



CHAIRS: Towards Full-Body Articulated Human-Object Interaction

Nan Jiang^{1,2,*†}, Tengyu Liu^{2,*}, Zhexuan Cao^{2,3†}, Jieming Cui²,
Yixin Chen², He Wang¹, Yixin Zhu¹, Siyuan Huang²

*Equal contributors [†] Work done during an internship at BIGAI  yixin.zhu@pku.edu.cn, syhuang@bigai.ai

¹ Peking University ² Beijing Institute of General Artificial Intelligence (BIGAI) ³ Tsinghua University

<https://jnnan.github.io/project/chairs/>

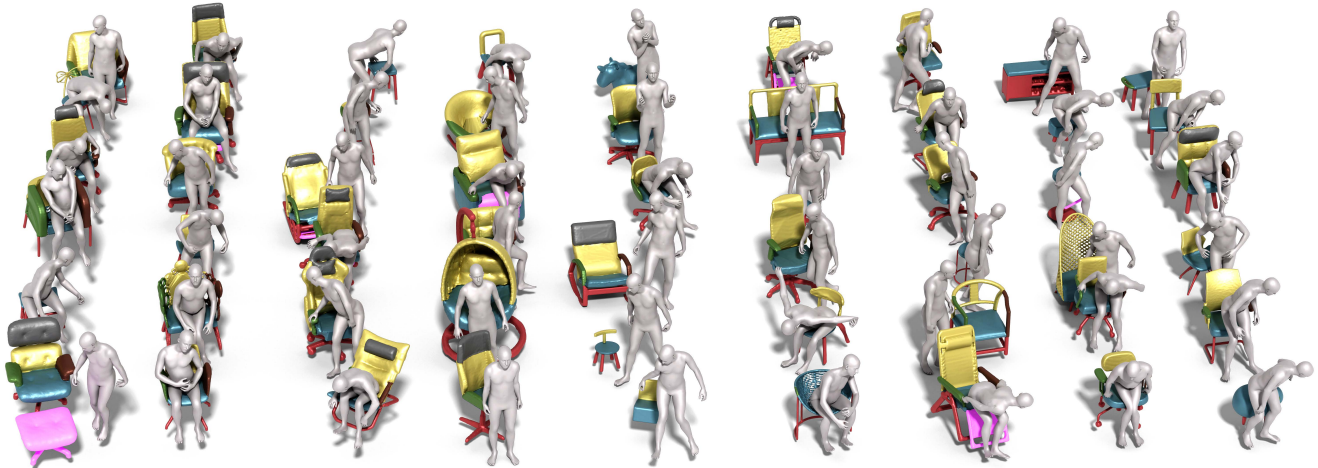


Figure 1. **Examples of the proposed CHAIRS dataset.** It contains fine-grained interactions between 46 participants and 81 sittable objects with drastically different kinematic structures, providing multi-view RGB-D sequences and ground-truth 3D mesh of humans and articulated objects for over 17.3 hours of recordings.

Abstract

*Fine-grained capturing of 3D Human-Object Interactions (HOIs) boosts human activity understanding and facilitates downstream visual tasks, including action recognition, holistic scene reconstruction, and human motion synthesis. Despite its significance, existing works mostly assume that humans interact with rigid objects using only a few body parts, limiting their scope. In this paper, we address the challenging problem of Full-Body Articulated Human-Object Interaction (f-AHOI), wherein the **whole** human bodies interact with articulated objects, whose parts are connected by movable joints. We present **Capturing Human and Articulated-object InteRactionS (CHAIRS)**, a large-scale motion-captured f-AHOI dataset, consisting of 17.3 hours of versatile interactions between 46 participants and 81 articulated and rigid sittable objects. CHAIRS provides 3D meshes of both humans and articulated objects during the entire interactive process, as well as **realistic and physically plausible** full-body interactions. We*

show the value of CHAIRS with object pose estimation. By learning the geometrical relationships in HOI, we devise the very first model that leverage human pose estimation to tackle the estimation of articulated object poses and shapes during whole-body interactions. Given an image and an estimated human pose, our model first reconstructs the pose and shape of the object, then optimizes the reconstruction according to a learned interaction prior. Under both evaluation settings (i.e., with or without the knowledge of objects' geometries/structures), our model significantly outperforms baselines. We hope CHAIRS will promote the community towards finer-grained interaction understanding. We will make the data/code publicly available.

1. Introduction

In computer vision and robotics, Human-Object Interaction (HOI) [29, 30, 62, 64] is the crux of modern fine-grained human activity understanding. In this work, we tackle a challenging problem of Full-Body Articulated Human-Object Interaction (f-AHOI), which requires (i) building on

kinematic-agnostic object representations for **articulated** objects, and (ii) modeling the fine-grained spatial-temporal interactions between objects and the human **whole-bodies**. Specifically, we address the problem of object pose estimation under f-AHOI, as the reconstruction of foreground 3D human poses is relatively easy from the front-view cameras.

Object pose estimation under f-AHOI is inherently challenging due to three primary reasons:

Lack f-AHOI datasets that captures human whole-bodies interacting with articulated objects Despite recent progress in 3D HOI, prior works either assume that the objects to be interacted with are rigid [2, 28, 47, 65], or the interactions involve only part of human bodies (*e.g.*, only hands [11] or upper limbs [14, 58]). These assumptions oversimplify daily interactions; humans use different body parts to interact with articulated objects composed of movable parts, such as cabinets and office chairs, calling for a dataset with a finer-grained level of interactions.

Large variance of object kinematic structures Objects related to f-AHOI show significant divergence in their kinematic structures, even within the same category; objects possess various numbers and types of parts and joints. Such diversity is in stark contrast to the articulated objects modeled in literature [11, 14, 31, 58], which assumes limited or no variety in kinematic structures. Reconstructing objects with diverse geometries and structures remains challenging.

Complex and subtle relations between human body parts and object parts Interacting with articulated objects involves complicated spatial and physical relationships, with severe occlusions and rich contacts that incapacitate conventional pose estimation methods that rely on pointcloud template-matching [19, 27, 39, 50, 65]. The contact-rich property also challenges capturing the fine details in reconstruction, as even small errors can result in implausible interactions such as penetration and floatation.

We devise the following three solutions to tackle the above three challenges, respectively.

To address the scarcity of f-AHOI dataset, we present CHAIRS, a large-scale f-AHOI dataset with multi-view RGB-D sequences. As shown in Fig. 1, CHAIRS includes 17.3 hours of diverse interactions among 46 participants and 81 sittable objects (*e.g.*, chairs, sofas, stools, and benches), 28 of which have movable parts; each frame includes 3D meshes of whole-body humans and objects. In this work, we focus on interactions with sittable objects; they are diverse in structure and contain distinct movable parts that afford various whole-body human interactions.

To model diverse kinematic structures, we extend the task of object pose estimation to the challenging setting of kinematic-agnostic pose estimation. Existing datasets [31, 52, 56] and methods [1, 27, 36, 48] for articulated objects assume similar or identical kinematic structure for intra-class objects; this assumption fails when dealing with real-

world daily objects. The kinematic structures in CHAIRS vary from a rigid stool with no articulation to swivel chairs with 7 movable parts. Specifically, we relax the assumption of limited kinematic structures to an open set of flexible but known structures. Given an observed image, an estimated human body from the image, and the kinematic structure of the object of interest, we aim to reconstruct the pose and shape of the object.

To disambiguate the complex and subtle relations during the whole-body interactions with articulated objects, we devise a novel pose estimation approach that leverages the fine-grained interaction relationships to reconstruct the interacting object. A common solution in the prior arts [2, 16, 65] is to manually label each object mesh with contact maps corresponding to human body parts. In comparison, our method exploits the complex and fine relationships with a reconstruction model and an interaction prior learned with conditional Variational Auto-Encoder (cVAE), which avoids the pre-defined knowledge through mundane annotation. Specifically, our approach first reconstructs coarse shapes and poses of the objects, then optimizes the details with the learned interaction prior.

Our **contributions** are four-fold. (i) We present CHAIRS, a large-scale multi-view RGB-D dataset with diverse and high-quality 3D meshes of human and articulated objects. (ii) We extend articulated object pose estimation to the challenging setting of f-AHOI. (iii) We devise an object pose estimation approach agnostic to the articulation structure. (iv) We propose a generic interaction prior that captures the fine-grained interactions with sittable objects and facilitates the pose estimation.

