A. Model

Why do we mainly focus on the articulated object pose estimation? Existing studies of HOI usually estimate the pose of humans and objects jointly, hoping the two estimations to improve each other. However, due to the imbalanced attention received by the human and articulated object pose estimation, we empirically observe that the object pose estimation is far from well-solved compared with human pose estimation, especially in scenarios where dense interactions and occlusions appear. Therefore, we mainly focus on improving the untouched articulated object pose estimation under human pose guidance in this paper, leveraging the mature and stable techniques of human pose estimation. Such motivation is similar to Ye *et al*. [61], which focuses on improving the reconstruction of interacting objects rather than the hand. Of note, our dataset still supports human pose estimation and encourages efforts that potentially improve it. Tab. 4 from the main text shows that incorporating the human pose information can significantly improve the object pose estimation performance, which verifies our assumption. The ground-truth human pose can further improve the object pose estimation by a large margin, demonstrating that further optimization of human poses is promising. It is regarded as one important step in our future work.

Coordinates for reconstruction and optimization Both object reconstruction and optimization are conducted in the human local coordinate centered at the pelvis bone of the SMPL model with the same orientation as the human root. We set a $2m \times 2m \times 2m$ cubic as the boundary for voxelization and interaction prior.

Details of Reconstruction Model The ResNet-101 to extract features from the input image outputs feature vectors of 1024 dimension. The feature vector is then mapped to $128 \times 8 \times 8 \times 8$ to form the input of the 3D blocks. Each 3D block consists of a 3D convolutional layer with a kernel size 3x3x3, a max-pooling layer and a batchnorm layer. The number of channels of the first 3D convolutional layer is 129, inlucding 128 for the object and 1 for the concatenated human occupancy. The chanels for the object are reduced by half after each block. The max-pooling layer has a pooling size of 2x2x2 and a stride of 2 in all dimensions.

Adapting D3D-HOI as baseline for our task The D3D-HOI method [58] is originally designed for hand-centric interactions, such as opening and closing a microwave, and contains manually defined optimization objectives, such as distance between hand and object. We make the following modifications to D3D-HOI to better fit the context of CHAIRS:

1. We replace the differentiable articulated object model in D3D-HOI by the pytorch_kinematics package (https://github.com/UM-ARM-Lab/pytorch_kinematics), which supports articulated objects with multiple links and joints.

Table A1: **Object reconstruction errors on BEHAVE dataset, with object kinematic structure and optimization.**

	Chair		Table		Yogaball		Suitcase	
	CD↓ (mm)	IOU↑ (%)	CD.↓ (mm)	IOU.↑ (%)	CD↓ (mm)	IOU↑ (%)	CD.↓ (mm)	IOU.↑ (%)
w/o HOI prior	134.5	11.35	161.6	10.53	106.37	30.53	161.0	**29.80**
w/ HOI prior	**127.3**	**14.22**	**152.2**	**12.86**	**98.79**	**33.75**	**158.4**	29.62

2. We changed the contact error in D3D-HOI to the distance between the hip joint and the center of the chair seat. Since the hip joint is usually higher than its nearby skin, we compute this error by adding a $20cm$ offset along the negative Y direction.
3. The orientation term in D3D-HOI encourages the human and the object to have opposite directions in "opening" and "closing" actions. We change this term to encourage the human to have the same orientation as the chair in the "sitting" case.

Data preparation for CHORE and PHOSA Since PHOSA requires predefined contact pairs as heuristics to reconstruct human-object interaction, we manually labeled each object mesh with contact maps corresponding to human body parts during the interaction. A part of the labeling results are shown in Fig. A2.

B. Additional Results

Generative model We evaluate the value of AHOI in CHAIRS by training conditional generative models [22] on both the CHAIRS dataset and the COUCH [67] dataset. Figure A3 shows that both models can generate realistic interactions with objects, and the model trained with CHAIRS can generate interacting poses with more full-body interactions. This observation confirms the value and the contribution of our CHAIRS dataset.

Qualitative results In Fig. A4, we qualitatively show more randomly selected results on the test set of CHAIRS. In general, our model predicts accurate object poses and shapes.

Qualitative comparisons We further compare reconstructions of our method against results from CHORE [56] and PHOSA [65] in Fig. A5. The qualitative comparison shows that our method can reconstruct interactions accurately.

In the wild In Fig. A6, we qualitatively evaluate the generalization power of our model with internet images and images captured in the wild.

