

Interactive Humanoid: Online Full Body Human Motion Reaction Synthesis With Social Affordance Forecasting and Canonicalization

Supplementary Material

A. Dataset Quality Visualizations

To demonstrate the quality of our dataset, we provide extensive visualizations. The figures below show sample data visualizations for CoChair and HHI, illustrating that our dataset maintains high quality in both contact and non-contact situations. **Additionally, for easier and more detailed inspection, we have saved numerous PLY files in the attached ZIP file.**



Figure 1. Quality Visualizations for HHI Dataset.

B. Dataset Construction for HHI and CoChair

Dataset Collection. We use 12 NOKOV motion capture cameras with 60fps for data collection. The cameras are arranged in a square formation on the ceiling, with four cameras on each side. For the body, each participant has to attach 53 markers. As for the hands, we attached 32 markers to the wrists, finger joints, and fingertips (16 markers per hand). For tracking objects during interactions, we affix 9

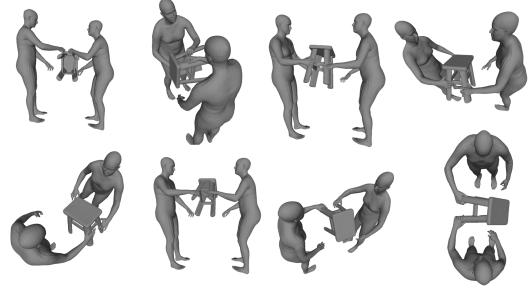


Figure 2. Quality Visualizations for CoChair Dataset.

markers to each object. During the interaction, only the actor knows the task and we allow them to decide whether or not to share this information with the reactor. Then the actor decides when to start the carrying process. This ensures the authenticity of the reaction when it responds. The actor initiates the motion first, and the reactor will respond accordingly.

Dataset annotation. We track 53 markers on the body and 16 markers on each hand. Followed by AMASS[7], we obtain SMPL-X[8] parameters for each frame through optimization. For each object, we mark the position of the attached marker on the scanned CAD model and calculate the transformation.

HHI: Human-Human Interaction. The HHI dataset is a large-scale full-body motion reaction dataset with clear action feedback which includes 30 interactive categories, 10 pairs of human body types, and a total of 5,000 paired interaction sequences. The HHI dataset has three characteristics. The first is *multi-human whole-body interaction* including body and hand interactions which is not supported in previous datasets. We believe hands are essential in multi-human interactions and can convey rich information during handshake, hug, and handover. The second is our dataset can distinguish *distinct actors and reactors*. For example, during handshakes, pointing directions, greetings, handovers, etc., We can identify the initiator of the action, which can help us better define and evaluate this problem. The third is our dataset has a *rich diversity*, which includes the types of interactions and reactions. We not only include 30 types of interactions between two humans but also provide multiple reasonable reactions for the same action by the actor. For example, when someone greets you, you can nod in response, respond with one hand, or respond with both hands. This is also a natural feature that humans possess when



Figure 3. We construct two datasets to support the research on reaction synthesis. HHI is the first large-scale whole-body motion reaction dataset with clear action feedback. CoChair is the first large-scale dataset for multi-human and object interaction. **Note:** Please refer to the demo data in the supplementary materials to check the annotation quality.

reacting, but previous datasets have rarely focused on this point and discussed it.

CoChair: Human-Object-Human Interaction. CoChair is a large-scale dataset for multi-human and object interaction consisting of 8 different chairs, 5 interaction patterns, and 10 pairs of different skeletons, totaling 3000 sequences. CoChair has two significant characteristics. CoChair is information asymmetry during collaboration. Each action has an actor/initiator (who knows the destination of the carrying) and a reactor (who does not know the destination), which can support our research on the reactor’s behavior patterns. The second is the diversity of carrying patterns. The dataset includes five carrying patterns: one-hand fixed carrying, one-hand mobile carrying, two-hand fixed carrying, two-hand mobile carrying, and two-hand flexible carrying.

C. Additional Generated Results

To more intuitively demonstrate the quality of our method, we provide additional generated results. The blue one is the actor, while the red one is the generated reactor. **For clearer demonstration, we have also saved numerous result videos in the ZIP file.**

D. Detailed Settings on Each Datasets

CoChair Datasets. We use 2000 sequences of data as the training set and 1000 sequences as the test set. The one initiating the motion is the actor and he/she knows the carrying destination, while the other person is the reactor.

HHI Datasets. We use 3000 sequences of data as the training set and 2000 sequences as the test set. The one initiating the action is the actor, while the other person is the reactor. It includes a total of 30 action categories.

