Reconstructing Close Human Interaction with Appearance and Proxemics Reasoning —Supplementary Material—

Buzhen Huang[1,2] Chen Li[4,5] Chongyang Xu[3] Dongyue Lu[2] Jinnan Chen[2]
Yangang Wang[1] Gim Hee Lee[2]

[1]Southeast University [2]National University of Singapore [3]Sichuan University
[4]IHPC, Agency for Science, Technology and Research, Singapore
[5]CFAR, Agency for Science, Technology and Research, Singapore

In this supplementary material, we first introduce the implementation details for reproducibility of our results. We also conduct more quantitative, qualitative and ablation experiments to demonstrate the effectiveness of our method. Finally, we discuss the protocol for dataset cleaning. More qualitative results can be found at https://www.buzhenhuang.com/works/CloseApp.html.

0.1. Implementation Details

The diffusion model contains a motion embedding layer, a positional encoding layer, and 4 transformer blocks. It also includes a ViT backbone and 2 linear layers to process the image and keypoints features. We use 100 and 5 diffusion timestep for the training and testing phases, respectively. DDIM sampling strategy [7] is adopted during the denoising process. In the optimization, we first diffuse the initial predicted motions to the first timestep $\mathbf{x}_t$, and then optimize the network parameters of proxemics prior to reconstruct the final motions. The proxemics prior is pretrained with AdamW [4] optimizer using a learning rate of 1e-4 on a single GPU of NVIDIA GeForce RTX 4090. The implementation is based on a machine with 64 GB memory. In our current implementation of our dual-branch optimization, the batch size is 1, and we optimize 16 frames in each interaction until the entire sequence is optimized. The optimization is completed when the overall loss is convergent, which typically takes 20 epochs for a single video. We set the maximum of the number of epochs to be 100.

0.2. More Comparisons and Results

We show more results on in-the-wild videos in Fig. 8 and Fig. 7. Some of them are samples from the proposed WildCHI dataset. The results demonstrate that our method can work well on diverse environments even with adverse lighting conditions. Although AutoTrackAnything cannot produce accurate masks for each individual in interactive cases, we find it is robust to segment all human related pixels from the image. It is sufficient for our method since we render two individuals in the same scene and use original RGB images as a constraint. With this strength, our method can reconstruct interactive humans from outdoor image with complex background and human clothes.

Figure 6. **Qualitative video on Hi4D dataset. It is an animatable figure, which can be viewed with Adobe Acrobat Reader.**

Method	MPJPE	Interaction	Joint PA-MPJPE
BUDDI	96.8	102.6	104.3
BUDDI-t	90.3	99.1	98.2
BUDDI-t w/ Appearance	88.1	96.0	96.6
CloseInt	63.1	81.4	72.6
Ours	59.1	80.2	70.2

Table 4. **Comparison with temporal baseline methods.** BUDDI-t is a temporal version of BUDDI, which can also be improved by the proposed appearance loss. Our method can outperform temporal baseline methods due to the proposed proxemic prior and appearance loss.

To further compare with optimization-based temporal frameworks for a fair comparison, we simultaneously optimize BUDDI on the entire sequence with additional temporal regularization (Eq. 8 in the main paper), which is named as BUDDI-t. As shown in Tab. 4 and Fig. 7, although the temporal information can improve BUDDI, our method still achieves better performance due to the proposed appearance

Method	PA-MPJPE	Joint PA-MPJPE
BEV	51.0	96.0
BUDDI	47.5	68.0
Ours	38.6	57.4

Table 5. **Quantitative results on CHI3D.**

loss and proxemic prior. We also adopt Joint PA-MPJPE proposed in BUDDI [6] as a metric, and the results in Tab. 4 show that our method is still superior.

In addition, we exclude the training data from CHI3D, and follow BUDDI to conduct a quantitative comparison on CHI3D S03. As shown in Tab. 5, our method can also outperform the baseline methods.

0.3. Abaltion on UV Map Size

We investigate the impact of Gaussian UV map size. Different UV map sizes result in different number of Gaussians. More Gaussians can promote better human appearances. In Tab. 6, we can find that body poses are worse with a smaller UV map, which indirectly highlights the importance of appearance constraint.

0.4. Physical Constraint

Physical constraint can prevent mesh penetrations in the close interaction. However, we find that it affects the pose accuracy. The reason is that this constraint introduces more local minimas and make it more difficult to find the optimal solution space. Consequently, it is also important for future works to design an accurate physical constraint that can be compatible with other constraints.

0.5. Temporal Information *vs.* Visual Appearance

In previous works [2], temporal information is used to model close interactions. However, visual appearance is also important as temporal information alone is not sufficient to address the visual ambiguity. Temporal information is effective in preventing high-frequency jittering (*e.g.*, when the 2D detections or 3D poses in a few non-consecutive frames are erroneous) and can result in improvement on the entire sequence. However, during the close interaction, incorrect detections always persist for multiple frames or even throughout the entire interaction process since existing models (*e.g.*, SAM and ViTPose) cannot clearly identify human semantics (*e.g.*, 2D keypoints and mask) in complex interactions. As shown in Fig. 1 of the main paper, the results are still not good despite the fact that we have already used a temporal version of SAM [5] (Autotrackanything). Specifically, temporal information cannot compensate for a large amount of unreliable 2D detections during the close interaction since human reconstruction methods rely on 2D observations to achieve model-image alignment. Although the temporal constraint is applied, BUDDI and CloseInt still overfit to the wrong 2D detections or produce oversmoothing motions during the close interaction, as shown in Tab. 4 and Fig. 7. In contrast, our method avoids 2D detections by modeling 3D appearance using 3DGS on all original RGB frames, which provide reliable 2D observations to constrain the interaction. Due to the appearance loss, our method outperforms the current SOTA temporal method, CloseInt. In addition, the results in Tab. 4 also demonstrate that the appearance loss can further improve BUDDI-t on Hi4D dataset.