Experimental results on the BEHAVE dataset We apply our method to the BEHAVE dataset [2] to evaluate the generalizability of the reconstruction and HOI prior model. We select four objects from the object list with rich full-body HOI, namely a chair, a square table, a yoga ball, and a suitcase. Our method is tested under the full object knowledge setting. We separately train object reconstruction and HOI prior models for each object. Different kinds of interaction

(*e.g.*, move and sit for the square table) are mixed up in one model. We show quantitative results in Tab. A1 and qualitative results in Fig. A7. We observe that although the metrics drop numerically, our model can still reconstruct the poses of the interacting objects.

C. Dataset

C.1. Data collection

Object gallery We render all objects in CHAIRS in Fig. A8. Parts are colored according to category.

Instructions Each participant was instructed to sit down before and after each instruction for synchronization. Participants can stand up and walk around while performing an instruction. All physical interactions were performed with the sittable objects. All other objects that appeared in the instructions (table, person, phone, *etc*.) required participants to interact by imaging their presence.

1. Pick up an object from the ground.
2. Talk to someone next to you.
3. Relax alone at home.
4. Listen to your friend talk while propping your head with your hand.
5. Sit and play with your phone.
6. Sit with your hands on the seat.
7. Think with your head lowered.
8. Your neck feels uncomfortable.
9. Grab a thing from the desk behind you.
10. Move the chair forward.
11. Lean on the back. Adjust or rock it if you can.
12. Move the chair.
13. Adjust the chair.
14. Sit with a twisted posture.
15. Sit with your feet on the footstep or the footrest.
16. Change the pose of your legs.
17. Stretch a little in the chair.
18. Change to another pose of sitting.
19. Adjust the height of the seat.
20. Walk around the chair and sit down.
21. Move, rock, or rotate the chair.
22. Your back feels uncomfortable.
23. Lean your head on the headrest. Adjust it if possible.
24. Stretch your back in the chair.
25. Talk to the person behind you.
26. Move the chair backward.
27. Lay in the chair.
28. Put your arms on the armrests. Adjust them if you can.
29. Move the chair to your left.
30. Move the chair to your right.
31. Adjust the seat.
32. Pick up a heavy object from the ground.

We only sample instructions that are *compatible* given an object. For example, "Lean on the back" is *not compatible* for all stools. Figure A9 shows diverse performances in CHAIRS.

Recruitment Due to the complex nature of data collection that requires physical presence at the scene while wearing MoCap suits, all participants were voluntary colleagues. Participants were compensated with a gift with a value of $4 USD for every 18 sequences recorded.

Body and hand shape We use optical trackers to record the positions of each participant's head, two hands, and two feet. We then optimize the body shape parameter β of the SMPLX model to fit the tracker positions. We rely on SMPLX's default hand shape parameter since our primary focus is not to model dexterous hand-object interactions.

Motion capture system We used a Noitom Virtual Production Solution (VPS) camera system and a Noitom Perception Neuron Studio IMU system. The cameras each have 1280x1024 resolution, 210 fps, <5ms latency, 3.6mm F#2.4 lens, 81 deg horizontal FoV, and 67 deg vertical FoV.

C.2. Post Processing

Spatial alignment Our data collection system consists of multiple pieces of hardware, including 4 Azure Kinect DK cameras and a hybrid MoCap system. Each camera and the MoCap system have their own coordinate systems. We use OpenCV and an Aruco checkerboard to register all cameras to the camera space of the left-most camera and align it with the MoCap's coordinate frame with an Iterative Closest Points (ICP) algorithm.

Given the transformation matrices of the Kinect cameras, we apply a custom ICP algorithm to refine both the multi-view point clouds and the registration of Kinect and Mocap. We base our method on plane-to-plane correspondences [44] to alleviate the sensitivity to outliers, disturbances, and partial overlaps. Given the source point set $P = \{p_i, i = 1, ..., N\}$ captured by the Kinect depth cameras and the target set $Q = \{q_i, i = 1, ..., M\}$ reconstructed from the MoCap system, the goal is to calculate the optimum transformation matrix T, such that $TP^T = Q^T$. Following point-to-point ICP [6], we first find the nearest points $\widetilde{q}_i$ in Q to each p_i in P. Next, we iteratively update T to minimize the Mahalanobis distance between P and Q:

$$T = \arg\min_{T} \sum_{i=1}^{M} d_i^T (C^Q_{n,\tilde{i}} + T C^P_{n,i} T^T)^{-1} d_i \qquad \text{(A1)}$$

where d_i is the corresponding Euclidean distance between p_i and $\widetilde{q}_i$, $C^Q_{n,\tilde{i}}$ and $C^P_{n,i}$ the covariance matrix calculate by the n nearest points around $\widetilde{q}_i$ in Q and p_i in P. Finally, we use Anderson Acceleration [12] for a faster convergence to a fixed point.

