arXiv:2401.10232v2 [cs.CV] 22 Jan 2025

ParaHome: Parameterizing Everyday Home Activities Towards 3D Generative Modeling of Human-Object Interactions

Jeonghwan Kim* Jisoo Kim* Jeonghyeon Na Hanbyul Joo
Seoul National University
{roastedpen,jlogkim,prom317,hbjoo}@snu.ac.kr
https://jlogkim.github.io/parahome

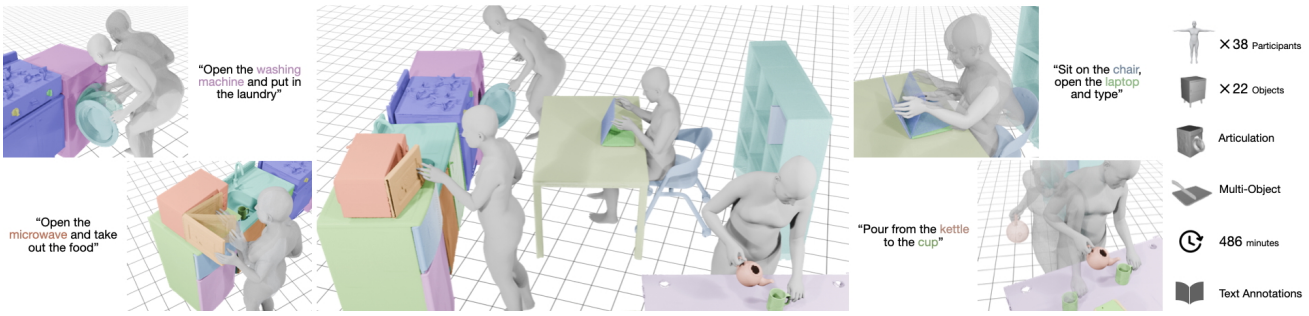

Figure 1. Our system captures the detailed 3D movements of the human body, hands, and diverse objects, along with text descriptions.

Abstract

To enable machines to understand the way humans interact with the physical world in daily life, 3D interaction signals should be captured in natural settings, allowing people to engage with multiple objects in a range of sequential and casual manipulations. To achieve this goal, we introduce our ParaHome system designed to capture dynamic 3D movements of humans and objects within a common home environment. Our system features a multi-view setup with 70 synchronized RGB cameras, along with wearable motion capture devices including an IMU-based body suit and hand motion capture gloves. By leveraging the ParaHome system, we collect a new human-object interaction dataset, including 486 minutes of sequences across 207 captures with 38 participants, offering advancements with three key aspects: (1) capturing body motion and dexterous hand manipulation motion alongside multiple objects within a contextual home environment; (2) encompassing sequential and concurrent manipulations paired with text descriptions; and (3) including articulated objects with multiple parts represented by 3D parameterized models. We present detailed design justifications for our system, and perform key generative modeling experiments to demonstrate the potential of our dataset.

1. Introduction

Our daily routines involve interactions with various objects, where we perform complex, sequential, and dexterous motions. Machines, however, struggle to mimic these behaviors, as the connection between human action and environmental responses (e.g., pulling a refrigerator handle to open it) is challenging to model. The major obstacle is the lack of large-scale datasets capturing 3D human-object interactions (HOIs) in natural and casual settings, including the 3D motions of human bodies, hands, and objects within a common spatio-temporal space. Existing datasets cover limited aspects: capturing human motion without objects [32, 47, 52, 70, 89], capturing human body motions in a static environment without dexterous hand manipulations [39, 88, 104], focusing on hand and object interactions in static postures without motions [4, 11, 36, 43], or considering relatively simple and atomic interactions such as grasping a single object [2, 26, 61, 92, 103, 117]. While a few recent datasets try to capture both body motion and dexterous hand manipulation, they are often captured in a less natural setup (e.g., a simple table setup), containing limited

*Equal contribution

action diversity [23, 49, 114].

