HOI-M^3: Capture Multiple Humans and Objects Interaction within Contextual Environment

Supplementary Material

Due to space limitations, we have to remove some details that are not included in the main paper in the appendix.

1. More Details of HOI-M^3 Dataset

In this section, we provide more details about HOI-M^3 dataset, including statistic analyses, data preprocess and hardware setup.

1.1. Dataset Statistic

HOI-M^3 provides a large volume of long human object interactions(HOI) (more than 10k frames HOI per sequences), which will be beneficial for long-term motion and HOI generation. To assess the dataset's diversity, we provide key statistics, including gender, height, weight, and object scale, illustrated in Figure 1. The results demonstrate the dataset's diversity in human body shapes and object scales.

1.2. Data Preprocess

For accurate object tracking, separating the target object from the background in a video sequence serves as a crucial cue for optimization. However, tracking an arbitrary object in diverse scenes is a non-trivial task. Following previous work, Track-Anything [16], we employ the Segment Anything Model (SAM)[7] to annotate the initial frame of each camera view. Subsequently, we utilize XMem[1] for video object tracking (VOS) on the subsequent frames.

1.3. Hardware Setup

Accurately capturing the motions of multiple humans and objects remains a challenging task, particularly in the presence of severe occlusions, a common occurrence in daily interactions within contextual environments. To address this challenge and capture realistic interaction sequences, we designed a custom room-like dome with a square-shaped multi-layer framework to house the RGB sensors. The system stands at a height of 2.9 m and has a side length of 7.8 m for its octagonal cross-section, as illustrated in Figure 3. To better align our capture setup with everyday scenarios, we opted for white backdrops instead of green ones to conceal the cable. We also provide more quality results sampled from HOI-M^3 dataset as shown in Figure 5.

2. How HOI-M^3 Contributes to the Community?

The HOI-M^3 dataset comprises various scenes depicting human-object interactions, accompanied by per-frame multiple human and object tracking. We believe our dataset addresses a significant gap in the literature on multiple human-object interactions. At the meanwhile, we anticipate that the dataset will serve as a valuable resource for various research directions. We propose the following challenges based on the HOI-M^3 dataset:

Multiple Person Pose and Shape Estimation. HOI-M^3 offers parametric model labels encompassing shape information and 3D skeletal positions. This provides a robust benchmark for multi-person scenarios, particularly in daily situations where individuals are frequently occluded by surrounding objects. We believe that HOI-M^3 serves as a reflective measure of each method's performance in such challenging scenarios.

Multiple HOI Capture. In recent years, significant advancements have been made in data-driven human motion capture, even for single HOI capture. However, there has been limited progress in monocular multiple HOI capture. The HOI-M^3 dataset addresses this gap by providing the largest and most accurate capturing labels paired with natural RGB images, enabling robust HOI supervision. Consequently, our dataset is well-suited for data-driven approaches in both monocular and multi-view settings, leveraging the precision of our ground truth annotations.

Multiple Human Motion Generation. We have witnessed remarkable advancements in diffusion techniques for generating lifelike human motions, progressing from single human motion [2, 6, 15, 18–20] to the recent exploration of two-human interactions [12]. Leveraging the extensive dataset of long-duration multi-human motions in HOI-M^3, we can offer accurate labels for multi-human interactions to facilitate this evolving task.

Multiple Interaction Generation. HOI-M^3 provides an extensive collection of diverse interaction sequences with synchronized ground truth capture. Motivated by the recent significant progress in Motion Generation (MoGen) tasks, we have demonstrated how our dataset contributes to this field in the main paper, particularly in the context of a novel task: Multiple Interaction Generation.

3. More Details of Monocular Multiple HOI Capture

3.1. Network Architecture

For a fair comparison, we do not choose large size of backbone; instead, we employ ResNet-34 [4], pre-trained on the ImageNet dataset [3], as the default backbone. All input