InterHuman Datasets.[5] According to the official splits, we have selected 5200 sequences as the training set and 1177 sequences as the test set. Since there is no obvious initiator of actions in this dataset, we assume that the first person is the actor and the other person is the reactor. The

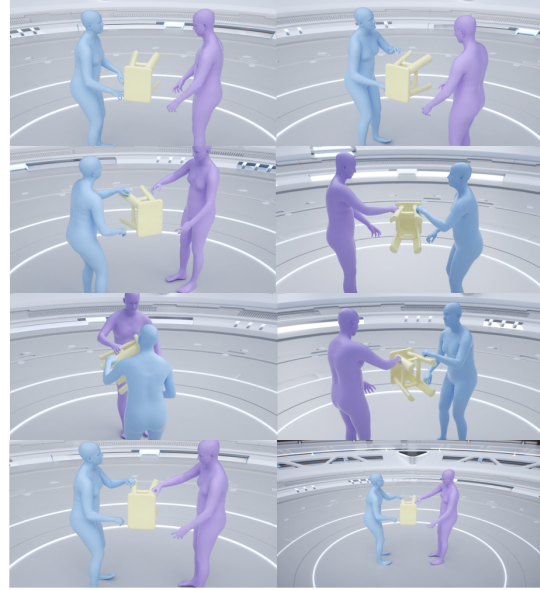


Figure 4. Result Visualizations for CoChair Dataset. The blue one is the actor, while the red one is the generated reactor.

dataset does not have fixed action categories, and each action corresponds to several textual descriptions.

Chi3D Datasets.[4] We have a total of 373 available data provided by officials. Among them, 257 are used as the training set and 116 as the testing set. We consider the person estimated from images as actors and the other one captured by motion capture devices as reactors.

E. Details of the baseline method

For all baseline methods, we entirely used the author’s code or made some modifications to adapt it to our task.

Progressively Generating Better Initial Guesses[6] uses Spatial Dense Graph Convolutional Networks and Temporal Dense Graph Convolutional Networks.

Spatio-temporal Transformer[1] is a transformer-based

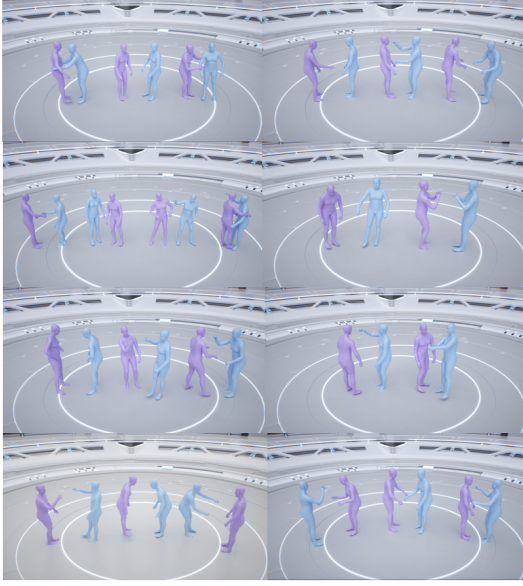


Figure 5. Result Visualizations for HHI Dataset.

architecture that uses attention to find temporal and spatial correlations to predict human motion.

InterFormer[3] consists of a Transformer network with both temporal and spatial attention to capturing the temporal and spatial dependencies of interactions.

InterGen-revised[5] is a powerful diffusion-based framework that can generate multi-human interaction based on the text description. We replace the CLIP branch with a spatio-temporal transformer to encode the actor’s motion.

F. Visualizations on Interhuman and Chi3D datasets

We show that our method still can generate reasonable reactions on existing human-human interaction datasets.



Figure 6. Visualization gallery of our method on InterHuman (left) and Chi3D(right). The deep black one is generated by our method.

G. More Details on Motion Forecasting Module

As shown in Fig.7, at the training stage, the humanoid reactor can access all motions of the actor. At the prediction stage in the real world, the humanoid reactor can only observe the past motions of the human actor. The forecasting module can anticipate the motions that the human will take.

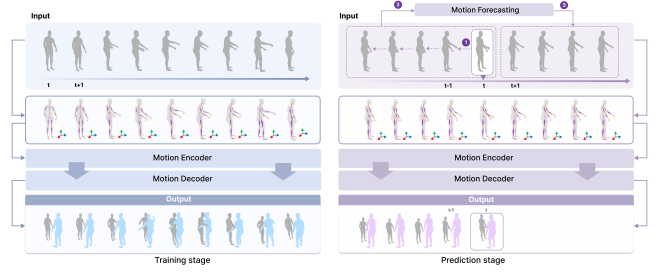


Figure 7. Social Affordance Forecasting.

We also visualized the results of motion forecasting as shown in Fig.8. Our method can provide reasonable predictions of whether there are objects or not. The imagination of these future behaviors can help the humanoid reactor to give more prompt and reasonable reactions. The left side shows the results generated by InterDiff[9], while the right side shows the results generated by HumanMAC[2].



Figure 8. Visualization results of Motion Forecasting with and without object.

H. Limitations

In the future, we will continue to enrich the categories of interactive objects to support research on generalization. Same as other kinematics-based methods, our methods also encounter problems such as floating feet, sliding, penetration, and other non-physically plausible problems. Moreover, our method does not separately consider hand movements. A straightforward solution would be to use a hierarchical generation strategy to separately handle body and hand movements, which is a potential future direction.

References

- [1] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, pages 565–574. IEEE, 2021. 2
- [2] Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. Humanmac: Masked motion completion for human motion prediction. *arXiv preprint arXiv:2302.03665*, 2023. 3
- [3] Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. Interaction transformer for human reaction generation. *IEEE Transactions on Multimedia*, 2023. 3
- [4] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7214–7223, 2020. 2
- [5] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023. 2, 3
- [6] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6437–6446, 2022. 2
- [7] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 1
- [8] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 1
- [9] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023. 3