Supplementary Material of NeuralDome: A Neural Modeling Pipeline on Multi-View Human-Object Interactions

1. More Details of Data Capturing

In this section, we discuss more details about our data capturing system and annotation process.

1.1. Data Preprocess

Accurate human-object segmentation is required for joint optimization of capturing, Instant-NSR [16] and Instant-NGP [9]. For each recording frame, we segment the human-object foreground mask by running background matting [12] and detect 2D joints (include hand) by using state-of-the-art 2D joint detectors, e.g. OpenPose [4].

1.2. Object Tracking

For object tracking, each object is represented by a rigid mesh by pre-scanned tenplate. To capture object motions, we attach 10 mm hemispherical markers with strong glue directly to the object's surface and use at least 4 markers for each object. Note that we empirically distribute them on the object to ensure at least 3 of them are always observed.

1.3. Time Synchronization

To avoid motion blur, we set the Optitrack system [1] to work at 120 FPS and set the RGB system to work at 60 FPS. Thus we need to resample the tracking data (120Hz) to the same frame rate as RGB (60Hz) to synchronize the RGB and marker system. Besides, we place an additional marker on the hand of the actor and block it as the flag of starting. Then we manually label the start frame in both recording timestamps.

1.4. Calibration

To align RGB and Optitrack into the unified world coordinate system w.r.t. to the scene, we need to register the Optitrack maker set to the 3D scene. First human annotator annotates 3 correspondences between the marker recording data and RGB world location and then estimates the marker-to-RGB rigid transformation using ICP [2, 17]. Besides, camera extrinsic parameters and rigid transformation are fixed during each recording.

1.5. Data Capture Protocol

We recruit 10 subjects (5 males and 5 females) that are between 18-40 years old and between 1.5-1.95m heights. Each subjects are recorded while interacting with 23 objects, according to their time availability. To ensure interact with objects as naturally as possible, each subjects are not instructed to do any actions.

2. Contribution to Community

Consisting of various human-object interacting scenes with rich labels covering capturing and rendering labels, our NeuralDome pipeline and HODome dataset fill an important gap in the literature and support many research directions. We propose the following challenges with HODome dataset:

Interaction Capturing. HODome dataset provided the largest accurate capturing label with paired natural RGB images for strong HOI supervision. Benefiting from our 76-view setting, our dataset is suitable for the monocular and multi-view settings. Moreover, our quantitative subset can be used for benchmarking thanks to our accurate ground truth and dense view validation.

Geometry Reconstruction. Joint human-object geometry Reconstruction is a challenging task and the existing dataset can not be used for the benchmark. Besides, existing publicly available data do not support learning an accurate data-driven model of human-object geometry. Thanks to our dense capture setting, NeuralDome can provide high-fidelity human/object geometry to enable this task.

Object-Occluded Pose and Shape Estimation. Existing public datasets (e.g. Human36m [6] and 3DPW [14]) mainly focus on capturing accurate human labels but ignore the object conditions. By contrast, HODome dataset provided accurate capturing labels with SMPL parameters and 3D keypoints under challenge object-occluded case.

Neural Rendering in HOI scenarios. With the human-object interaction sequences captured by our dense cameras, there are lots of interesting and meaningful directions to explore. First, it's interesting to extend our layer-wise human-object representation to weaker settings that are closer to real life, e.g., sparse views, without accurate object poses.

Besides, generalizable neural rendering techniques should be further developed to support HOI scenarios where occlusions are inevitable Moreover, our HODome can naturally enable building photo-realistic neural avatars that support object-aware deformation in HOI scenarios.

3. More Details of Human-object Tracking

In Section 4.1 of the paper, we have described the joint tracking between humans and objects. Here we elaborate more details and mathematical formulations.

3.1. Tracking Initialization

Object Tracking. We consider each object $V \in \mathbb{R}^m$ as a rigid body mesh model with m vertices. And we only need to estimate the translation $T_t \in \mathbb{R}^3$ and rotation $R_t \in \mathbb{SO}(3)$ with respect to its pre-scanned template on each frame t. The 3D location of the object mesh on per-frame is represented as,

$$V_t(R_t, T_t) = R_t \mathcal{O}(c_t, p_t) + T_t, \tag{1}$$

where $\mathcal{O}(c_j)$ represents the p_j part of category c_j mesh template. T_t and R_t is rigid transformation estimated from a per-frame marker set using Rigid-ICP.