2. Related Work

3D Human-Object Interaction (HOI) HOI research has evolved from detecting interactions in 2D images [4, 13, 29, 30, 40, 62, 64] to reconstructing [5, 16, 44, 54, 58, 63, 66] and generating [17, 21, 51, 53, 57] 3D interactions in 3D scenes. Notably, PiGraph [44] captures human daily activities, Rosinol *et al.* [43] represent the interactions with a graph structure, and Hassan *et al.* [16, 58] reconstruct 3D human-scene interactions. However, these works rely on visual observations to collect ground-truth 3D poses, which leads to inaccurate reconstruction under partial observation. Meanwhile, MoCap systems [2, 11, 47] provide fine-grained 3D interactions between humans and 3D objects. In particular, GRAB [47] and ARCTIC [11] focus on interactions with small objects, such as grasping and holding, whereas BEHAVE [2] captures the interactions with daily objects. However, most existing works focus on either rigid objects or articulated objects but in the domain of hand-object interactions. In comparison, our CHAIRS dataset provides realistic *whole-body* interactions (*e.g.*, move the bench, relax in the chair) with diverse articulated objects.

Table 1. Comparisons between CHAIRS and other HOI datasets.

Dataset	# object	# participants	# instance	# hours	fps	# view	articulated objects	human	annotation type
PROX [16]	/	20	/	0.9	30	1	No	Whole-body	single-kinect
GRAB [47]	51	10	4	3.8	120	0	No	Whole-body	mocap
BEHAVE [2]	20	8	6	0.14	30	4	No	Whole-body	multi-kinect
ARCTIC [11]	10	9	1	1.2	30	8+1	Yes	Two hands	mocap
D3D-HOI [58]	24	5	/	0.6	3	1	Yes	Whole-body	manual
CHAIRS (Ours)	81	46	32	17.3	30	4	Yes	Whole-body	mocap

Articulated Human-Object Interaction (AHOI) Articulated Human-Object Interactions (AHOIs) build on part-level object representations and model the fine-grained spatial-temporal interactions between human and articulated objects [14]. To date, the most relevant works are D3D-HOI [58], ARCTIC [11], and 3DADN [41]. Specifically, D3D-HOI [58] collects a video dataset of humans interacting with containers such as microwaves and refrigerators, ARCTIC [11] collects a motion-captured RGB-D dataset of hand-object interactions with articulated objects, whereas 3DADN [41] annotates movable object parts from internet videos as 3D planes with rotations. Of note, all objects only have one revolute joint connecting two rigid parts, and all interactions captured focus only on hand-object interactions such as “open” and “close.” In comparison, we take one step further to study the whole-body AHOIs; most body parts interact with diverse articulated objects.

Contact-Rich HOI f-AHOI requires a more detailed HOI understanding. Despite the rapid growth of literature in 3D HOI, only a few involve full-body contacts either by reconstruction [16] or generation [15, 53, 67]. However, these prior arts are limited to interactions with static scenes and limited interactions. In comparison, our CHAIRS dataset contains diverse articulated objects and interactions.

Articulated Object Pose Estimation Estimating rotation and translation (*i.e.*, 6-DOF pose estimation) of rigid objects has recently attracted significant attentions [3, 9,

19, 24, 37, 39, 49]. Template-based methods are commonly adopted approach [20, 25, 50, 60] and have spurred a series of recent works in articulated object pose estimation [8, 27, 33]. Other methods rely on regression models [1] or implicit functions [23, 36, 48, 59]. Despite recent progress, these methods are based on a simplified assumption of consistent kinematic structures within each object category. Hence, the pose estimation models are designed and trained to estimate the attributes and states of a fixed set of joints. Although recent datasets on articulated objects [31, 32, 52] contain different kinematic structures, the diversity of kinematic structures is not the primary focus and thus is still limited. To overcome these shortcomings, we collect the CHAIRS dataset with diverse kinematic structures and devise models to handle 3D objects with various parts and kinematics.

3. The CHAIRS Dataset

A major obstacle in modeling AHOIs is the absence of accurate 3D annotations. In this work, we present CHAIRS, a large-scale AHOI dataset with multi-view RGB-D sequences. CHAIRS provides high-quality 3D meshes of humans and articulated objects during interactions, collected with an inertial-optical hybrid motion capture (MoCap) system and optimized for superior realism and physical plausibility. Tab. 1 shows the detailed comparison between CHAIRS and previous HOI datasets.

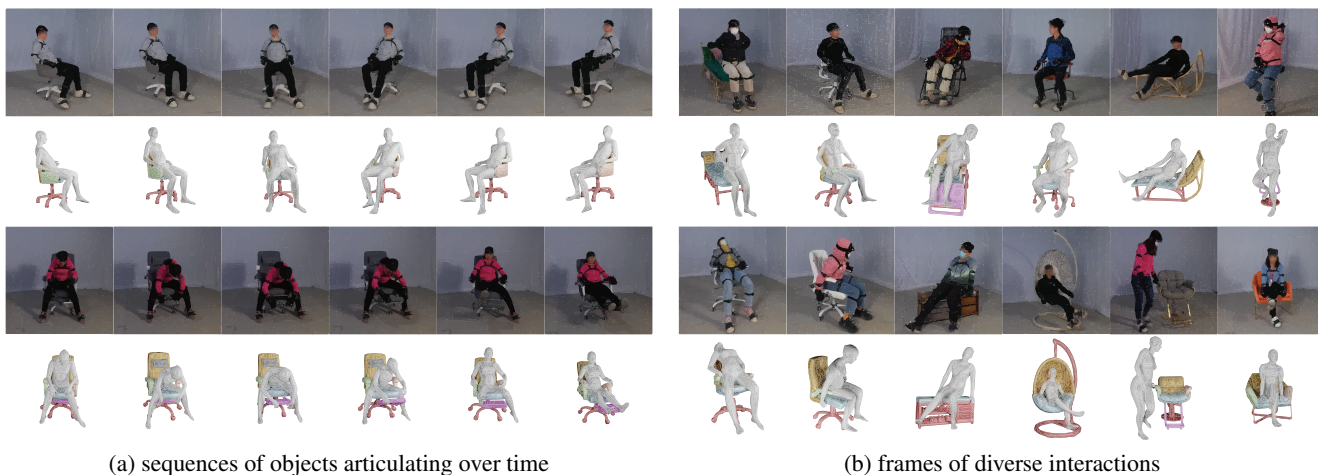


Figure 2. Examples from the proposed CHAIRS dataset. CHAIRS captures versatile AHOIs from carefully calibrated multi-view RGB-D cameras and provides fine-grained 3D meshes for both humans and articulated objects. We show (a) RGB frames and ground-truth meshes of AHOIs in sequences and (b) diverse types of AHOIs.

3.1. Data Collection

Summary CHAIRS has a total of 1390 sequences of articulated interactions between human and sittable objects, such as chairs, sofas, stools, and benches. Fig. 2 shows exemplar sequences of CHAIRS and object gallery. For each object, we asked 6 participants to record three sequences of interactions with it, yielding 18 sequences for each object. In each sequence, a participant was asked to perform 6 different actions. The actions were randomly chosen from a list of 32 interactions (*e.g.*, move the stool forward, relax on the sofa, spin the chair); see Supplementary Material for details. We ensure the data diversity with 46 participants and 81 objects. Only high-level instructions were provided to the participants to ensure natural performances.

Object Gallery CHAIRS features object collections with rich appearances and kinematic structures. The objects were selected and purchased online by maximizing the style variance; 28 of them have at least one articulated joint. We scanned the 3D meshes of each object with the Scaniverse app on an iPad Pro (11-inch, 2nd generation) and manually refined the geometries to remove artifacts. We define eight object functional parts and use the annotation tool [35] to segment the 3D meshes accordingly. When interacting with an object, participants were only provided with instructions compatible with the given object.

Camera and Hardware Setup As shown in Fig. 3, all the sequences were captured exclusively in a controlled laboratory setup, with a designated area of 5m×4m where all actions were fully visible to the cameras. Four multi-view front-facing Kinect Azure DK cameras were set up towards performed interactions. The cameras were well-calibrated and synchronized. To ensure high-quality ground-truth poses for both humans and objects, we adopted a commercial inertial-optical hybrid MoCap system in addition to the Kinect setup; see details in the next section.