Temporal alignment Observed images and poses in CHAIRS come from two independent systems (*i.e.*, MoCap and Kinect) without clock synchronization. Since both systems run steadily at 30 Hz, the two recorded data streams

have a constant difference in time. We use a time-lagged cross-correlation (TLCC) [45] algorithm to align the two systems temporally.

Specifically, we first extract the heights of the subject's head and two hands from both systems. For our MoCap system, we can directly read the joint positions with forward kinematics. For the Kinect cameras, we obtain the human joint positions with the Kinect Body Tracker SDK. Next, we compute the first-order differential on each sequence and compute the time offset between the differentials of each joint using TLCC. Finally, by measuring the peak of the TLCC correlation, we obtain three offsets (one for each joint); we use the median of the three offsets as our final temporal offset.

D. Compliance

List of code, data, models used, and their licenses We used the following assets. Please find the licenses of corresponding assets in the directories inside square brackets.

- SMPL-X [38] model and body [license/smplx-model,license/smplx-body.txt]
- ExPose [7] model and code [license/expose.txt]
- FrankMocap [42] model and code [license/frankmocap.txt]
- PARE [27] model and code [license/pare.txt]
- Category-Level Articulated Object Pose Estimation [28] model and code [*No license information found.*]
- Metropoly rigged 3D people (used in main paper Fig.3 and supplementary video)
- D3D-HOI [58] code [*No license information found.*]
- iStock [https://www.istockphoto.com] images used for in-the-wild evaluations. [license/istock.txt]