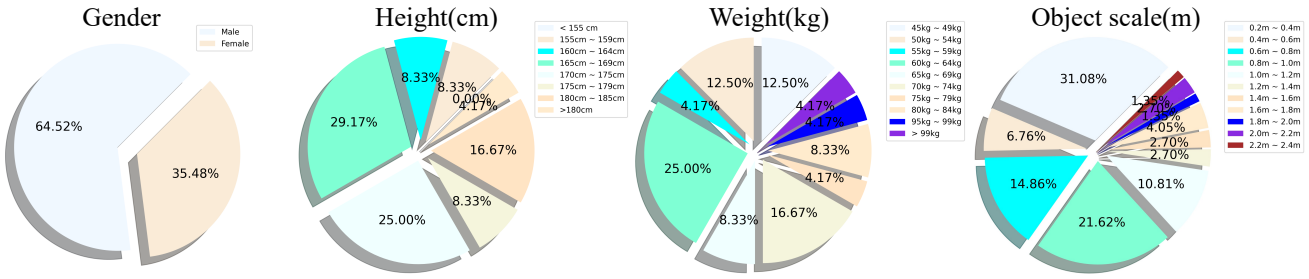


Figure 1. Statistics of HOI-M^3 humans and objects.

images were padded to the standardized size of 512×512. Each prediction head attached to the backbone comprises a $3 \times 3 \times 256$ convolutional layer, BatchNorm, ReLU, and another $1 \times 1 \times c_0$ convolutional layer, where c_0 represents the output size.

3.2. Loss Function

To supervise the network, we have developed individual loss functions for different maps. The network is supervised by the weighted sum of the body pose loss L_{theta}, the body shape loss L_{beta}, the object pose loss L_{object}, the 3D keypoints loss L_{3D}, the 2D keypoints loss L_{2D}, the center keypoint heatmap L_{hm}, and the depth loss of humans and objects L_{depth}.

Human Object Center Loss. We employ a heatmap representing the 2D human body center and object center in the image, which is represented as a Gaussian distribution in the human-object position. The center keypoint heatmap L_{hm} is derived as follows:

$$L_{\text{hm}} = \|C_m^{\text{pred}} - C_m^{\text{gt}}\|_2, \tag{1}$$

where $C_m^{\text{pred}} \in \mathbb{R}^{128\times128}$ is the predicted center heatmap, and $C_m^{\text{gt}} \in \mathbb{R}^{128\times128}$ is the ground truth of C_m^{pred}.

Human Parameter Loss. Through the parameter sampling process, we enforce the human parameter loss L_{theta} and L_{beta} to match each ground truth body with a predicted parameter result for supervision. The body pose loss L_{theta} and the body shape loss L_{beta} are derived as follows:

$$\begin{aligned} L_{\text{theta}} &= \|\theta^{\text{pred}} - \theta^{\text{gt}}\|_1, \\ L_{\text{beta}} &= \|\beta^{\text{pred}} - \beta^{\text{gt}}\|_1, \end{aligned} \tag{2}$$

where $\theta^{\text{gt}} \in \mathbb{R}^{24\times3}$ and $\beta^{\text{gt}} \in \mathbb{R}^{10}$ denote the ground truth of the model's parameters. $\theta^{\text{pred}} \in \mathbb{R}^{24\times3}$ and $\beta^{\text{pred}} \in \mathbb{R}^{10}$ denote the predicted parameter results sampled from each center position of the human. Here we use the ℓ_1 norm, following previous work [14, 17].

Object Pose Loss. Similar to Human Parameter, we sample the object's 6D pose from each object center with a predicted parameter result for supervision. The object pose loss L_{object} is derived as follows:

$$L_{\text{object}} = \|R^{\text{pred}} - R^{\text{gt}}\|_1, \tag{3}$$

where $R^{\text{pred}} \in \mathbb{R}^{3\times2}$ denotes predicted object rotation, and $R^{\text{gt}} \in \mathbb{R}^{3\times2}$ denotes the ground truth of the rotation.

Depth Loss. Besides the local representation of humans and objects, another key component is depth. Here we impose each subject's depth as follows:

$$L_{\text{object}} = \|Z_{\text{center}}^{\text{pred}} - Z_{\text{center}}^{\text{gt}}\|_1, \tag{4}$$

where $Z_{\text{center}}^{\text{pred}} \in \mathbb{R}$ denotes the predicted depth of humans or objects, and $Z_{\text{center}}^{\text{gt}} \in \mathbb{R}$ denotes the ground truth of the depth.

