Closely Interactive Human Reconstruction with Proxemics and Physics-Guided Adaption —Supplementary Material—

Buzhen Huang[1,2] Chen Li[1] Chongyang Xu[3] Liang Pan[2] Yangang Wang[2] Gim Hee Lee[1]

[1]National University of Singapore [2]Southeast University [3]Sichuan University

In the supplementary material, we first provide the details of VQ-VAE, motion representation, and training procedures to help the reproduction of the experimental results. More comparisons, analyses, and qualitative experiments are also conducted to further demonstrate the superiority of the proposed method. Finally, we show some failure cases to illustrate the limitations of the current method. We also provide a supplementary video for this work.

1. VQ-VAE

A naive training of VQ-VAE suffers from codebook collapse [3]. To avoid the limitation, we adopt exponential moving average (EMA) and codebook reset (Code Reset) to improve the codebook utilization. Specifically, EMA updates the codebook smoothly: $C^t \leftarrow \lambda C^{t-1} + (1-\lambda)C^t$, and C^t is the codebook in the current iteration. $\lambda = 0.99$ is a exponential moving constant. Code Reset finds inactivate codes during the training and reassigns them according to input data.

For the model architecture, the encoder of the discrete interaction prior consists of a motion embedding layer, a positional encoding layer, and 4 transformer blocks. The decoder has a motion decoding layer and 4 transformer blocks. Each codebook has a size of 256×256.

2. Motion representation

Previous works [7] always represent human motion in a canonical space, and the global rotation and translation are obtained by accumulating local angular and linear velocities. This representation cannot be directly applied to multi-person scenarios since it does not maintain the person-to-person spatial relationships. To address this problem, we use the root position of character a in the first frame as origin and transform the interactive motions to the new coordinate. Consequently, the joint positions and velocities are kept in the world frame. In addition, the two-person interactions satisfy commutative property, which means the interaction $\{\mathbf{x}^a, \mathbf{x}^b\}$ and $\{\mathbf{x}^b, \mathbf{x}^a\}$ are equivalent.

Dataset	w/o proj. loss	Ours
3DPW	73.8	70.6
Hi4D	64.2	63.1

Table 1. **Ablation on projection loss gradients.** "w/o proj. loss" denotes our model without the projection loss gradients. The number are MPJPE.

Method	Accel↓	A-PD↓
GroupRec	25.2	1.34
Ours	10.7	1.15

Table 2. Accel and A-PD are acceleration error and average penetration depth, respectively.

Timesteps	step=1	step=3	step=5	step=10
MPJPE	68.2	64.4	63.1	63.0

Table 3. **Ablation on number of duffusion timesteps.** The ablation is conducted on Hi4D.

3. Implementation details

The diffusion model has the same stracuture as the VQ-VAE encoder, which contains a motion embedding layer, a positional encoding layer, and 4 transformer blocks. The number of diffusion timesteps is set to 100 during the training stage. In the inference, we adopt DDIM sampling strategy [5] with 5 timesteps to achieve the distribution adaption. For the projection loss gradients, we use ViT pose [6] to predict 2D poses. To train the diffusion model, 25% of projection loss gradients and image features are randomly masked. The frame length of interactive motions is 16, and the batch size for VQ-VAE and diffusion model are 256 and 32, respectively. All the models are trained with AdamW [2] optimizer using a learning rate of 1e-4 on a single GPU of NVIDIA GeForce RTX 4090.