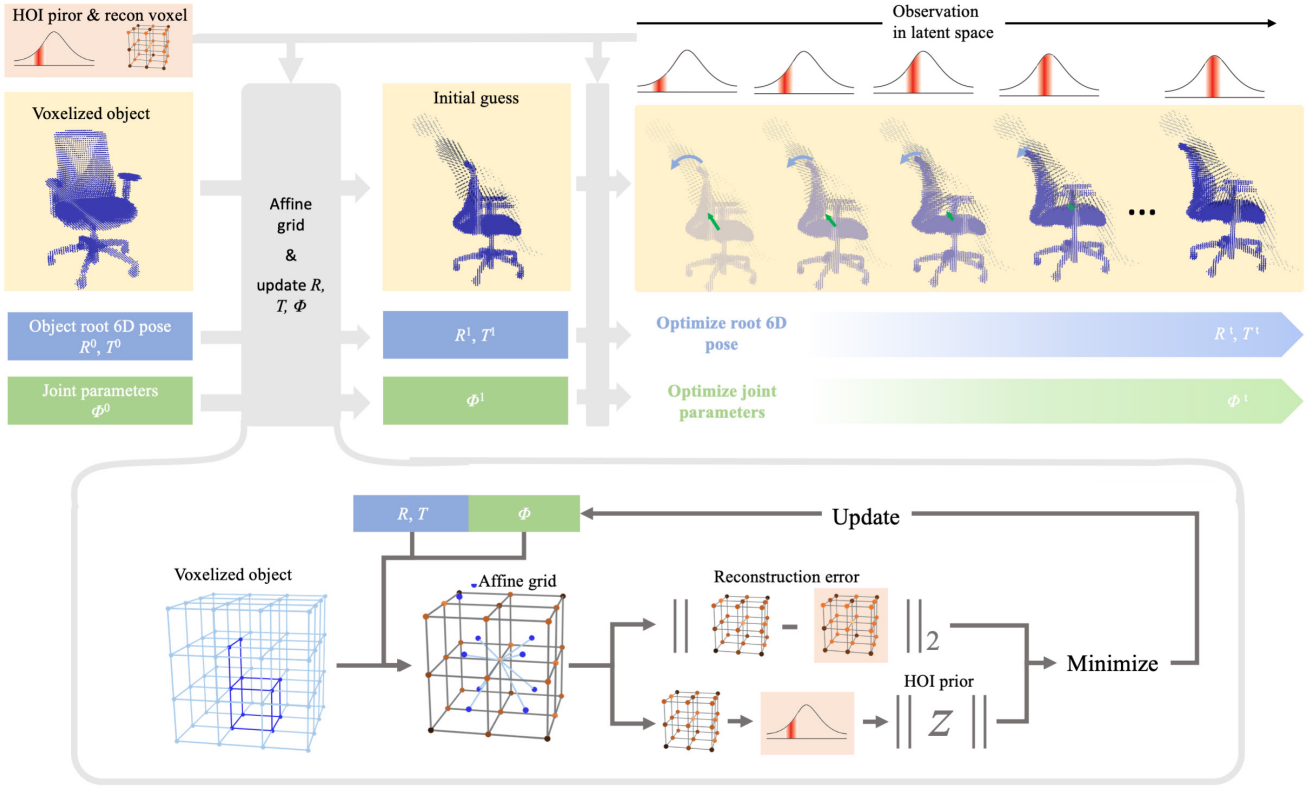

Figure A1: **Detailed diagram of the optimization process.**

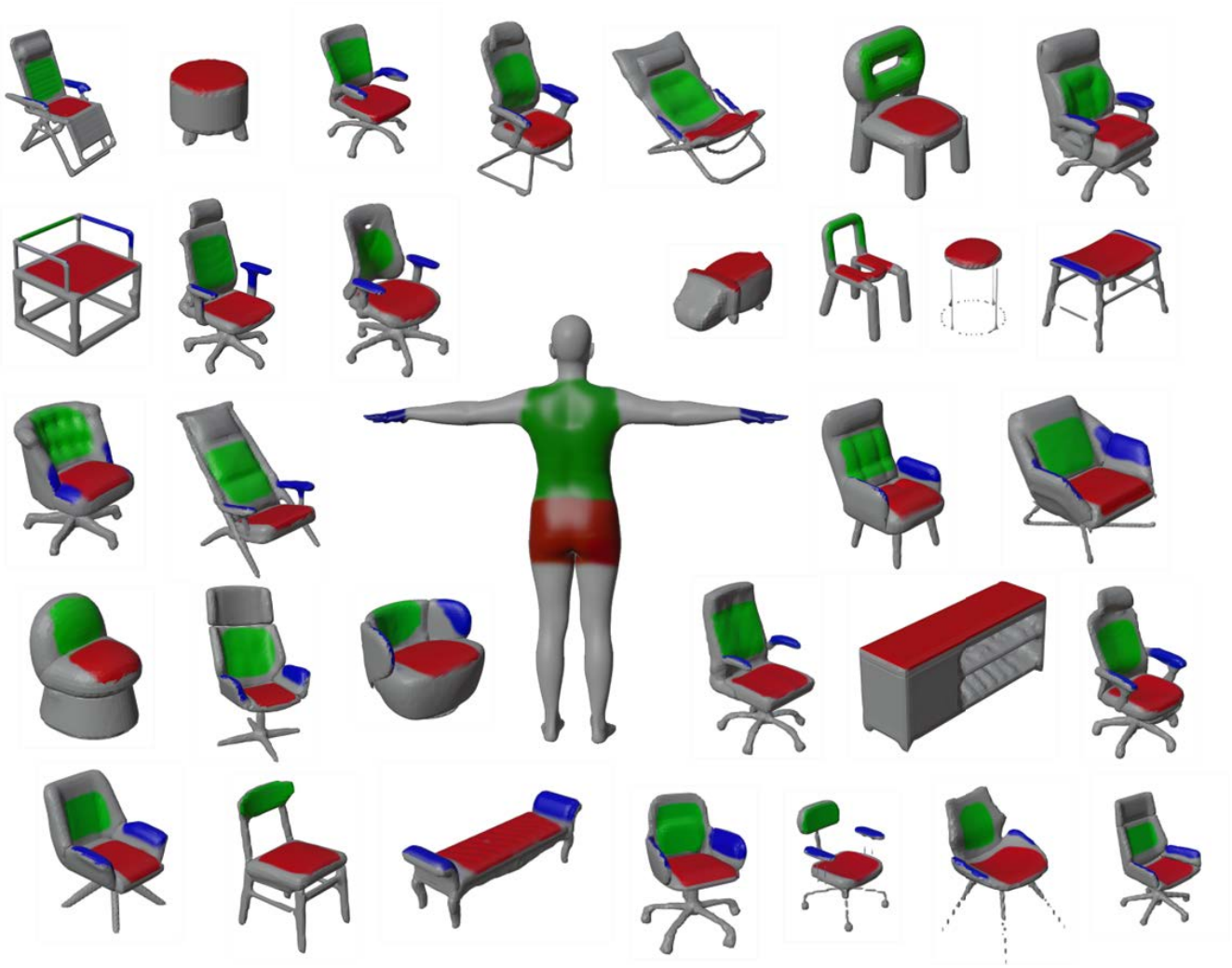

Figure A2: **Labeled contact maps** We use three colors to show the mappings of the surfaces on