3.2. Joint optimization for human-object tracking

We used SMPL-X [10] as the body model, which provides a differentiable function $\mathcal{M}(\cdot)$ to control an artist-created mesh with $N = 10475$ vertices and $K = 54$ joints. We estimate body shape and pose over the whole sequence from multi-view RGB videos in a frame-wise manner. Recall that we imposed several regularization terms in Eq. (1) to ensure plausible interactions. Here we elaborate the mathematical formulation of each term.

Following the previous method [3, 5], we compute the index of vertex where the body is in contact with objects, and enforce contact between SMPL-X and object explicitly as the following term:

$$E_{\text{contact}} = \|\mathbb{1}_t^{\text{contact}}(V_t(R_t, T_t) - \mathcal{M}(\beta_t, \theta_t, \psi_t, \gamma_t))\|_2^2, \tag{2}$$

where $\mathbb{1}_t^{\text{contact}}$ denote as binary indicator matrix computed from the contact map. Note that marker-based tracking is quite accurate but remains some bias due to the error caused by camera alignment. Thus we impose a human-object silhouettes loss term using a differentiable render [7] to refine the 3D object and SMPL-X to human-object foreground masks:

$$E_{\text{homask}} = \| \sum_{1}^{76} (I_j^{\text{homask}} - DR(V_t, \mathcal{M})\|_2^2, \tag{3}$$

where DR denote as differentiable rendering and I_j^{homask} denote as human-object mask computed from background matting. Finally, we also impose a marker corresponding team to prevent local minimal,

$$E_{\text{marker}} = \|\mathbb{1}^{\text{marker}} V_t(c_j, p_n) - s)\|_2^2, \tag{4}$$

where $\mathbb{1}^{\text{marker}}$ denote as a binary indicator matrix that selects the vertices on the object mesh V_t at its marker location and s is the tracking data by the marker.

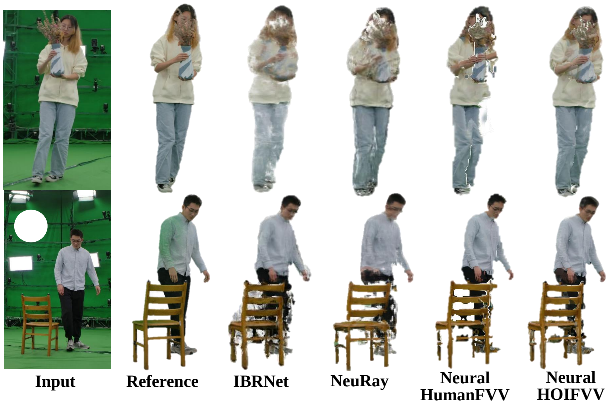

Figure 1. We provide more quality result on sparse-view rendering tasks and evaluate on IBRNet [15], NeuRay [8], NeuralHumanFVV [13] and our baseline NeuralHOIFVV.

4. More Details of NeuralHOIFVV

To validate that the HODome dataset is able to support novel view synthesis under sparse view settings, we propose a naive method called NeuralHOIFVV. We adapt the neural texture blending method introduced in [16]. Instead of using precise depth, we use the reprojection of 6-view PIFu [11] trained on HODome to generate the coarse depth maps of the target view and input views. After getting the coarse depth map of the target view, we use it to warp the input images and input coarse depth maps into the target view. Then we use the same network as introduced in [16] to predict the two channels' feature maps $W = (W_1, W_2)$ representing the blending weights of warped images. Note that different from [16], we do not have a coarse rendering image at the target view generated by the textured mesh. Our blending result is obtained only by using the blending map W and the warped images. We provide also provide more quality result on sparse-view rendering tasks and evaluate on IBRNet [15], NeuRay [8], NeuralHumanFVV [13] and our baseline NeuralHOIFVV as shown in Fig. 1.