Although the appearance modeling also relies on temporal information, the strength of appearance in disentangling visual ambiguities cannot be achieved by purely temporal constraints.

0.6. Image Features and 2D Keypoints

Some large-scale datasets (*e.g.*, Inter-X [8] and InterHuman [3]) do not contain paired RGB images. To leverage these datasets for learning the proxemic prior, we can only project the 3D joints to the image plane and then use the 2D pose as a condition for the diffusion model. We conduct an experiment on different conditions of the diffusion model in Tab. 7. Images contain more information (*e.g.*, body shape and ordinal relationship) for reconstruction, and the 2D keypoints are not required when RGB images are available. However, 2D keypoints can help to leverage knowledge from pure 3D datasets. Therefore, we use both these 2D observations as our condition.

0.7. Robustness of appearance loss

The appearance loss does not require high-quality textures, and it can work as long as the textures of the two humans are distinguishable as shown in Fig. 9. So, the loss is effective in most scenarios. However, it may fail when the two individuals are wearing the similar color clothing. We have to rely on other constraints (2D keypoints, proxemics, and temporal information) to achieve the reconstruction in such special cases. In addition, the complex light-conditions (*e.g.*, shadows) and cloth deformations can affect the reconstructed textures as we discussed in the limitation, a light- and pose-dependent design for the appearance may alleviate the negative impacts.

0.8. Contact Information

A contact constraint may promote more accurate interactions in our method. However, it is difficult to identify the contact regions of two closely interactive humans from monocular RGB images. We do not use contact constraint since current methods cannot predict reliable contact information from in-the-wild videos. As demonstrated in [1], humans can only achieve 49.9% consistency at 17 regions partitioning on FlickrCI3D dataset, while the model's predictions are even worse, at only 24.8%. Therefore, [1] uses ground-truth contact annotations during the optimization.

Figure 7. **Samples from the proposed WildCHI dataset**, which contain complex interactions. BUDDI-t is a temporal version of BUDDI, which cannot produce satisfatory results on these in-the-wild videos. **It is an animatable figure, which can be viewed with Adobe Acrobat Reader.**

Figure 8. Our method can work well on diverse environments even with adverse lighting conditions.

In contrast, our method leverage appearance and proxemics information can also obtain satisfactory results.

0.9. Protocol for Dataset Cleaning

Our method cannot guarantee completely satisfactory reconstruction. To ensure the quality of the proposed dataset, we first evaluate the annotations of the entire sequence with re-projection error, and then manually check each frame with the rendered image and a 3D GUI. When the rendered result or 3D interaction is incorrect, we mark the frame as invalid. Finally, we select 100 sequences based on motion diversity and quality, and 96.1% frames are valid, which demonstrate that our method is relatively robust. The final dataset contains different and complex interaction types, with each sequence performed by different subjects.

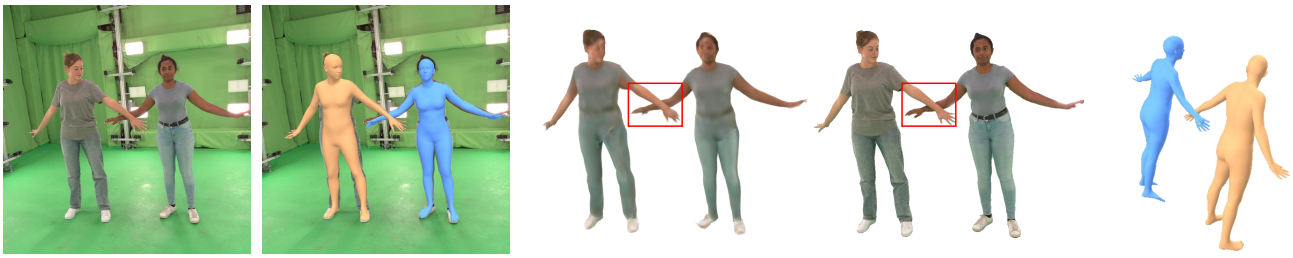

Figure 9. Although the reconstructed texture is not good, the appearance loss can work as long as the textures of the two humans are distinguishable.

Method	MPJPE	PA-MPJPE	MPVPE	Interaction	A-PD
128 × 128	61.80	46.42	74.89	80.65	0.74
256 × 256	60.64	45.84	73.60	80.02	0.78
512 × 512	59.06	44.29	71.99	80.18	0.81

Table 6. **Ablations on UV map size.** The ablation is conducted on Hi4D dataset.

Condition	2D keyp.	Image	2D keyp. + Image
Hi4D	71.2	63.1	63.0
Hi4D + Inter-X	66.3	–	62.1

Table 7. **Ablation on conditions of diffusion model.** 2D keypoints can help to employ prior knowledge from large-scale dataset like Inter-X.

References

[1] 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. 2

[2] Buzhen Huang, Chen Li, Chongyang Xu, Liang Pan, Yangang Wang, and Gim Hee Lee. Closely interactive human reconstruction with proxemics and physics-guided adaption. In *CVPR*, 2024. 2

[3] 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

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

[5] Roman Lyskov. Autotrackanything, 2024. 2

[6] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. In *CVPR*, 2024. 2

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

[8] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, Yunhui Liu, Wenjun Zeng, and Xiaokang Yang. Inter-x: Towards versatile human-human interaction analysis. *arXiv preprint arXiv:2312.16051*, 2023. 2