human bodies and objects that frequently get in touch during interactions.

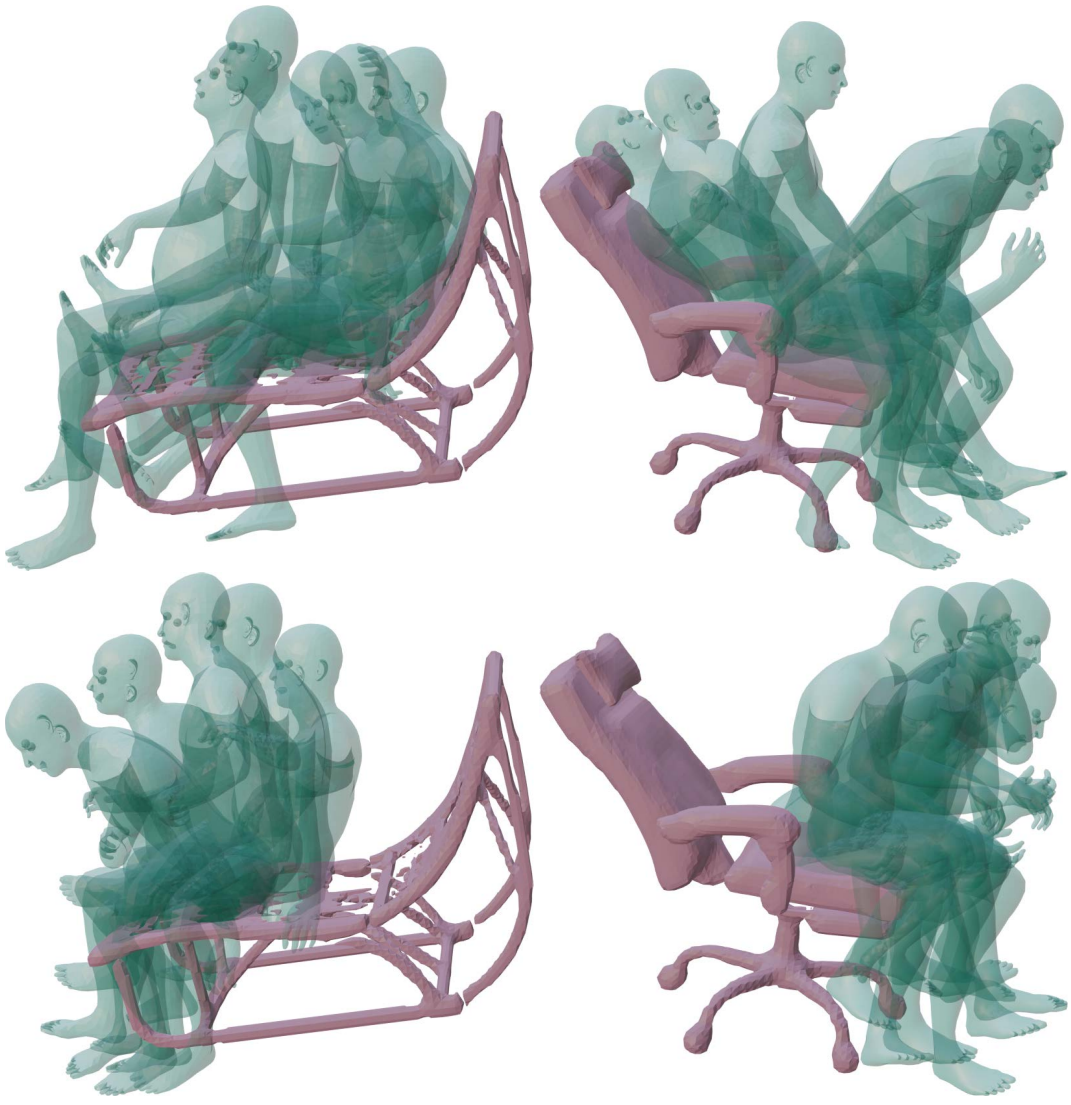

Figure A3: **Generated interacting human poses.** Top: model is trained on CHAIRS; bottom: model is trained on COUCH [67].