Additional Loss. In addition to imposing supervision on each regression target, we also utilize some intermediate supervised signals for training, such as 2D keypoints and 3D keypoints of humans:

$$\begin{aligned} L_{\text{2D}} &= \|P_{\text{2D}}^{\text{pred}} - P_{\text{2D}}^{\text{pred}}\|_1, \\ L_{\text{3D}} &= \|P_{\text{3D}}^{\text{pred}} - P_{\text{3D}}^{\text{pred}}\|_1, \end{aligned} \tag{5}$$

where $P_{\text{2D}}^{\text{pred}} \in \mathbb{R}^{24\times2}$ and $P_{\text{3D}}^{\text{pred}} \in \mathbb{R}^{24\times3}$ denote predicted 2D and 3D keypoints, and $P_{\text{2D}}^{\text{gt}} \in \mathbb{R}^{24\times2}$ and $P_{\text{3D}}^{\text{gt}} \in \mathbb{R}^{24\times3}$ denote the ground truth of 2D and 3D keypoints.

4. More Details of Multiple Interaction Generation

Our diffusion models encompass both a forward diffusion process and a reverse diffusion process. The forward diffusion process progressively introduces Gaussian noise to the original data x_0. In this case, we employed a transformer model architecture as our denoising network, comprising four self-attention blocks. Each self-attention block consists of a multi-head attention layer followed by a position-wise feed-forward layer. Illustrated in Figure 2, our denoising network incorporates several feature embeddings. Specifically,

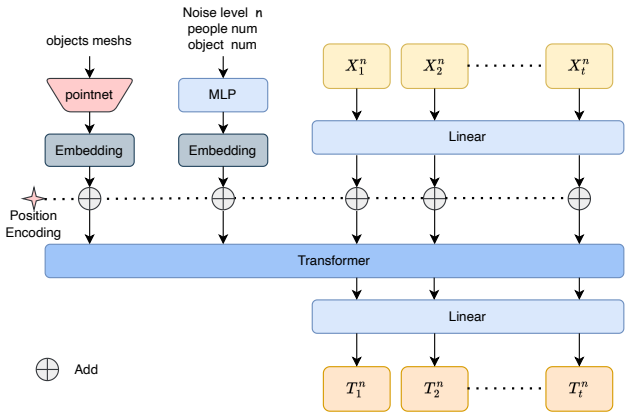


Figure 2. Model architecture of denoising network.

it includes embeddings from object meshes and condition signals of noise levels n, human numbers, and object numbers, which are then concatenated together as input to our transformer model.

5. Experiment

5.1. Ablation Study

To comprehensively evaluate the components of inertial-aided multi-object tracking, we perform an additional qualitative analysis of various constraint terms. It is important to note that we lack ground truth specific to tracking, so our evaluations are qualitative in nature. In figure 6, we present the quality results obtained by ablating different components. Specifically, "w/o collision," "w/o IMU Init," and "w/o offscreen loss" denote the results obtained without using the collision constraint term, without employing the IMU as initialization, and without utilizing the offscreen term $E_{\text{offscreen}}$, respectively. The results demonstrate that the offscreen term $E_{\text{offscreen}}$ effectively prevents degenerate results. Furthermore, without IMU initialization, recovering the object's rotation from the human-object mask becomes challenging, and our collision loss ensures realistic interactions between humans and objects.

5.2. More Benchmarks

Monocular 3D Human Pose and Shape Estimation In addition to the two benchmarks for novel data-driven tasks and their corresponding strong baselines presented in the main paper, we also introduce additional benchmarks for a prevalent vision task: monocular 3D human pose and shape estimation. To ensure a fair comparison with existing works, we conduct several experiments on our datasets. For evaluation metrics, we utilize mean per joint position error (MPJPE), procrustes aligned mean per joint position error (PA-MPJPE), the percentage of correct keypoints (3DPCK), and area under curve (3D-AUC) to assess the performance of 3D pose due to their common usage. Additionally, we employ per vertex error (PVE) to evaluate body mesh estimation ability. Furthermore, we report the percentage of correct keypoints after procrustes alignment (PA-3DPCK) and area under curve after procrustes alignment (PA-3DAUC) on our dataset. We believe that our dataset currently stands as the most comprehensive benchmark in terms of evaluation metrics. The main results are presented in Table 1, indicating that conducting tests in scenarios involving multiple persons within multiple object occlusions poses a significant challenge compared to results obtained from other datasets.