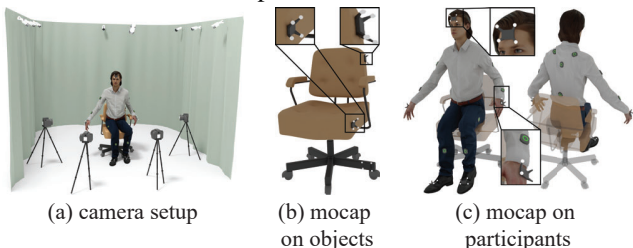


Figure 3. **The camera and hardware setup of data collection for the CHAIRS dataset.** We (a) set up 4 front-facing RGB-D cameras along with a set of motion capture cameras around the capturing site, (b) attach hybrid trackers to movable object parts, and (c) place 5 hybrid trackers and 17 IMUs on participants.

3.2. Motion Capture (MoCap) System

Hybrid MoCap Our MoCap system contains a MoCap suit with 5 hybrid trackers and 17 wearable Inertial Measurement Units (IMUs), a pair of gloves with 12 IMUs

each, an additional set of hybrid trackers, and a set of 8 high-speed cameras. A hybrid tracker is a rigid assembly of 4 optical markers and an IMU that can measure accurate 6D poses of itself under severe occlusion. We illustrate our data collection setup in Fig. 3. When capturing the pose of a human or object part, we can either use an IMU to record its global orientation or a hybrid tracker to record its 6D pose.

Articulated Object Capture Collecting the articulated pose of an object during interactions involves three steps. First, we arrange the object to its canonical pose and attach a hybrid tracker to each of its movable parts. Next, we compute the relative transformation between the trackers and the object part. During recording, we calculate the ground-truth 6D pose of each object part in real time based on the trackers’ pose. Finally, we fit the rigid parts to the object’s kinematic structure for high-quality object poses.

Human Body Capture We adopt the SMPL-X [38] representation for human poses and shapes. Participants were asked to wear a MoCap suit with 17 IMUs, a pair of MoCap gloves, and 5 hybrid trackers mounted on their heads, hands, and feet. Of note, the hybrid trackers capture 6D poses, whereas IMUs only measure global orientations. We optimize the human model’s shape parameters such that the reconstructed SMPL-X mesh aligns with the hybrid tracker positions. The MoCap system produces real-time estimated human poses and shapes during recording.

3.3. Post-processing

Data Alignment Kinect cameras and the MoCap system have separate 3D coordinates and clocks. We align the 3D coordinates of Kinect sequences with MoCap reconstructions based on plane-to-plane correspondences [45], which alleviate the sensitivity to outliers, disturbances, and partial overlaps. We align the temporal sequences from Kinect and MoCap using time-lagged cross-correlation [46], a typical approach to synchronize two sequences that shift relatively in time.

Penetration Removal Due to the limited number of sensors and discrepancies in limb lengths, implausible contacts and penetrations still exist in captured 3D interactions. To address this issue, we fix the physical glitches

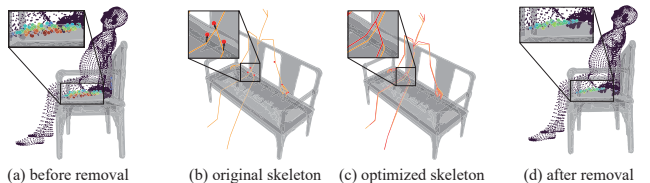


Figure 4. **Illustration of the penetration removal process.** (a)(d) Small purple points denote human vertices without penetration, whereas large colored points are those with penetration. Red points denote the most significant penetration, and blue barely in contact. (b)(c) Yellow lines denote the original skeleton, red markers the target joints to be optimized, and red lines the optimized skeleton.

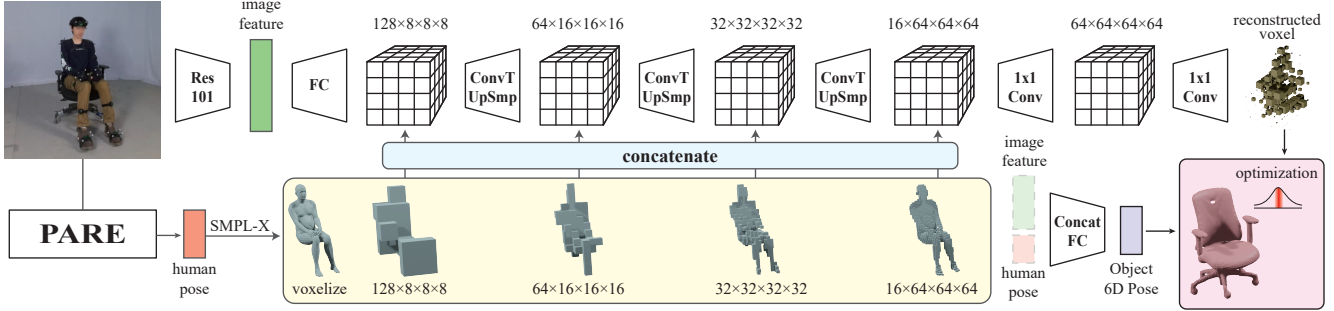


Figure 5. **The overall architecture of our model.** The reconstruction model uses the predicted voxelized human to guide the pose estimation of the interacting object. We further regress the root 6D pose of the object using the image feature and the SMPL-X parameters. We utilize both predictions and an interaction prior to optimize the final estimated pose.

with a carefully designed optimization algorithm as shown in Fig. 4. Given a parameterized human body and an articulated object point cloud, we first compute the penetration depths between the human and the object point cloud. Next, we use the transpose of the linear-blend-skinning weights of SMPL-X to aggregate the maximum penetration depth and direction to the human skeleton joints; this information is used to calculate a target skeleton that offsets the penetration. Finally, we run gradient-based optimization to fit the human model to the new skeleton while keeping the human pose parameter close to the MoCap reconstruction.

Privacy Protection We blur the faces [34] of all participants to hide identities and informed all participants that they can remove themselves from CHAIRS at any time.

4. Articulated Object Pose Estimation

CHAIRS can support a wide range of AHOI tasks, including detection, motion generation, physics-based analysis, or even language-guided motion generation with additional annotations. We showcase the value of CHAIRS on articulated object pose estimation. Despite recent progress in articulated object pose estimation [11, 14, 58] and HOI reconstruction [5, 47, 55, 67], articulated object pose estimation remains unaddressed in the challenging setting of f-AHOI. Specifically, our setting requires the model to accurately estimate the pose of the articulated objects in the context of heavy occlusion and dense contact.

4.1. Task Definition

Given an observed image I , the parameterized human model $H = (\beta, \theta_b, \theta_h, R_b, T_b)$, and the meshes $X = \{X_i, i = 1, \dots, N\}$ of the interacting object that has N parts, the task is to estimate the object pose $O = \{(R_i, T_i), i = 0, \dots, N\}$, where $\beta \in \mathbb{R}^{10}$, $\theta_b \in \mathbb{R}^{21 \times 6}$, $\theta_h \in \mathbb{R}^{30 \times 6}$, and $R_b \in \mathbb{R}^6$ and $T_b \in \mathbb{R}^3$ are the shape and pose parameters of the SMPL-X [38] model. ($R_0 \in \mathbb{R}^6$, $T_0 \in \mathbb{R}^3$) is the object root pose, and $\{(R_i \in \mathbb{R}^6, T_i \in \mathbb{R}^3)\}$ denotes the global rotation and translation for each part X_i . We use the orthogonal 6D representation [68] for the rotations in both human and object poses.

4.2. Model Architecture

We propose an interaction-aware object pose estimation model that leverages fine-grained geometric relationships in HOIs and the interaction priors. Our method contains two stages: given an image and estimated SMPL-X [38] parameters, we first estimate the object occupancy grids and root pose with a reconstruction model. Then, we optimize the reconstructed human-object pair with a learned interaction prior. Fig. 5 illustrates the overall framework of our model, and Fig. 6 shows the interaction prior model.

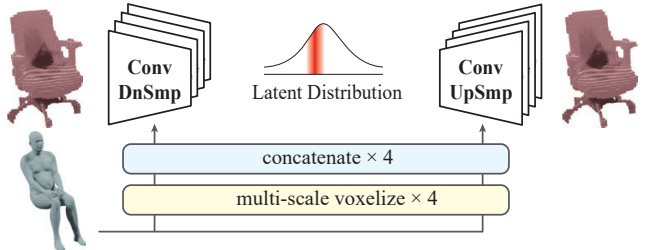


Figure 6. **An illustration of the interaction prior model.** It is a cVAE that generates object voxels conditioned on human voxels. We minimize the norm of the latent code during optimization.