Figure A4: **Additional qualitative results of our model on the test set of CHAIRS.**

Figure A5: **Qualitative comparisons.** From left to right: RGB image, CHORE reconstruction, CHORE reconstruction from second view, PHOSA reconstruction, PHOSA reconstruction from second view, **our reconstruction**, and **our reconstruction from second view**. Results show a clear advantage of our method in modeling interactions.

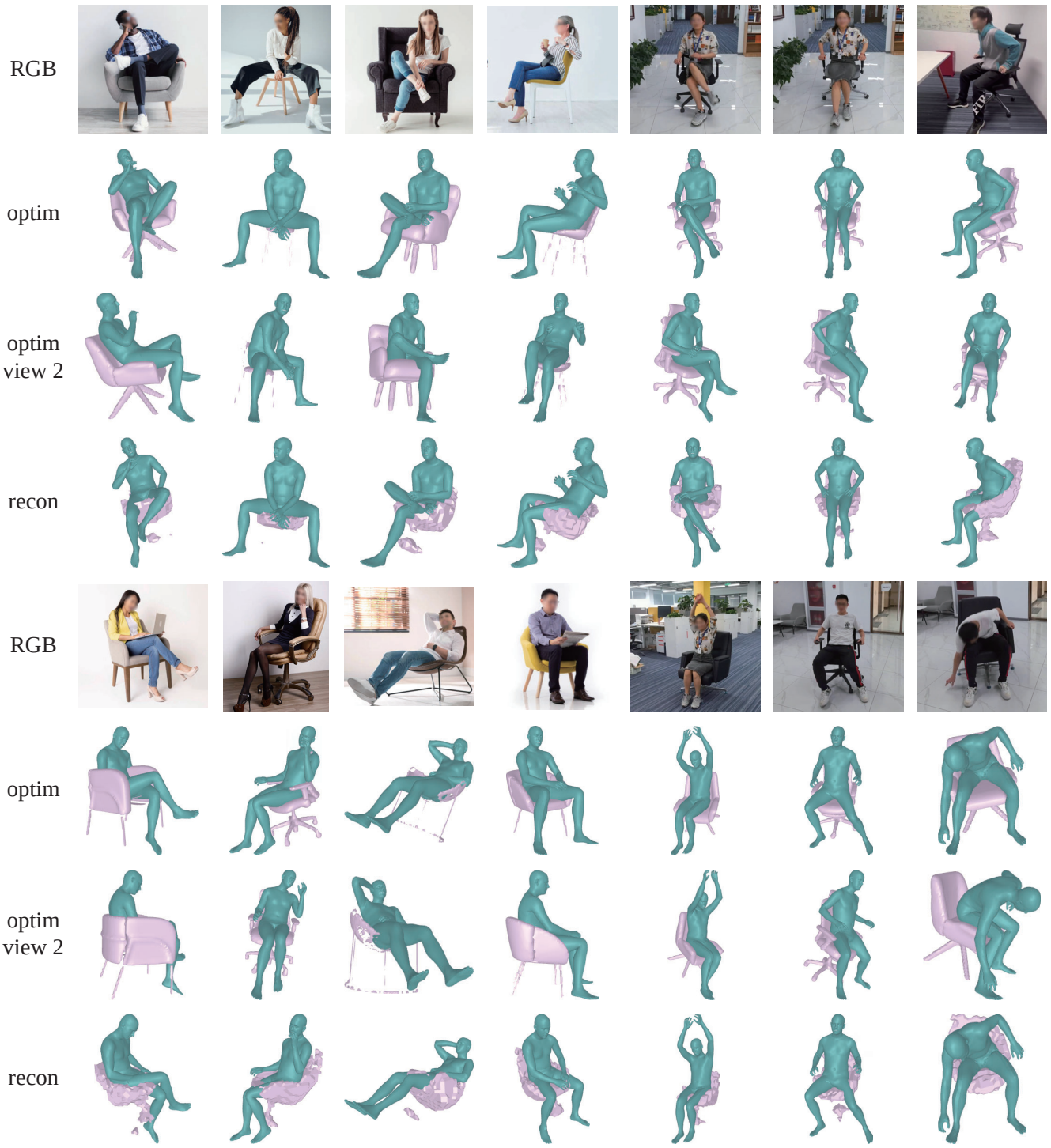

Figure A6: **Qualitative results of running our model on images captured in the wild.**