5. More Experiment Results

To better evaluate the components of joint optimization, we further do an additional quality analysis of different constraint terms. Note that we have no ground truth of the specific tracking, thus we conduct qualitative evaluations

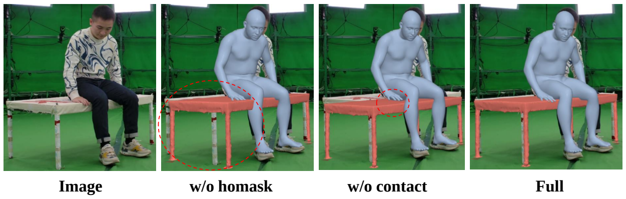

Figure 2. **Qualitative evaluation**

only. Fig. 2 shows the quality result by ablating different components. "w/o homask" and "w/o contact" respectively denotes the result obtained without using the human-object masks term E_{homask}, and without using the contact term E_{contact}. It demonstrates that the term of human-object masks E_{homask} can effectively alleviate the bias caused by calibration and alignment and our contact map further ensures the realistic interaction between humans and objects. We also provide more quality results sampled from our HODome dataset as shown in Fig. 3. Fig. 5 provides the gallery of data examples captured by our multi-view HODome with 76 synchronized high-resolution RGB cameras and Optitrack system. Our dataset includes a variety of human-object under various interactions. Fig. 4 shows the objects of data sampled from our HODome dataset.

References

[1] Naturalpoint, inc. motion capture systems. https://optitrack.com/. 6. 1

[2] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992. 1

[3] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 2

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1

[5] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, pages 281–299. Springer, 2022. 2

[6] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 1

[7] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018. 2

[8] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *CVPR*, 2022. 2

[9] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 1

[10] 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. 2

[11] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2

[12] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2291–2300, 2020. 1

[13] Xin Suo, Yuheng Jiang, Pei Lin, Yingliang Zhang, Minye Wu, Kaiwen Guo, and Lan Xu. Neuralhumanfvv: Real-time neural volumetric human performance rendering using rgb cameras. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6222–6233, 2021. 2

[14] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 1

[15] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 2

[16] Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, et al. Human performance modeling and rendering via neural animated mesh. *arXiv preprint arXiv:2209.08468*, 2022. 1, 2