References

[1] Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 1

[2] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *CVPR*, 2023. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[5] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 4

[6] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Gmd: Controllable human motion synthesis via guided diffusion models. *arXiv preprint arXiv:2305.12577*, 2023. 1

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

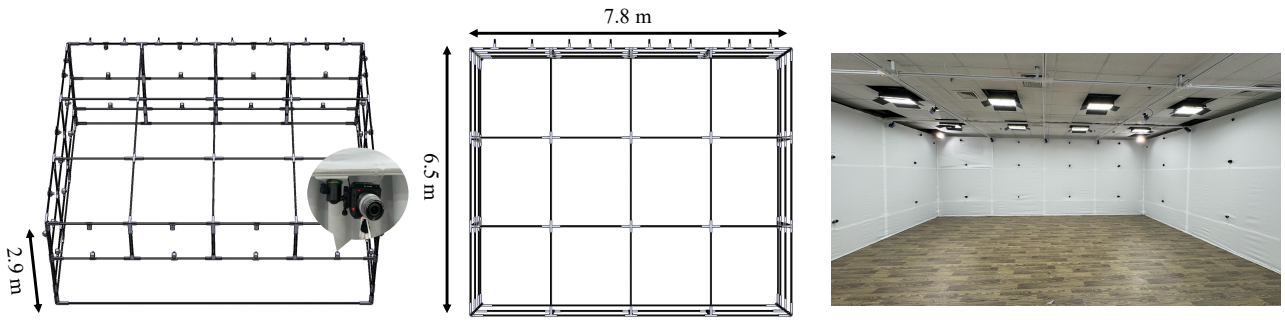
[8] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, pages 11127–11137. IEEE, 2021. 4

[9] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, pages 2252–2261, 2019. 4

[10] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, pages 3383–3393, 2021. 4

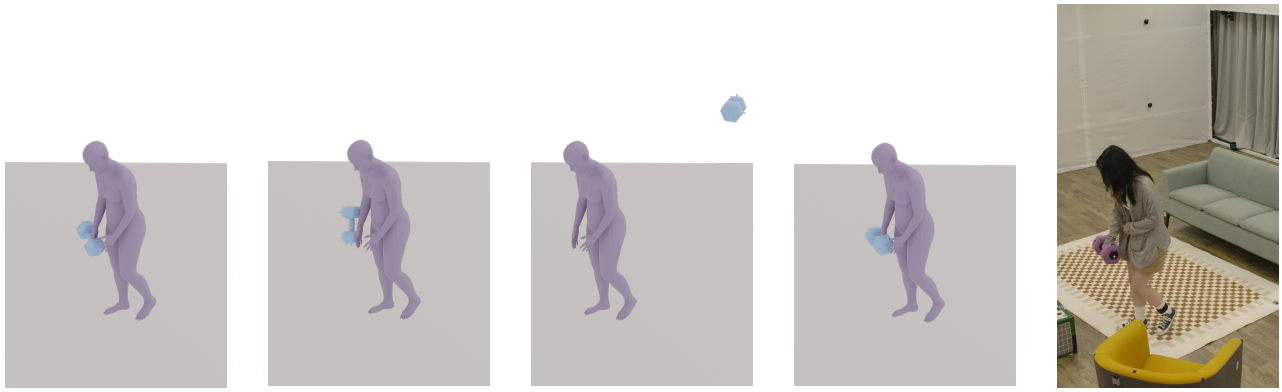
[11] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 4

[12] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion



a) Perspective view b) Top view c) Capture system

Figure 3. Hardware setup.



w/o collision constraint w/o IMU Init w/o offscreen loss Ours Image

Figure 4. Qualitative evaluation.

Method	MPJPE↓	PA-MPJPE↓	3DPCK↑	PA-3DPCK ↑	3DAUC↑	PA-3DAUC↑	PVE↓
HMR [5]	324.60	187.69	13.64	50.68	3.57	16.71	404.49
SPIN [9]	**309.81**	160.56	16.97	53.74	5.11	23.33	357.00
HybrIK [10]	326.86	**127.74**	**18.95**	**68.79**	**6.74**	**29.96**	**335.55**
PARE [8]	325.64	188.65	9.63	**46.50**	1.77	14.49	403.25
BalancedMSE [13]	331.93	152.40	13.85	56.35	4.02	26.06	346.16
CLIFF [11]	332.47	161.09	14.31	58.39	4.25	23.54	413.70

Table 1. Monocular 3D human pose and shape estimation benchmark. The best results are in **bold**.

generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023. 1

[13] Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced mse for imbalanced visual regression. In *CVPR*, 2022. 4

[14] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, pages 11179–11188, 2021. 2

[15] 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

[16] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos, 2023. 1

[17] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *CVPR*, pages 546–556, 2021. 2

[18] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3d environments. In *2020 International Conference on 3D Vision (3DV)*, pages 642–651. IEEE, 2020. 1

[19] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *CVPR*, pages 6194–6204, 2020.