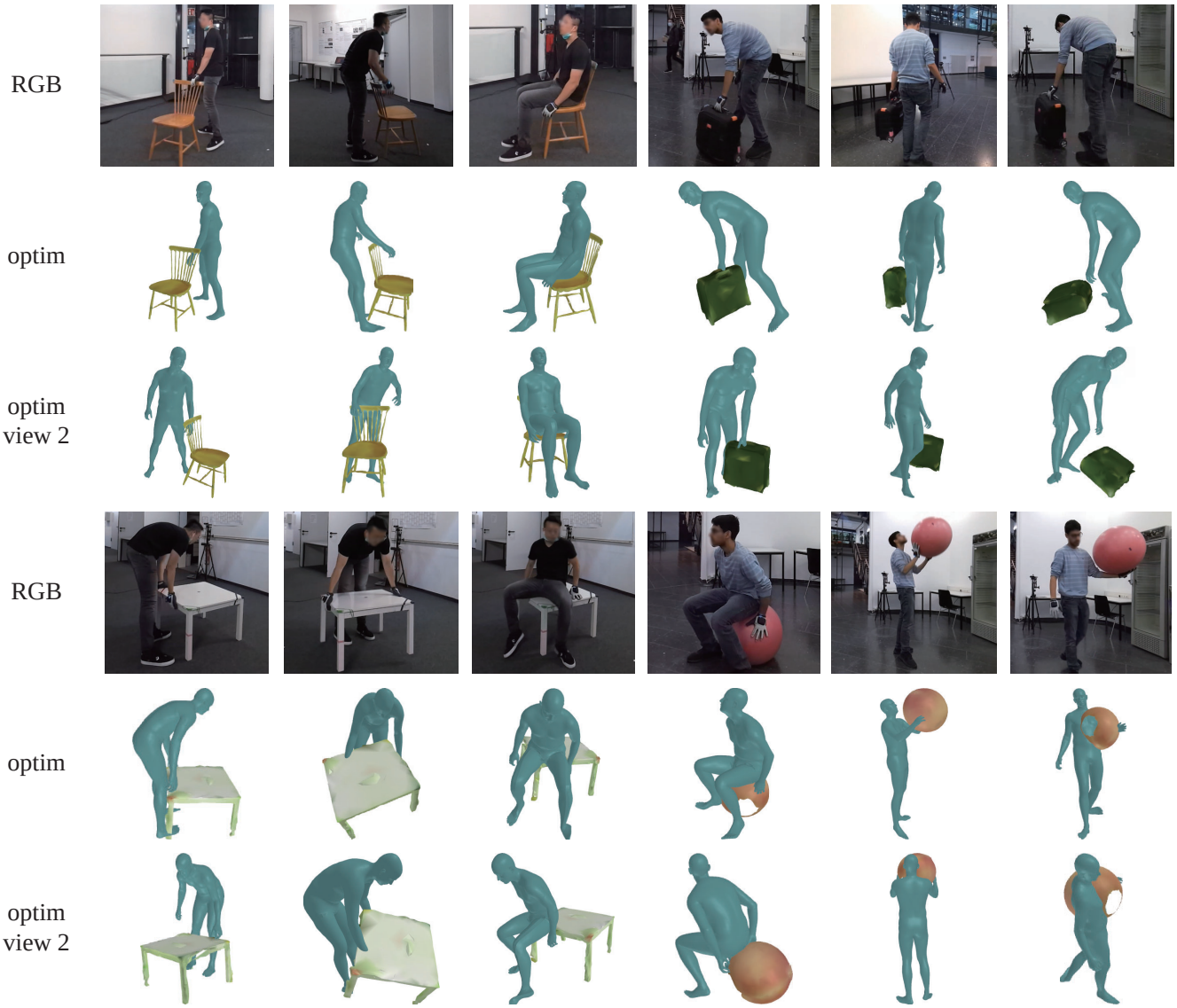

Figure A7: **Qualitative results of running our model on images from the BEHAVE [2] dataset.**

Figure A8: **Sittable objects in CHAIRS.** The first six rows are the objects in the training set, whereas the last row shows the ones in the test set.