[17] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. 1

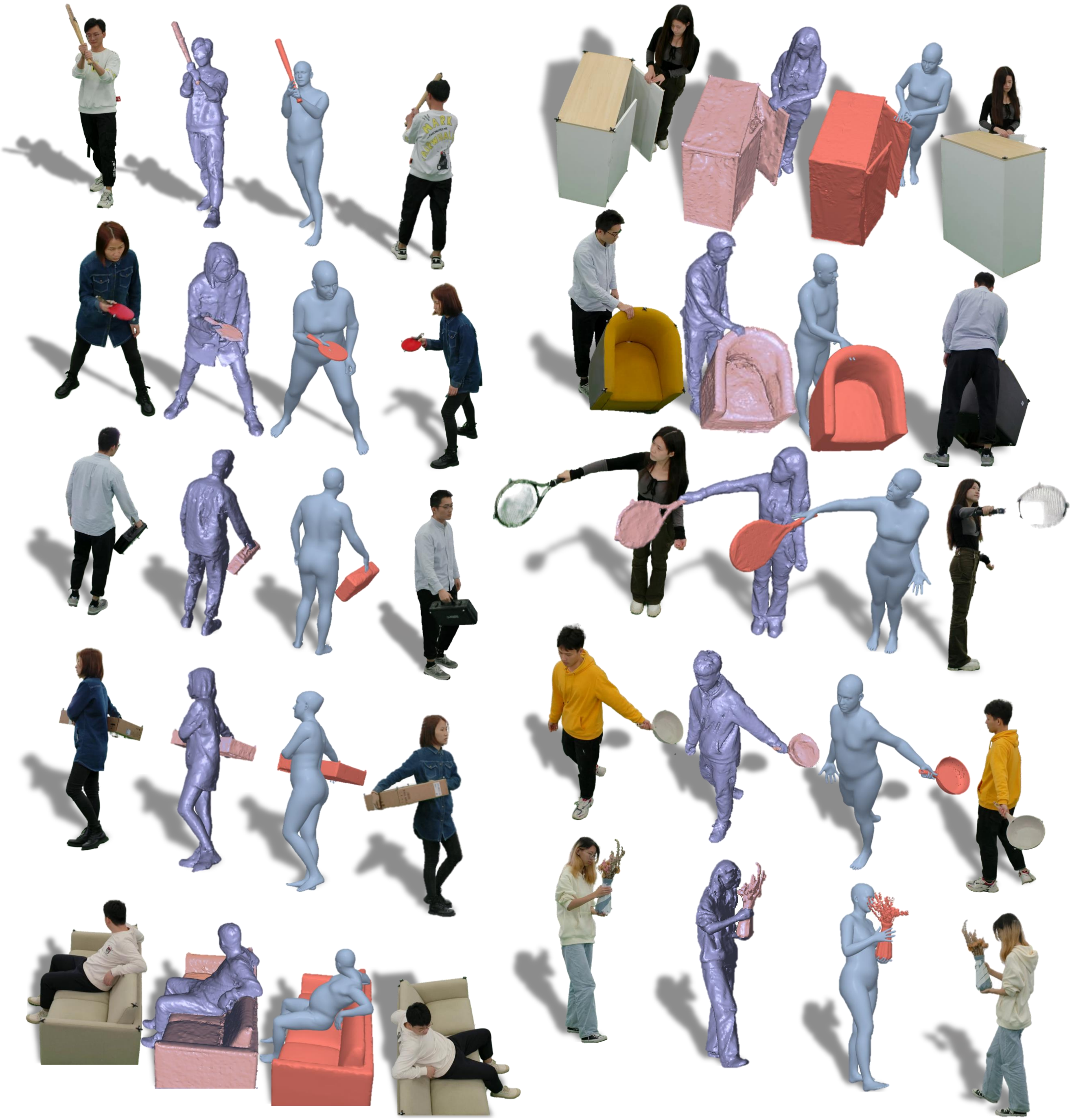

Figure 3. **More quality results.**

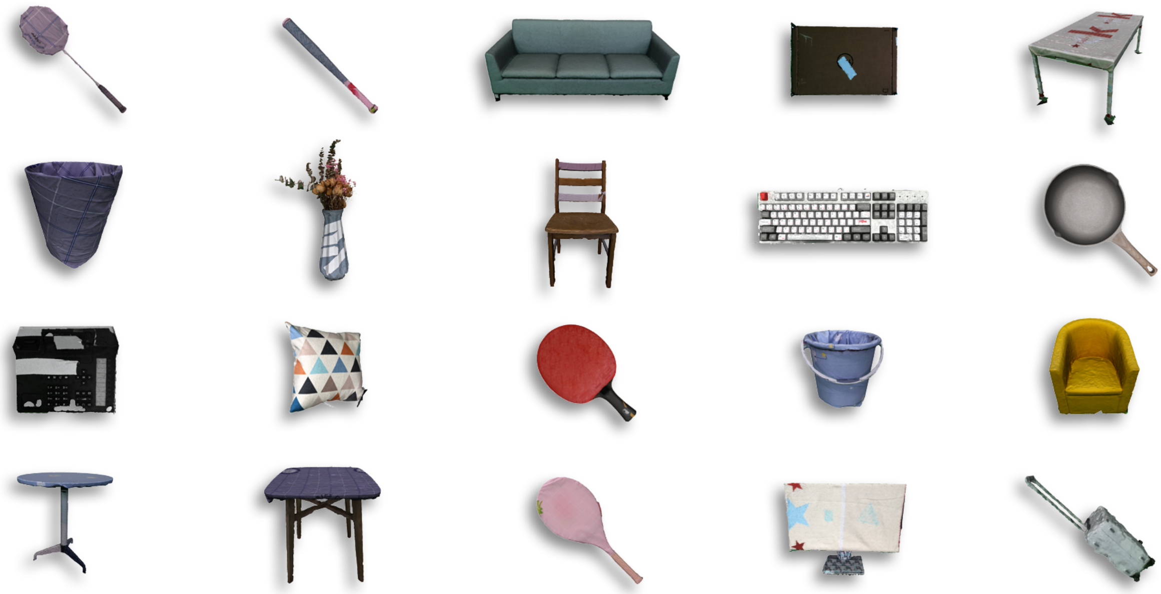

Figure 4. The objects sampled from our HODome dataset