4.3. Object Reconstruction and Pose Initialization

Given an observation I , we estimate the human pose and shape using an off-the-shelf estimator and voxelize the estimated human shapes H' using Kaolin [22] to four different resolutions. To better utilize the geometric relationship between the human-object pair, we estimate the object shape and pose with the guidance of the human pose. Specifically, we first extract the ResNet-101 [18] features from the image and estimate the object voxel from the image features with a 3D decoder, which is composed of three 3DConvT layers and upsampling layers at different resolutions, and two 1x1 3DConv layers. Next, we concatenate the convolutional feature grids with the human voxels at each resolution to enhance the human pose guidance. The last 3DConv layer produces the estimated object occupancy grid \mathcal{V}'_O . We finally concatenate the image features extracted from ResNet-101 and the SMPL-X parameters, and use an additional MLP to

regress the root pose (R'_0, T'_0) of the object. We also use this root pose as the initialization for the optimization.

To train the reconstruction model, we first initialize the human shape estimator with the pre-trained weights from the PARE model [38] and fine-tune it on CHAIRS. Next, we freeze the weights of the PARE model and train the reconstruction model with the object pose estimation loss $\mathcal{L}^{\mathcal{O}}$, which is the L1 loss on object voxels.

4.4. Interaction Prior

To capture the fine-grained relationship between humans and interacting objects, we propose a cVAE-based interaction prior model, which learns the conditional distribution of object occupancy given the human shape.

Specifically, the condition to the prior cVAE is a multi-resolution voxelized human, and the goal is to reconstruct the voxelized object. We use 3DConvNets as the encoder and decoder. During training, we feed the voxelized object through the encoder to get the object features at different scales. The object features are concatenated with the multi-resolution human voxels in each corresponding layer, and an MLP is utilized to estimate the latent Gaussian distribution $\mathcal{N}(\mu, \sigma)$. Next, we sample the latent code $z \sim \mathcal{N}(\mu, \sigma)$ by re-parameterization and decode it with the decoder. Finally, we concatenate the feature grids at each layer in the decoder with the corresponding human voxel condition.

We train the prior model on CHAIRS with four losses:

$$\mathcal{L}_P = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{pene}} + \mathcal{L}_{\text{contra}}, \quad (1)$$

where $\mathcal{L}_{\text{recon}}$ and \mathcal{L}_{KL} are the standard reconstruction and KL divergence loss, respectively. $\mathcal{L}_{\text{pene}}$ is the penetration loss that penalizes voxel grids occupied by both humans and objects. $\mathcal{L}_{\text{contra}}$ maximizes the distance of latent variables between the original data and augmented noisy data. We augment part of the training data with random noises.

4.5. Pose Optimization with Interaction Prior

To reconstruct the fine-grained human-object relation and recover the final object poses, we utilize an additional optimization stage based on the initialized poses using the kinematic information and the interaction prior. Specifically, given the object’s CAD model and URDF, the estimated SMPL-X parameters H' , and object voxels $\mathcal{V}'_{\mathcal{O}}$ estimated from the reconstruction model, we initialize the object model $\hat{\mathcal{O}}$ with the estimated root transformations and random part states, and iteratively update the object model $\hat{\mathcal{O}}$ ’s parameters by minimizing the objective $\mathcal{J}_{\text{recon}} + \mathcal{J}_z$:

$$\mathcal{J}_{\text{recon}} = \|V(\hat{\mathcal{O}}) - \mathcal{V}'_{\mathcal{O}}\|_2; \quad \mathcal{J}_z = \|\text{Enc}(H', \hat{\mathcal{O}})\|, \quad (2)$$

where $V(\cdot)$ is the voxelization function, $\mathcal{J}_{\text{recon}}$ term penalizes the distance between the voxelized object model and the estimated object voxels, and \mathcal{J}_z constrains the norm of

the latent predicted by the cVAE encoder to be small, which regularizes the estimated interaction to be close to the prior. The overall process of pose optimization with interaction prior is illustrated in Fig. 7.

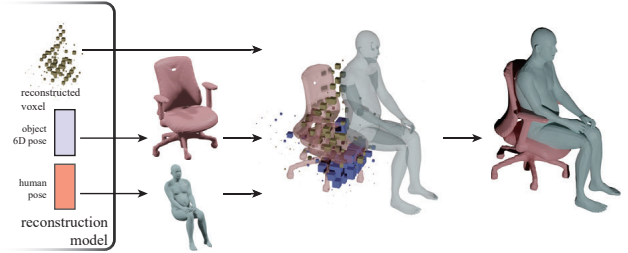


Figure 7. An illustration of pose estimation with interaction prior. Starting with the reconstruction output, we optimize the object according to the reconstructed voxel and interaction prior.

5. Experiments

Experimental Settings We split CHAIRS into training, testing, and evaluation sets; 70% of objects are used for training, 20% for testing, and the rest for evaluation. We evaluate the performance of our model under two different settings: with (*w/ opt*) and without optimization (*w/o opt*). In the *w/ opt*. setting, we report the chamfer distance between the objects posed with ground truth and estimated transformation parameters. In the *w/o opt*. setting, however, we do not have the estimated transformation parameters. We therefore report the chamfer distance between the ground-truth object mesh and the mesh obtained by running the marching cube algorithm on the reconstructed voxels.

Evaluation Metrics We evaluate object pose estimation with the mean rotation and translation errors of each object part, and evaluate the object shape reconstruction with the chamfer distance and intersection over union (IoU). We finally evaluate the reconstructed f-AHOI with the penetration depth and contact scores between the human and the object. We compute the penetration depth for a human-object pair as the maximum depth of the object’s surface inside the human’s body. This metric is zero if there is no penetration. The contact value is the shortest distance between the human and the object. We clip the contact value to [0,20cm] for human-object pairs that are far away.

Baseline Methods We compare the performance of articulated object pose estimation with two object reconstruction methods LASR [59] and ANCSH [27] as baselines; we use the depth map as the input to ANCSH. Both methods are *fine-tuned* on CHAIRS. We further compare our model with D3D-HOI [58] that jointly estimates the human and object poses. We modified the optimization objectives of D3D-HOI to better fit the data distribution of CHAIRS.

5.1. Results and Analyses

Tab. 2 shows the quantitative results. Incorporating the geometrical relationships, our model significantly improves

