ReGenNet: Towards Human Action-Reaction Synthesis *Appendix**

Liang Xu[1,2] Yizhou Zhou[3] Yichao Yan[1†] Xin Jin[2†] Wenhan Zhu[1] Fengyun Rao[3]
Xiaokang Yang[1] Wenjun Zeng[2]

[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China
[2]Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[3]WeChat, Tencent Inc.

https://liangxuy.github.io/ReGenNet/

A. Details of building the datasets

In this section, we provide more details of the definitions of asymmetric actions and the actor-reactor order labeling process. Taking the Chi3D-AS dataset as an example, it contains 8 action categories, *i.e.*, "Grab", "Handshake", "Hit", "HoldingHands", "Hug", "Kick", "Posing" and "Push". For all these actions, there always exists a person acts and the other reacts. Even for the interaction of "Handshake", a man actively reaches out his hands and the other one consequently reaches out his hands to complete the handshaking.

To annotate the actor-reactor order of the NTU120 and Chi3D datasets, we first build an annotation tool based on Python3, tkinter and tkVideoPlayer. Then we render the motion sequences of NTU120 and Chi3D to videos. The annotation tool displays the motion sequence videos together with the corresponding RGB videos. The annotators select the actor-reactor order of the given two-person interactions based on both the motion and the RGB videos. After the initial annotation process, the results are double-checked manually. The failed extracted motion sequences by PyMAF-X [14] for NTU120 are removed from the final datasets.

B. The choice of the datasets

We choose Chi3D [2] for its ground-truth SMPL-X parameters with subtle hand gestures, which is suitable for the **detailed** feature of human-human interactions. The NTU120 dataset [8] is chosen for that it is currently the largest human-human interaction dataset with 26 action categories and diverse reaction patterns. Take the action "kick" as an example, the reaction could be "move back", "raise the feet", "jump backward" or "turn the shoulders". Moreover, NTU-X [12] also verifies that it is feasible to extract SMPL-X parameters for NTU120, yet NTU-X is not open-source. The InterHuman dataset [7] is suitable for our setting, while it only contains human body parameters without dexterous hand movements.

C. Effects of pose estimation results

To *quantitatively measure the effects brought by the noise of the pose estimation models*, we adopt the PyMAF-X to extract the SMPL-X parameters from the pure RGB videos of the **Chi3D dataset** and then use the estimated pose results for reaction generation. The experimental results in Tab. C.1 show consistent improvements of our proposed ReGenNet over the baselines.

D. The technical details of the baselines

For the Chi3D-AS and NTU120-AS datasets, we choose cVAE [6], AGRoL [1], MDM [11], and MDM-GRU [11] as our baselines for that they are state-of-the-art methods for human scene/object interaction generation and conditional human motion generation. We adopt the 6D rotation representation of the SMPL-X parameters as inputs and the frame length is set to 60 and 150 for NTU120-AS and Chi3d-AS.

1) cVAE: Conditional VAE framework is widely adopted in human scene/object/human interaction generation. We adopt the codes of ACTOR [9] as the baseline cVAE model and modify the motion of the actor as a condition. We use the AdamW optimizer with a fixed learning rate of 0.0001 and train the model for 1,000 epochs. 2) AGRoL: We adopt the codes from [1]. The features of the action and reaction sequences are simply concatenated together and fed to the MLP backbone. Same as in AGRoL, we build the MLP network with 12 blocks and the latent dimension is 512. We train the model for 600K steps by the AdamW optimizer with a fixed learning rate of 0.0001. 3) MDM: We adopt the codes from [11] and use the Transformer encoder-only backbone as the baseline. The features of the action and reaction are concatenated together and summed with a stan-

†Corresponding authors

Method	Train conditioned				Test conditioned			
	FID↓	Acc.↑	Div.→	Multimod.→	FID↓	Acc.↑	Div.→	Multimod.→
Real	$0.17^{\pm 0.01}$	$1.000^{\pm 0.0000}$	$5.28^{\pm 0.11}$	$20.26^{\pm 0.65}$	$1.06^{\pm 1.94}$	$0.570^{\pm 0.0056}$	$7.37^{\pm 1.17}$	$12.38^{\pm 1.33}$
cVAE	$22.04^{\pm 3.85}$	$0.800^{\pm 0.0008}$	$9.08^{\pm 0.47}$	$13.22^{\pm 0.16}$	$\mathbf{19.95^{\pm 19.37}}$	$0.374^{\pm 0.0044}$	$8.78^{\pm 0.61}$	$\mathbf{11.38^{\pm 0.75}}$
AGRoL	$36.54^{\pm 2.75}$	$0.937^{\pm 0.0001}$	$7.02^{\pm 0.16}$	$14.66^{\pm 0.24}$	$21.09^{\pm 24.31}$	$0.370^{\pm 0.0058}$	$6.68^{\pm 1.09}$	$10.49^{\pm 1.82}$
MDM	$22.92^{\pm 1.29}$	$0.979^{\pm 0.0000}$	$6.34^{\pm 0.14}$	$15.76^{\pm 0.41}$	$26.21^{\pm 14.11}$	$0.364^{\pm 0.0053}$	$6.49^{\pm 0.50}$	$9.83^{\pm 1.01}$
MDM-GRU	$24.86^{\pm 1.72}$	$0.908^{\pm 0.0000}$	$6.23^{\pm 0.12}$	$15.37^{\pm 0.36}$	$25.60^{\pm 11.58}$	$0.392^{\pm 0.0066}$	$6.33^{\pm 0.54}$	$9.59^{\pm 1.01}$
ReGenNet	$\mathbf{0.34^{\pm 0.01}}$	$\mathbf{1.000^{\pm 0.0000}}$	$\mathbf{5.52^{\pm 0.11}}$	$\mathbf{20.06^{\pm 0.48}}$	$22.88^{\pm 14.12}$	$\mathbf{0.414^{\pm 0.0064}}$	$\mathbf{7.03^{\pm 0.79}}$	$10.75^{\pm 1.37}$