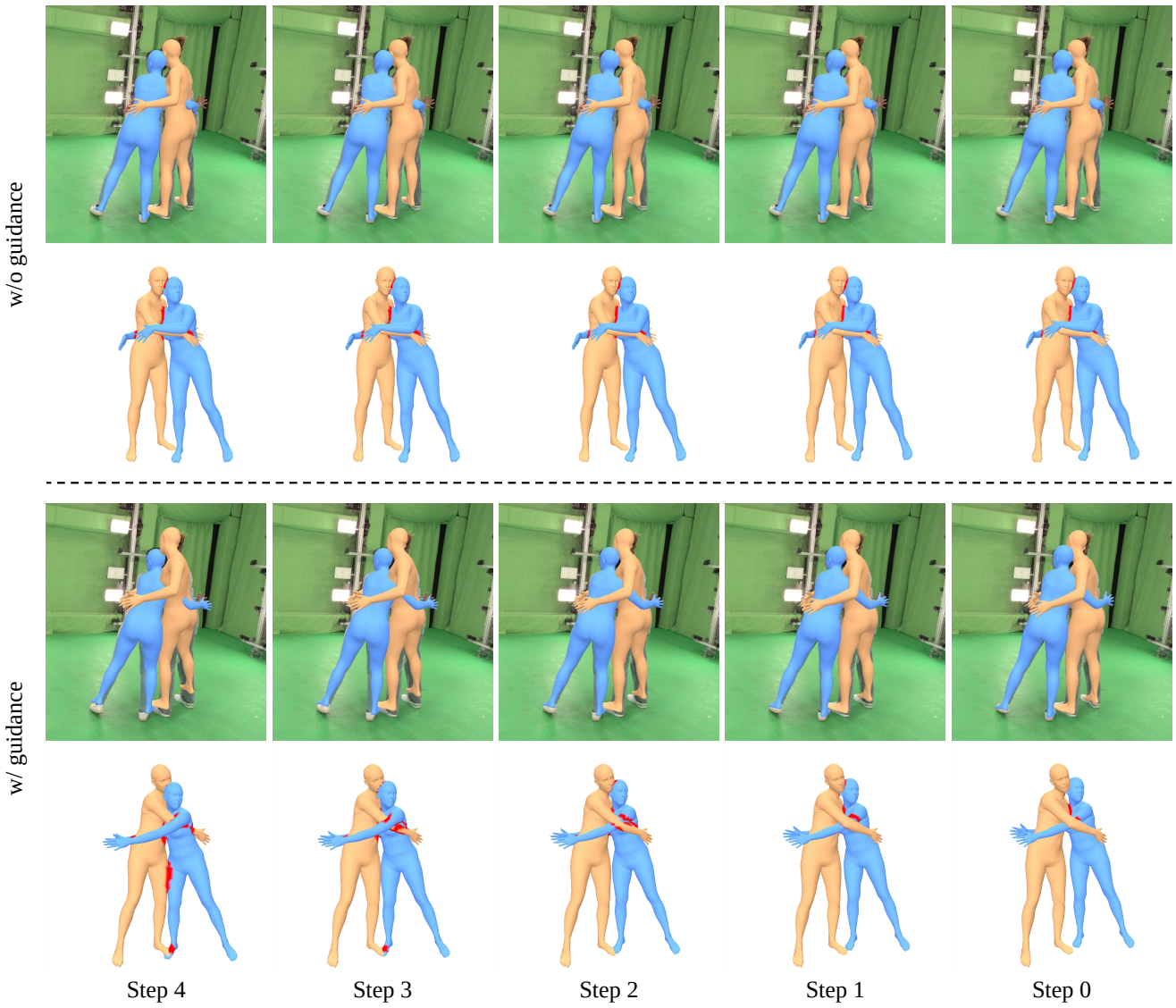

Figure 1. Comparison between the models with and without the guidance. Without the guidance, the reverse diffusion process has limited effect in improving the interations.

4. Extended experiments

Projection loss gradients. We also investigate the projection loss gradients, which provide signals to guide the reverse diffusion process and enforce the 3D models to be consistent with image observations. The term can improve joint accuracy for common scenarios. However, it has limited effect on severely occluded cases since the current state-of-the-art 2D pose detectors cannot produce reliable results for ambiguous images. In Tab. 1, Hi4d has more complex interactions than 3DPW, and the performance on 3DPW dataset significantly decreases without the projection loss gradients.

Diffusion guidance. In Fig. 1, we show the intermediate results in each timestep during the inference phase. The inference contains 5 timesteps with the denoising diffusion implicit models [5]. Although the reverse diffusion has the same timesteps, we find that the denoising model without the guidance produces slight changes and the final results still show severe penetrations. In contrast, our model can refine the interactions and alleviate penetrations, which demonstrates the importance of proposed guidance.

Diffusion timesteps. We analyze the impact of different timesteps in Tab. 3. The performance increases with more timesteps at first and then becomes stable. To balance the accuracy and efficiency, we use 5 timesteps in the inference phase.

Penetration and acceleration error. In Tab. 2, we further use the average penetration depth (A-PD) [4], which reflects the degrees of inter-penetration, to evaluate the body contact

Figure 2. More qualitative results on Hi4D, 3DPW and CHI3D datasets. Our method can reconstruct closely interactive humans with plausible body poses, natural proxemic relationships and accurate physical contacts from single-view inputs

and penetration, and our method can produce better performance. Our method also outperforms GroupRec [1] on the acceleration error due to the proposed velocity loss and temporal architecture.

More results. We also show more qualitative results in Fig. 2. Our method can reconstruct closely interactive humans with plausible body poses, natural proxemic relationships and accurate physical contacts from single-view inputs.

References

[1] Buzhen Huang, Jingyi Ju, Zhihao Li, and Yangang Wang. Reconstructing groups of people with hypergraph relational reasoning. In *ICCV*, pages 14873–14883, 2023. 3

[2] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[3] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *NeurIPS*, 32, 2019. 1

[4] Yu Rong, Jingbo Wang, Ziwei Liu, and Chen Change Loy. Monocular 3d reconstruction of interacting hands via collision-aware factorized refinements. In *3DV*, pages 432–441, 2021. 2

[5] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2

[6] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 1

[7] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052*, 2023. 1