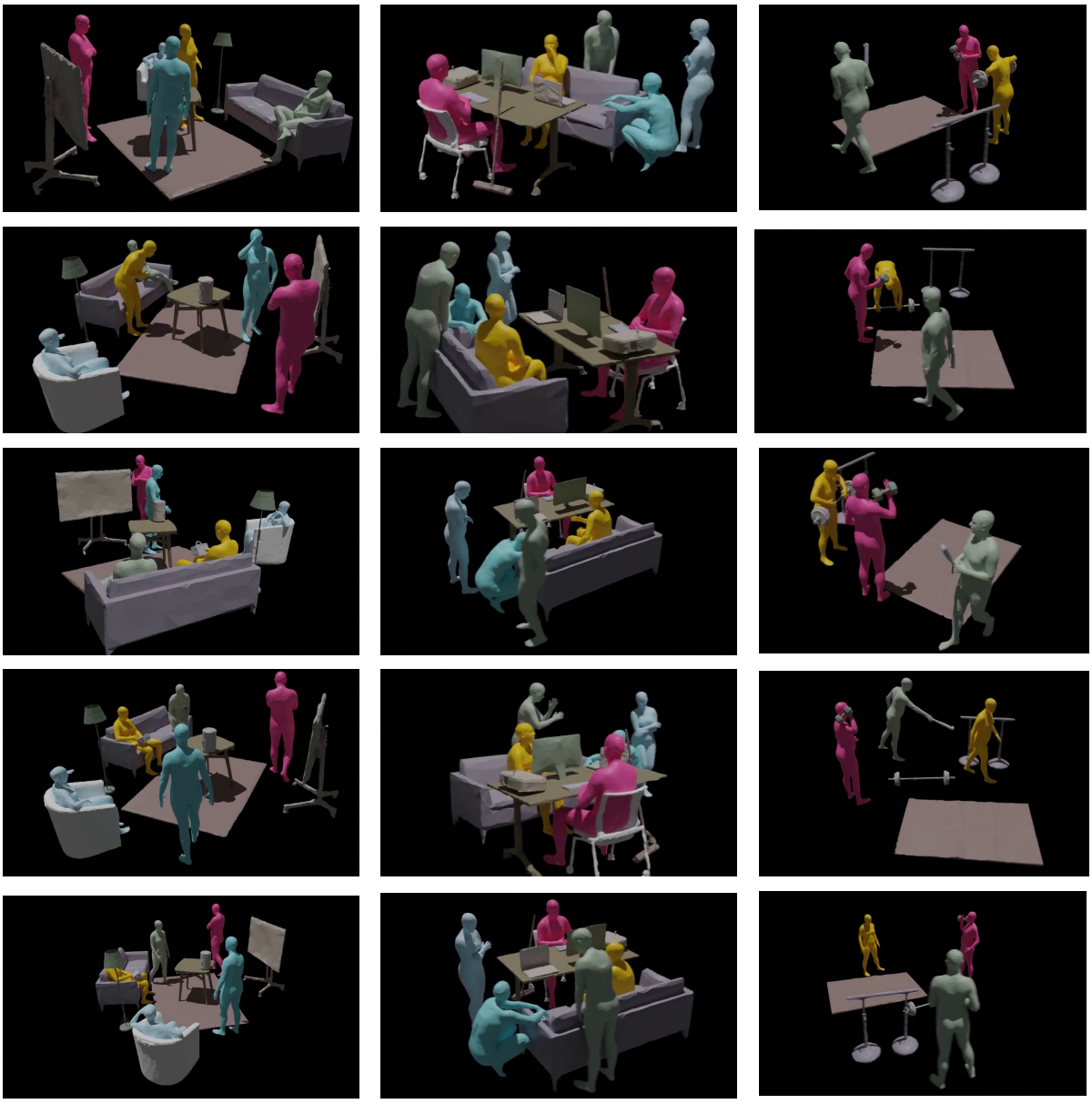


Figure 5. More quality results.

[20] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *CVPR*, pages 3372–3382, 2021. 1

Figure 6. Data examples were captured by our system.