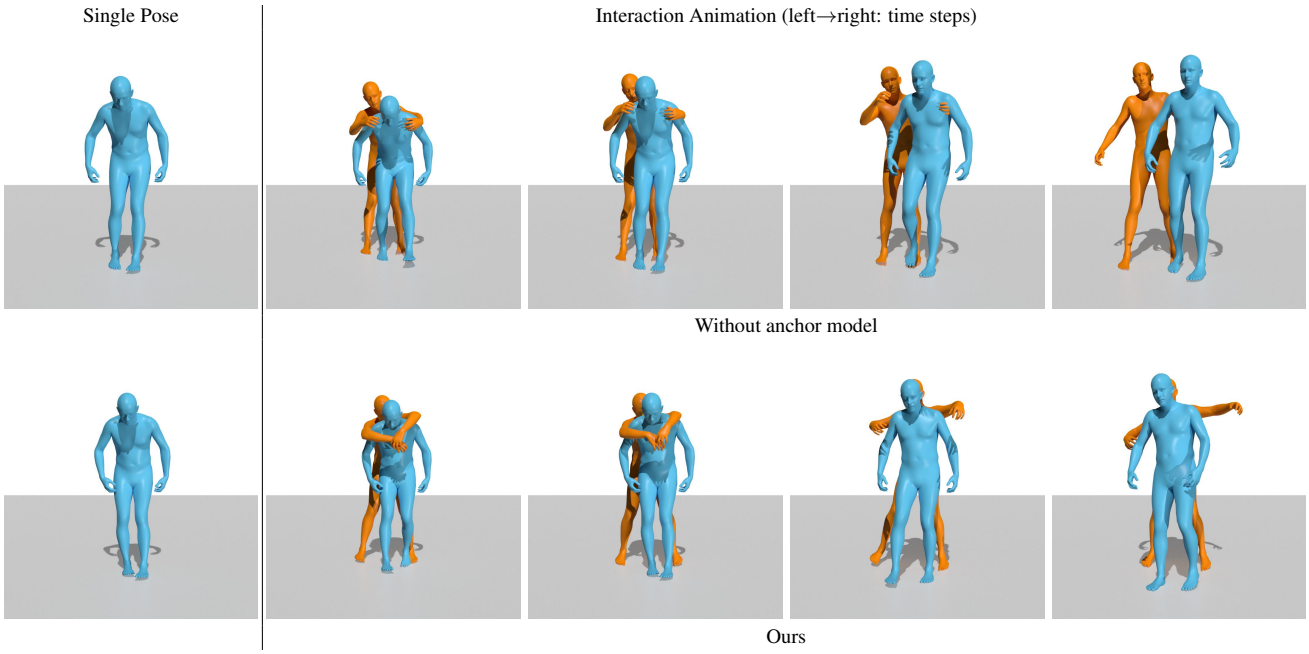


Figure 11. Single pose-to-interaction motion synthesis comparison on Inter-X dataset [22]. Compared to the end-to-end without interactive pose as anchor (w/o anchor) model, our method synthesizes more realistic interactive poses, leading to more natural human interactions.

Figure 12. Generated two-person interaction videos from state-of-the-art video diffusion model [10]. The generative video suffers from unrealistic and temporally inconsistent interactive motions.

intermediate interactive pose representation, our approach ensures consistent interaction and maintains accurate contact between the two individuals.

Single pose-to-interaction motion synthesis. We present single pose-to-interaction motion synthesis results on the Inter-X [22] and Dual-Human [3] datasets in Fig. 10. As shown, our method generates appropriate interactive poses from various input poses while effectively capturing vivid underlying human dynamics. A qualitative comparison with the end-to-end baseline without interactive pose as anchor (Tab. 4 of main paper) is provided in Fig. 11.

D. Future Work

As discussed in the Introduction of the main paper, existing video diffusion models [1, 6, 7, 10] can generate human images over time; however, the resulting motions often lack temporal consistency and realism, as shown in 12, where we applied state-of-the-art image-to-video diffusion models to generate videos from a human interaction image in Fig. 1 of the main paper. The generated videos suffer from unrealistic and inconsistent motions. A promising application of our work is to use the generated motion as a conditioning signal for pose-guided human video diffusion models [8, 23, 24]. By providing high-quality motion input tailored to the input image, this approach could significantly enhance the realism of the generated motion. We consider this an exciting direction for future research.

References

[1] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 10

[2] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *NeurIPS*, 2023. 5

[3] Qi Fang, Yinghui Fan, Yanjun Li, Junting Dong, Dingwei

Wu, Weidong Zhang, and Kang Chen. Capturing closely interacted two-person motions with reaction priors. In *CVPR*, 2024. 5, 6, 9, 10

[4] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, 2020. 4

[5] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022. 1

[6] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 10

[7] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022. 10

[8] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, pages 8153–8163, 2024. 10

[9] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 1

[10] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 10

[11] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *IJCV*, 2024. 1, 2, 5, 7

[12] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *NeurIPS*, 2024. 5

[13] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *TOG*, 2015. 5, 7

[14] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. In *CVPR*, 2024. 1, 4

[15] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 5, 7

[16] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 1

[17] Li Siyao, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. *arXiv preprint arXiv:2403.18811*, 2024. 5, 6

[18] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[19] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, 2022. 4

[20] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 1

[21] A Vaswani. Attention is all you need. *NeurIPS*, 2017. 1

[22] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *CVPR*, 2024. 5, 6, 7, 9, 10

[23] Jingyun Xue, Hongfa Wang, Qi Tian, Yue Ma, Andong Wang, Zhiyuan Zhao, Shaobo Min, Wenzhe Zhao, Kaihao Zhang, Heung-Yeung Shum, et al. Follow-your-pose v2: Multiple-condition guided character image animation for stable pose control. *arXiv preprint arXiv:2406.03035*, 2024. 10

[24] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *ECCV*, 2024. 10