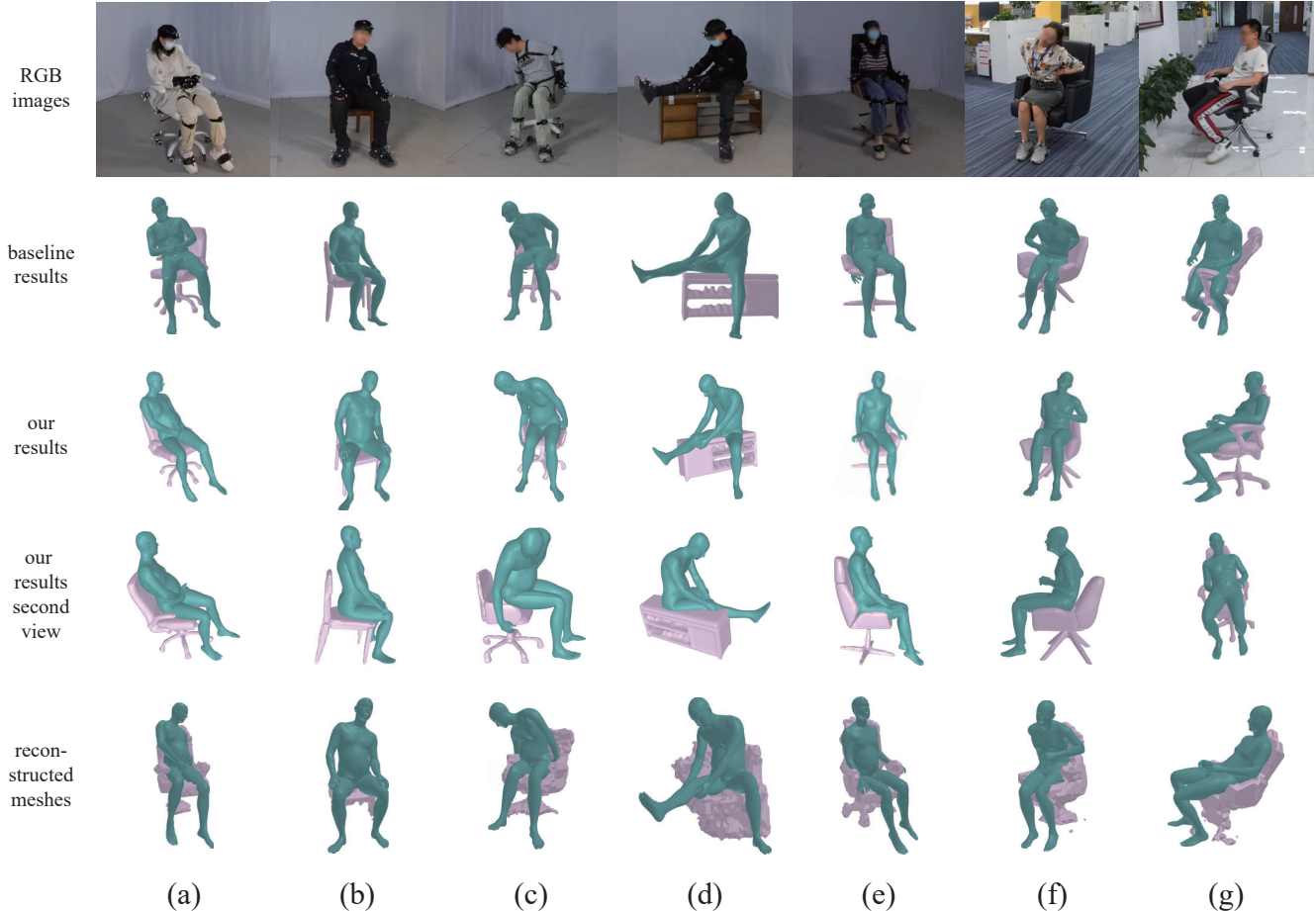


Figure 8. **Qualitative results of our model.** (a)-(e) Results on CHAIRS test set. (f)-(g) Results on images taken in the wild. Baseline results are obtained from D3D-HOI [58]. We show optimized human and object poses in the third and fourth row, and visualize the mesh obtained by running marching cube on the reconstructed voxels in the last row.

the performance of pose estimation and shape reconstruction compared with existing methods. More specifically, in the *w/o opt.* setting where the object is unknown, our model outperforms the SOTA method LASR, by a wide margin. Although our model is surpassed by D3D-HOI and ANCSH, they both assume known object structures. Our model notably outperforms all existing baselines when we provide the object structure to our model in the *w/ opt.* setting.

We show qualitative results in Fig. 8. Columns (a)-(e) show reconstruction results on the test set. We visualize the

Table 2. **Comparisons against existing methods.** *: method requires knowledge of object structure and/or geometry; †: method does not require object-related knowledge.

Method	Object				HOI	
	Rot.↓ (°)	Transl.↓ (mm)	CD↓ (mm)	IoU↑ (%)	Pene.↓ (mm)	Cont.↓ (mm)
LASR† [59]	/	/	205.2	/	/	/
ANCSH* [27]	/	/	90.36	/	/	/
D3D-HOI* [58]	27.31	119.2	126.9	16.60	7.472	1.163
Ours (w/o opt.)†	/	/	160.2	11.03	4.530	2.720
Ours (w/ opt.)*	19.35	66.23	72.30	21.57	1.143	1.562

reconstructed mesh before optimization with the marching cube. We observe that our model can reconstruct plausible and accurate interactions before optimization, and the optimization step further improves interaction details.

5.2. Ablations

We verify the design of our model with three ablation studies and report the quantitative results in Tab. 3.

Prior We remove the interaction prior model and optimize object poses by minimizing only $\mathcal{L}_{\text{recon}}$. We observe a large drop in performance in both * and † settings and con-

Table 3. **Ablation of interaction, prior, and contrastive loss.**

Method	Object				HOI	
	Rot.↓ (°)	Transl.↓ (mm)	CD↓ (mm)	IoU↑ (%)	Pene.↓ (mm)	Cont.↓ (mm)
Full†	/	/	160.2	11.03	4.530	2.720
– prior†	/	/	165.3	10.52	4.377	3.295
Full*	19.35	66.23	72.30	21.57	1.143	1.562
– prior*	19.97	83.39	87.90	18.81	1.749	2.081
– contr.*	21.52	81.90	87.28	18.93	1.265	2.393
– inter.*	17.88	69.53	78.12	19.50	1.022	2.320

firm that the interaction prior plays a vital role in estimating the object pose accurately. Note that both settings have an optimization step, and the only difference is that the * model has access to the object geometry and structure during optimization. We observe a drop in penetration when the prior model is removed in the † setting, while the contact value increases by a much larger margin. This indicates our interaction prior model pulls the object toward the human when they are not in contact.

Contrast We remove the contrastive loss $\mathcal{L}_{\text{contra}}$ when training the prior model. We observe similar results as in the –prior experiment. This result shows that contrastive loss is crucial to learning a robust interaction prior.

Interaction We remove the concatenated human voxel in 3DConv layers in both the reconstruction model and the interaction prior model. This eliminates the interaction awareness of our model. We observe slight degradation across all object reconstruction metrics, showing the significance of interaction awareness in our model. We also observe that the contact value increases while penetration drops. This is similar to the –prior ablation in *w/o opt.* setting, which shows that the interaction awareness is also pulling the human and object towards each other. Finally, we observe an unexpected low rotation error, which we attribute to the rotation symmetries in the dataset.

In summary, we conclude that all three components contribute significantly to object pose and shape reconstruction.

Failure Cases Our model fails to estimate the correct orientation of object parts in two typical scenarios. The most common scenario is rotation symmetry, wherein the object is geometrically similar under certain rotations. Rotation symmetry is common in spherical and cylindrical object parts, such as the base of a stool or a round seat. Fig. 9a shows an example of rotation symmetry. Existing methods [10, 50] bypass this issue with (i) multiple equally-correct ground truths and (ii) a min-of-N loss that calculates the smallest distance to any of the ground truths. However, this method requires a carefully designed classification of the symmetry type for each object.

We attribute another common failure to interaction symmetry; the way a person interacts with an object is identical when the object is in different poses. Interaction symmetry confuses our model when the visual module fails to differentiate poses. We show in Fig. 9b that our model leverages fine geometrical relations to reconstruct natural interactions despite the false prediction of the object pose.

In-the-wild Generalization We curate a small set of images captured in our daily scenes to test the model’s generalizability. Fig. 8(f-g) shows two qualitative results in an office and demonstrates that the proposed model generalizes to images taken outside laboratory settings. The model fails to predict an accurate object pose in the last column when the person is not interacting with the object. Please see ad-

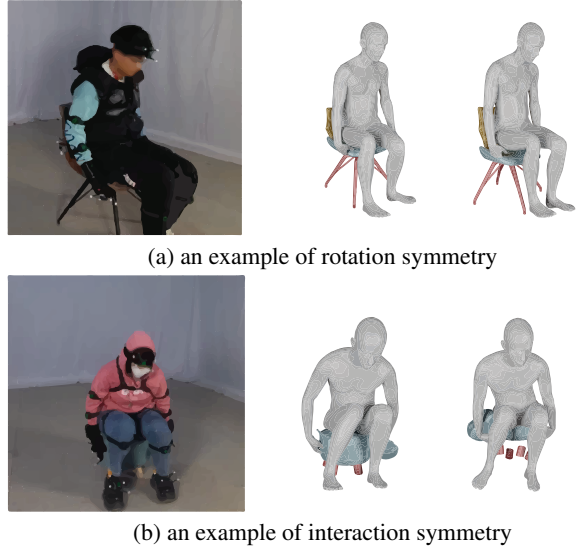


Figure 9. **Common failure cases caused by symmetry.** The left meshes are ground truths, whereas the right are the model predictions. (a) Rotation-symmetrical object yields a large rotation error but a small visual error. (b) Interaction symmetry occurs when both the body and legs of the puppy stool are flipped, yet the predicted interactions and structure look reasonable.

ditional results and analyses in the Supplementary Material.