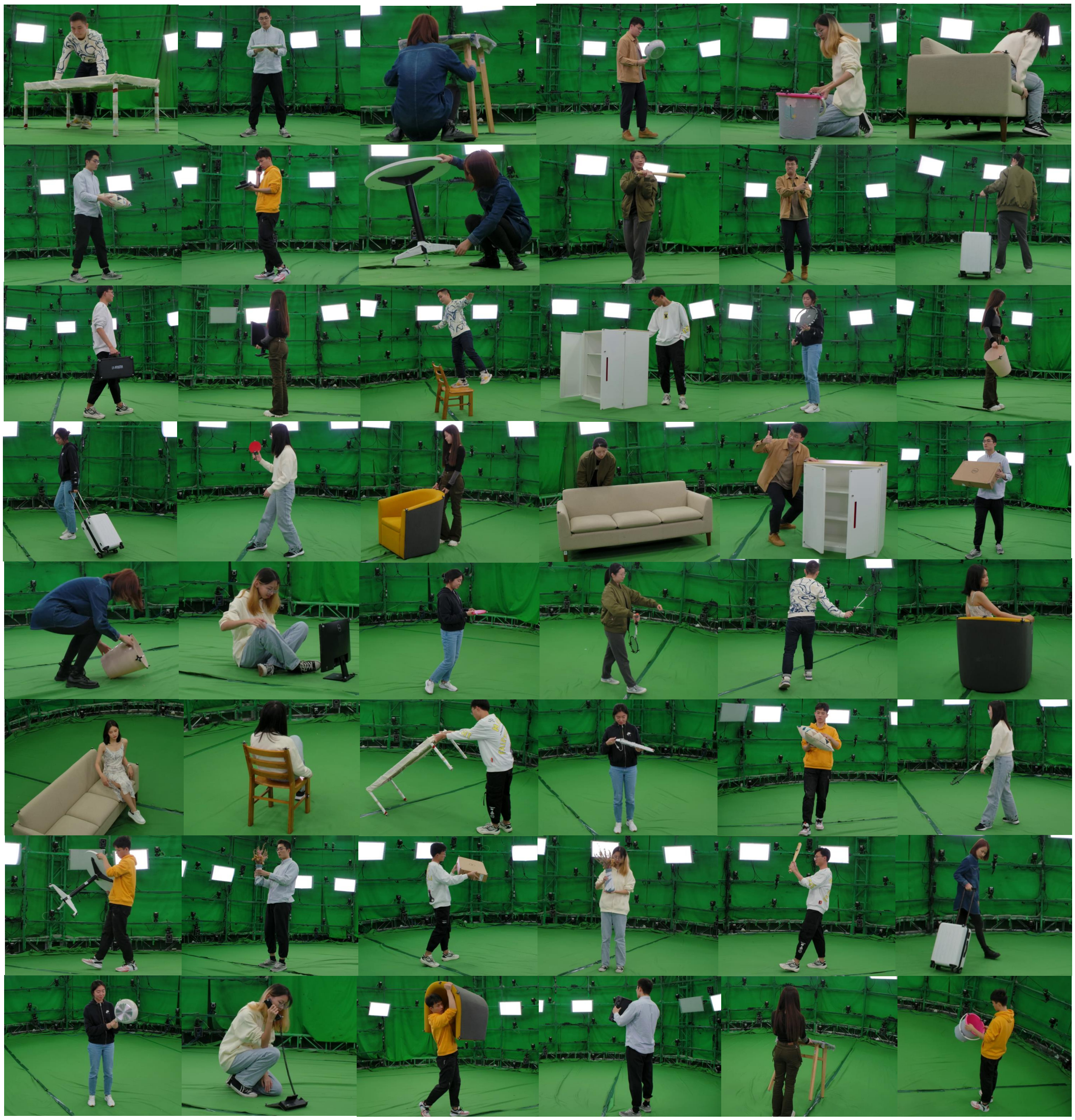

Figure 5. Data examples were captured by our multi-view HODome with 76 synchronized high-resolution RGB cameras and Optitrack system. Our dataset includes a variety of human-object under various interactions