Table C.1. Results on the Chi3D dataset with estimated pose results. ± indicates 95% confidence interval, → means that closer to Real is better. **Bold** indicates best result and underline indicates second best.

dard positional embedding before being fed into the Transformer encoder blocks of 8 layers. We train the model for 600K steps by the AdamW optimizer with a fixed learning rate of 0.0001. 4) MDM-GRU: We adopt the codes from [11] and use the implemented GRU backbone as the baseline. The features of the action and reaction are added together and fed into the GRU blocks of 8 layers. We train the model for 600K steps by the AdamW optimizer with a fixed learning rate of 0.0001. For the text-conditioned setting, we adopt T2M [4], MDM [11], MDM-GRU [11], RAIG [10] and InterGen [7] as baselines. We also adopt the 6D rotation representation of the SMPL-X parameters as inputs and the frame length is set to 150.

E. Details of the metric calculations

For the action-conditioned human reaction generation, we follow the prior works in human motion generation, Action2Motion [3], ACTOR [9] and MDM [11] to calculate the Frechet Inception Distance(FID) [5], action recognition accuracy, diversity and multi-modality. We borrow the code from the ACTOR [9]. Firstly, we train the action recognition model based on a slightly modified version of ST-GCN [13]. The ST-GCN model takes the 6D rotation representation of the SMPL-X parameters as input and outputs the action classification results. We train the NTU120-AS and Chi3D-AS datasets for 100 epochs with 64 batch size and 0.0001 learning rate. We generate 20 times of 1000 motion sequences with different random seeds and report the average together with the confidence interval at 95%. The definition of each metric is as follows:

1) FID: The features are extracted from the generated motions and the real motions. Then the FID is calculated between the feature distribution of the generated motions and the distribution of the real motions; 2) Action recognition accuracy: We use the pre-trained ST-GCN model to classify the generated motions and calculate the accuracy; 3) Diversity: which measures the variance of the generated motions across all action categories. Given the motion feature vectors of generated motions and real motions as $\{v_1, \cdots, v_{S_d}\}$ and $\{v'_1, \cdots, v'_{S_d}\}$, the diversity is defined as $Diversity = \frac{1}{S_d}\sum_{i=1}^{S_d} ||v_i - v'_i||_2$. $S_d = 200$ in our experiments. 4) Multi-modality: which measures how much the generated motions diversify with each action type. Given a set of motions with C action types, for c-th action, we randomly sample two subsets with size S_l, and then extract the feature vectors as $\{v_{c,1}, \cdots, v_{c,S_l}\}$ and $\{v'_{c,1}, \cdots, v'_{c,S_l}\}$, the multimodality is defined as $Multimod. = \frac{1}{C \times S_l}\sum_{c=1}^{C}\sum_{i=1}^{S_l} ||v_{c,i} - v'_{c,i}||_2$. $S_l = 20$ in our experiments.

For the text-conditioned human reaction generation, we follow the prior work of [4] and adopt the 1) Frechet Inception Distance (FID) [5] to measure the latent distance between real and generated samples, 2) diversity to measure the latent variance, 3) multimodality (MModality) to measure the diversity of the generated results for the same text, 4) R Precision to measure the accuracy of retrieving the ground-truth description from 31 randomly mismatched descriptions, and 5) MultiModal distance (MM Dist) to calculate the latent distance between generated motions and texts. We train a motion feature extractor together with a text feature extractor in a contrastive paradigm to align the features of texts and motions. We run all the evaluations 20 times (except MModality for 5 times) and report the averaged results with the confidence interval at 95%.

F. Boarder Impacts

In AR/VR and gaming applications, a well-trained non-player character (NPC) who can react properly conditioned on your body movements is highly demanded. Our model can be applied to generate plausible human reactions in real time for these applications. We believe our work will foster future research in this direction.

References

[1] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. *arXiv preprint arXiv:2304.08577*, 2023. 1

[2] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, pages 7214–7223, 2020. 1

[3] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *ACM Multimedia*, pages 2021–2029. ACM, 2020. 2

[4] 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*, pages 5152–5161, 2022. 2

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, pages 6626–6637, 2017. 2

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[7] 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. 1, 2

[8] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *T-PAMI*, 42(10):2684–2701, 2019. 1

[9] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *CVPR*, pages 10985–10995, 2021. 1, 2

[10] Mikihiro Tanaka and Kent Fujiwara. Role-aware interaction generation from textual description. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15999–16009, 2023. 2

[11] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 1, 2

[12] Neel Trivedi, Anirudh Thatipelli, and Ravi Kiran Sarvadevabhatla. Ntu-x: an enhanced large-scale dataset for improving pose-based recognition of subtle human actions. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–9, 2021. 1

[13] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, pages 7444–7452. AAAI Press, 2018. 2

[14] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *arXiv preprint arXiv:2207.06400*, 2022. 1