6. Conclusion

We promote HOI towards articulated, fine-grained, and part-level direction with (i) a novel dataset, CHAIRS, (ii) a challenging problem of object reconstruction under f-AHOI, and (iii) a strong baseline. The CHAIRS dataset captures a large-scale collection of whole-body AHOIs with diverse and natural interactions and wildly different sittable objects. The object reconstruction problem removes the oversimplified assumption of kinematic consistency, and our model leverages fine-grained interaction relationships to rule out ambiguities.

Limitations While our model can accurately reconstruct articulated objects under heavy occlusions from f-AHOI, its performance depends heavily on the interaction that created such ambiguity. The performance of our model drops significantly when there is no f-AHOI. In addition, Our model does not leverage the interaction prior to improve human pose estimation. Similar to our approach, the same interaction prior will likely improve human pose estimation in hard cases when the human is heavily occluded.

Societal Impacts CHAIRS and f-AHOI bring in new opportunities to understand how humans interact with the environment. We firmly believe that a solid understanding of f-AHOIs in the future would empower intelligent agents in real-life applications, such as assistive robots in healthcare and elderly care services, as well as indoor service robots that clean and arrange furniture. Meanwhile, we are aware of the insecure use of f-AHOIs understanding in

surveillance technology that could lead to the invasion of privacy; we blur all faces to remove personally identifiable information in our dataset.

Acknowledgments We thank Zhiyuan Zhang for his technical support during his internship at BIGAI. This work is supported in part by the National Key R&D Program of China (2021ZD0150200) and the Beijing Nova Program.

References

- [1] Ben Abbatematteo, Stefanie Tellex, and George Konidaris. Learning to generalize kinematic models to novel objects. In *Conference on Robot Learning (CoRL)*, 2019.
- [2] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [3] Markus Braun, Qing Rao, Yikang Wang, and Fabian Flohr. Pose-rcnn: Joint object detection and pose estimation using 3d object proposals. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2016.
- [4] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiakuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *International Conference on Computer Vision (ICCV)*, 2015.
- [5] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [6] Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and Vision Computing*, 10(3):145–155, 1992.
- [7] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, 2020.
- [8] Karthik Desingh, Shiyang Lu, Anthony Pipari, and Odest Chadwicke Jenkins. Factored pose estimation of articulated objects using efficient nonparametric belief propagation. In *International Conference on Robotics and Automation (ICRA)*, 2019.
- [9] Thanh-Toan Do, Ming Cai, Trung Pham, and Ian Reid. Deep-6dpose: Recovering 6d object pose from a single rgb image. *arXiv preprint arXiv:1802.10367*, 2018.
- [10] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [11] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Articulated objects in free-form hand interaction. *arXiv preprint arXiv:2204.13662*, 2022.
- [12] Haw-ren Fang and Yousef Saad. Two classes of multisecond methods for nonlinear acceleration. *Numerical Linear Algebra with Applications*, 16(3):197–221, 2009.
- [13] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [14] Sanjay Haresh, Xiaohao Sun, Hanxiao Jiang, Angel X Chang, and Manolis Savva. Articulated 3d human-object interactions from rgb videos: An empirical analysis of approaches and challenges. In *International Conference on 3D Vision (3DV)*, 2022.
- [15] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *International Conference on Computer Vision (ICCV)*, 2021.
- [16] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *International Conference on Computer Vision (ICCV)*, 2019.
- [17] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [19] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [20] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, and Vincent Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *International Conference on Computer Vision (ICCV)*, 2011.
- [21] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017.
- [22] Krishna Murthy Jatavallabhula, Edward Smith, Jean-Francois Lafleche, Clement Fuji Tsang, Artem Rozantsev, Wenzheng Chen, Tommy Xiang, Rev Lebaredian, and Sanja Fidler. Kaolin: A pytorch library for accelerating 3d deep learning research. *arXiv preprint arXiv:1911.05063*, 2019.
- [23] Zhenyu Jiang, Cheng-Chun Hsu, and Yuke Zhu. Ditto: Building digital twins of articulated objects from interaction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [24] Jia Kang, Wenjun Liu, Wenzhe Tu, and Lu Yang. Yolo-6d+: single shot 6d pose estimation using privileged silhouette information. In *IEEE International Conference on Image Processing (ICIP)*, 2020.
- [25] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *International Conference on Computer Vision (ICCV)*, 2017.
- [26] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

- [27] Xiaolong Li, He Wang, Li Yi, Leonidas J Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [28] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [29] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [30] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [31] Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Yang Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [32] Liu Liu, Han Xue, Wenqiang Xu, Haoyuan Fu, and Cewu Lu. Toward real-world category-level articulation pose estimation. *Transactions on Image Processing (TIP)*, 31:1072–1083, 2022.
- [33] Frank Michel, Alexander Krull, Eric Brachmann, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Pose estimation of kinematic chain instances via object coordinate regression. In *British Machine Vision Conference (BMVC)*, 2015.
- [34] Asmaa Mirkhan. Blurry faces. <https://github.com/asmaamirkhan/BlurryFaces>, 2020.
- [35] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [36] Jiteng Mu, Weichao Qiu, Adam Kortylewski, Alan Yuille, Nuno Vasconcelos, and Xiaolong Wang. A-sdf: Learning disentangled signed distance functions for articulated shape representation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [37] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [38] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [39] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [40] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *European Conference on Computer Vision (ECCV)*, 2018.
- [41] Shengyi Qian, Linyi Jin, Chris Rockwell, Siyi Chen, and David F Fouhey. Understanding 3d object articulation in internet videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [42] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *International Conference on Computer Vision (ICCV)*, 2021.
- [43] Antoni Rosinol, Arjun Gupta, Marcus Abate, Jingnan Shi, and Luca Carlone. 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans. *arXiv preprint arXiv:2002.06289*, 2020.
- [44] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Pigraphs: learning interaction snapshots from observations. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016.
- [45] A. Segal, Dirk Hhnel, and S. Thrun. Generalized-icp. In *Robotics: Science and Systems (RSS)*, 2009.
- [46] Chenhua Shen. Analysis of detrended time-lagged cross-correlation between two nonstationary time series. *Physics Letters A*, 2015.
- [47] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020.
- [48] Wei-Cheng Tseng, Hung-Ju Liao, Lin Yen-Chen, and Min Sun. Cla-nerf: Category-level articulated neural radiance field. In *International Conference on Robotics and Automation (ICRA)*, 2022.
- [49] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [50] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [51] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [52] Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qinpeng Zhao, and Kai Xu. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [53] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [54] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

- [55] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision (ECCV)*, 2022.
- [56] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [57] Jingwei Xu, Huazhe Xu, Bingbing Ni, Xiaokang Yang, X. Wang, and Trevor Darrell. Hierarchical style-based networks for motion synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.
- [58] Xiang Xu, Hanbyul Joo, Greg Mori, and Manolis Savva. D3d-hoi: Dynamic 3d human-object interactions from videos. *arXiv preprint arXiv:2108.08420*, 2021.
- [59] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [60] Jiaolong Yang, Hongdong Li, Dylan Campbell, and Yunde Jia. Go-icp: A globally optimal solution to 3d icp point-set registration. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015.
- [61] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What’s in your hands? 3d reconstruction of generic objects in hands. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [62] Hangjie Yuan, Mang Wang, Dong Ni, and Liangpeng Xu. Detecting human-object interactions with object-guided cross-modal calibrated semantics. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022.
- [63] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes: The importance of multiple scene constraints. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [64] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [65] Jason Y. Zhang, Sam PePOSE, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020.
- [66] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *International Conference on Computer Vision (ICCV)*, 2021.
- [67] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European Conference on Computer Vision (ECCV)*, 2022.
- [68] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

A. Model

Why Articulated Object Pose Estimation? Existing studies of HOI usually estimate the pose of human and object 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 [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. A1 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.

Table A1. Comparisons against optimizing the articulated object poses with different human pose priors. GT denotes using ground truth human poses to optimize the object poses, No inter. denotes not considering human-object interaction prior.

Method	Human		Object	
	MPJPE↓ (mm)	PA-MPJPE↓ (mm)	CD.↓ (mm)	IOU.↑ (%)
No inter.	/	/	87.90	18.81
PARE [26]	81.09	47.19	73.79	21.66
PARE(finetune)	74.50	43.99	72.30	21.57
GT	0	0	65.50	23.16

Contrastive Loss We expect our interaction prior to capturing a more general human-object interaction relationship by a conditional Gaussian distribution, where the latent codes of reasonable and common human-object spatial relationships should locate closer to the mean of the computed Gaussian than those of the unreasonable ones. To encourage this desired behavior, we devise a contrastive loss to train the interaction prior model alongside the usual penetration, reconstruction, and KL-divergence losses. Given an observed human H and voxel \mathcal{V}_O of the object O from the training data as a positive example (H, \mathcal{V}_O) , we first generate the corresponding negative example (H, \mathcal{V}'_O) by random perturbation to the object.