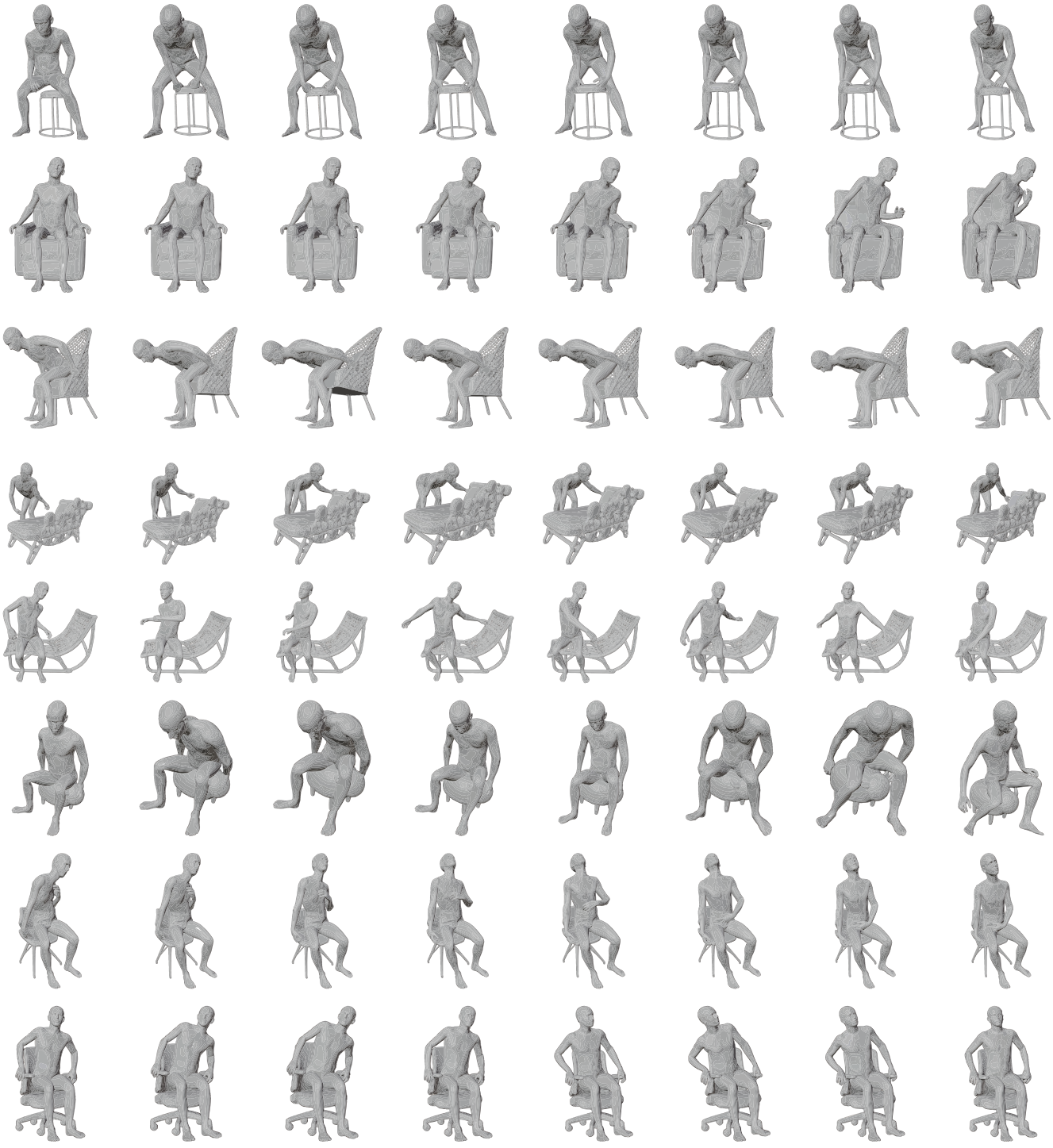

Figure A9: **Performances of different participants on different objects with the same instruction.** The first four rows show four performances of the instruction "Move the chair." The second participant rotated the chair with a small angle. The last four rows show four performances of the instruction "Stretch a little in the chair."