More specifically, we add random noise to the root and articulated poses of O and obtain \mathcal{V}'_O by voxelization. We then define the contrastive loss as $\mathcal{L}_{\text{contra}} = \max(0, \|\text{Enc}(\mathcal{V}_O, H)\| - \|\text{Enc}(\mathcal{V}'_O, H)\|)$, where Enc represents the conditional encoder of our proposed cVAE-

based prior model. $\mathcal{L}_{\text{contra}}$ pushes the latent codes of perturbed human-object pairs away from the distribution centroid.

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.

Optimization Details Fig. A1 illustrates the object pose optimization process. Given the estimated object voxel, the estimated human, the object shape, and the kinematic structure of the object, the goal of the optimization is to fit the object to the human body under the following conditions:

- (1) The optimized object should match the object voxel estimated from the monocular image.
- (2) The spatial relationship between the object and the human agrees to the interaction prior.

The parameters to be optimized are the root 6D pose of the object, denoted as R, T , and its joint parameters (if any), denoted as Φ , that control the rotation and shift of the parts under kinematic constraints. In this work, we assume that a joint (except the root) can be revolute (rotate along one axis), prismatic (shift along one axis), or both. The revolute-prismatic joint (such as the joint that links the base and seat of an office chair) is constrained to rotate and shift along the same axis.

For the optimization, we first initialize the root pose R, T using the estimated root 6D pose from the object pose estimation model and all the joint parameters Φ to zero. Then the parameters R, T and Φ are optimized by minimizing the reconstruction loss $\mathcal{J}_{\text{recon}}$ and interaction prior loss \mathcal{J}_z through gradient descent.

Note that the discrete $(0, 1)$ voxel occupancy is not differentiable w.r.t. the object parameters R, T , and Φ through deformations or transformations. Thus, we resample the voxel occupancy using trilinear interpolation given the affined $(0, 1)$ voxel grid, allowing the gradients to flow and update the root and joint parameters. After optimization, the parameters can be directly applied to get an updated 3D object model and further a finer object representation (e.g., mesh) with less geometric error.

Model Details Please check the code in [https://github.com/jnnan/chairs/blob/main/optimize_cvae_part.py].

Adapting D3D-HOI as Baseline 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:

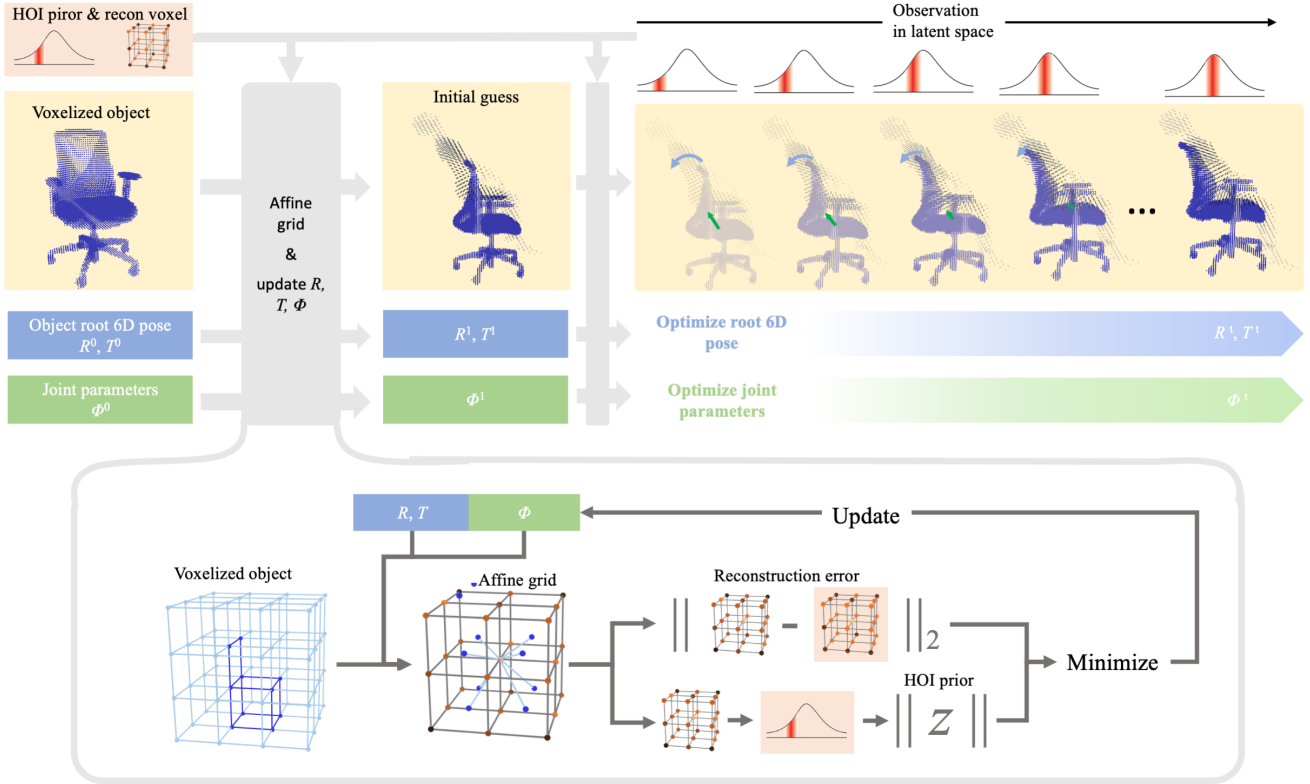


Figure A1. Detailed diagram of the optimization process.

Table A2. 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

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.
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 add a 20cm offset along the negative Y direction when computing this error.
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.

B. Additional Results

Qualitative Results In Fig. A2, we qualitatively show more randomly selected results on the test set of CHAIRS. In general, our model predicts accurate object poses and shapes.

In the Wild In Fig. A3, we qualitatively evaluate the generalization power of our model with three videos captured in the wild. We show four frames from each video. Our model generalizes well in the first row when the background is relatively clean. Our model fails to predict accurate object poses in some samples in the second and third rows. We hypothesize these are caused by the noisy backgrounds or the camera views that are out of the training data distribution.

Experimental Results on 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 yogaball, 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. A2 and qualitative results in Fig. A4. We observe that although the metrics drop numerically, our model is still able to reconstruct the poses of the interacting objects.



Figure A2. Additional qualitative results of our model on the test set of CHAIRS.



Figure A3. Qualitative results of running our model on images captured in the wild.

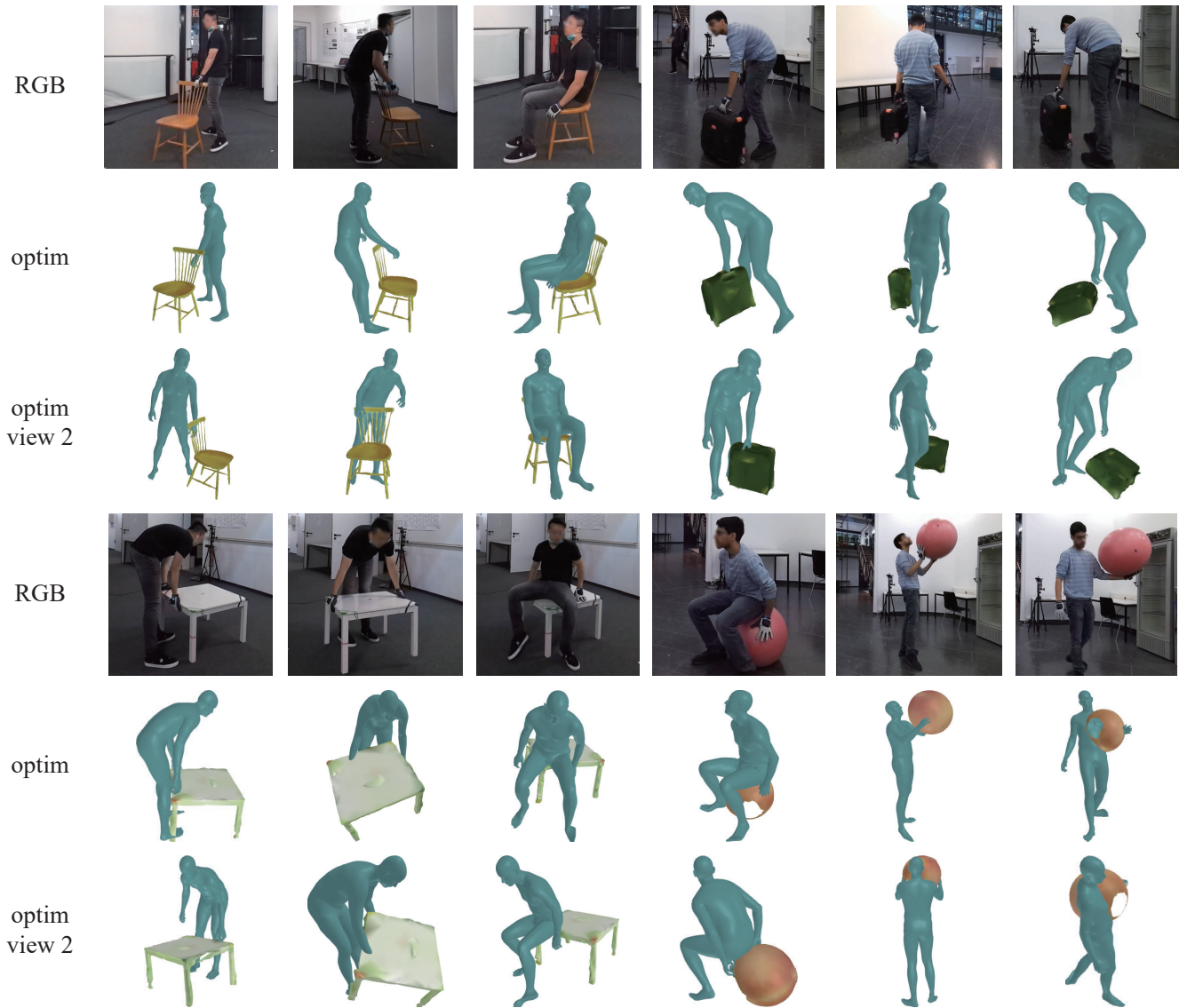


Figure A4. Qualitative results of running our model on images from the BEHAVE [2] dataset.

C. Dataset

C.1. Data Collection

Object Gallery We render all objects in CHAIRS in Fig. A5. 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.



Figure A5. **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.

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 A6 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 the head, two hands, and two feet of each participant. We then optimize the body shape pa-

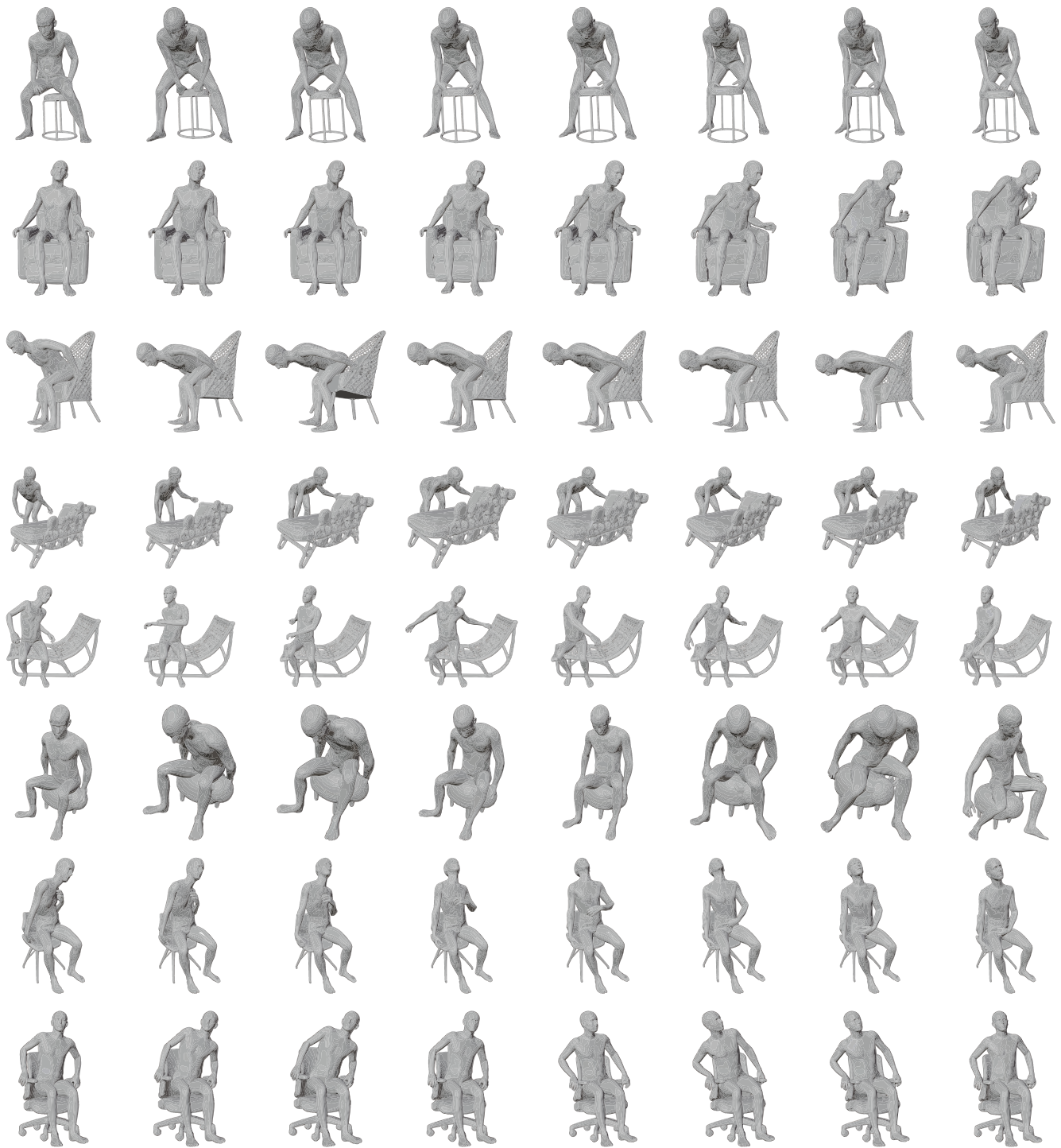


Figure A6. Performances of different participants on different objects with the same instruction. The first four rows show four performances of the instruction “Move the chair”. The second participant rotated the chair with a small angle. The last four rows show four performances of the instruction “Stretch a little in the chair.”

parameter β of the SMPLX model to fit the tracker positions. We rely on SMPLX’s default hand shape parameter since it is not our primary focus 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-viewpoint clouds and the registration of Kinect and Mocap. We base our method on plane-to-plane correspondences [45] to alleviate the sensitivity to outliers, disturbances, and partial overlaps. Given the source point set $P = \{p_i, i = 1, \dots, N\}$ captured by the Kinect depth cameras and the target set $Q = \{q_i, i = 1, \dots, M\}$ 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 \tilde{q}_i in Q to each p_i in P . Next, we iteratively update T to minimize the Mahalanobis distance between P and Q :

$$T = \arg \min_T \sum_{i=1}^M d_i^T (C_{n,\tilde{q}_i}^Q + TC_{n,i}^P T^T)^{-1} d_i, \quad (\text{A1})$$

where d_i is the corresponding Euclidean distance between p_i and \tilde{q}_i , C_{n,\tilde{q}_i}^Q and $C_{n,i}^P$ the covariance matrix calculate by the n nearest points around \tilde{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 sys-

tems run steadily at 30 Hz, the two recorded data streams have a constant difference in time. We use a time-lagged cross-correlation (TLCC) [46] 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. We obtain the human joint positions with the Kinect Body Tracker SDK for the Kinect cameras. 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 GitHub 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 [26] model and code [license/pare.txt]
- Category-Level Articulated Object Pose Estimation [27] model and code [No license information found.]
- Metropoly rigged 3D people (used in main paper Fig.3 and supplementary video) [license/animation-model.txt]
- D3D-HOI [58] code [No license information found.]
- iStock [<https://www.istockphoto.com>] images used for in-the-wild evaluations. [license/istock.txt]