To solve such fundamental challenges in collecting high-quality HOI data in a natural home environment, we present a novel capture system *ParaHome*, along with a large-scale dataset including diverse and natural human-object interactions. Our system allows participants to freely interact with multiple objects in a room environment as shown in Fig. 1. Obtaining such signals is challenging, since the system should handle severe occlusions during interactions and multiple scales (big body movements across the room and subtle hand motion). As a solution, we design our ParaHome system by combining a multiview camera system equipped with 70 synchronized cameras, wearable IMU-based motion capture suit and gloves. The multi-camera system tracks the 3D movements of objects, capturing both rigid and articulated motions, as well as the global rigid transformations of the human body and hands in the camera space. The human body and hand motions are captured using wearable motion capture equipments that are occlusion-free enabling occlusion-robust capture of human object interaction. As combining heterogeneous systems is challenging, we present multiple hardware and algorithmic solutions to robustly reconstruct human motions and object movements in a common 3D spatiotemporal space. Through experiments, we verify the robustness of our system, providing justifications of our design choices.

Leveraging our ParaHome system, we collect a large-scale human-object interaction dataset in the home activity scenarios. Our dataset contains data from 38 participants, 22 objects, 208 captures, and 486 minutes of sequences in clock time, which are publicly available. We design the capture scenario to reflect various natural activities that are commonly observed in our daily lives, where each participant performs sequential and casual interactions with multiple objects concurrently, in their own unique styles given a common verbal guidance in each capture. The resulting data captures detailed spatio-temporal relations between humans and multiple objects within 3D parametric spaces, where we provide corresponding 3D human mesh model (SMPL-X) [78], object-specific parametric models with their contact cues, and descriptive text annotaions, as shown in Fig. 1.

Our newly captured dataset provides new research opportunities for modeling the correlations between human motions and 3D object movements in natural HOI scenarios. To demonstrate this, we formulate our generative modeling pipeline in the context of motion synthesis, and explore two possible example problems: text-conditioned motion synthesis and object-guided motion synthesis.

2. Related Work

3D Motion Capture System and HOI Datasets. Pioneered by the work of Kanade et al. [53], several systems propose to reconstruct human behaviors [31, 72, 73, 81] with multiview capture system. With such system, early approaches pursue to capture human body motions markerlessly [5, 6, 12, 15, 18, 24, 25, 28, 56, 84, 91, 97, 98]. For more robust human motion tracking, some works combined IMU sensors into their system [17, 35, 54, 60, 63, 77].

Such multi-view systems are used with additional cues to capture hand-object interactions to deal with issues such as severe occlusions or smaller scale of hands. For example, some methods use RGB-D images with the optimization techniques

[36, 37], manual annotation [11, 129], or pretrained models [87]. Some other methods use marker-based mocap systems [38], magnetic sensors [26], or mocap gloves [19], and some rely on synthetic data to model the interactions [43, 48]. Other than hand-object interactions, body-object interactions are captured with manual annotations [2], IMUs combined with RGB data to track objects[119], IMUs for body pose and RGB for tracking body with object [85, 123]. Considering both body and hands, scale difference is a critical issue. As a resolution, methods using depth estimation and segmentation [103], mocap systems [61, 92], multi-view cameras [117, 118], and RGB-D setups [46] are used. To account for articulated objects rather than rigid objects, egocentric RGB data is used with manual annotation to compute hand and object pose [67]. Some approaches use multi-view setup with markers attached to the surface of body and object to capture interaction with sittable object [49] and others [23, 68, 113, 114] use table setting to restrict body movements or set a placeholder in the synthetic environment [50]. We summarize the key aspects of existing datasets and ours in supp. mat.

Human Motion Synthesis. Various approaches are proposed to reconstruct static hand-object interactions using dataset-based regression [21], spatio-temporal consistency [65], and utilize prior knowledge of hand-object contact [9, 30] or interaction fields [110, 126]. Such methods have been extended to include body and rigid object poses via visibility cues [102, 105, 106]. On top of that, recent approaches use an additional condition to generate human body motion, either text annotations [34, 51, 83, 95, 121] or scene contexts [8, 40, 66, 74, 99–101, 104, 112]. Some methods rather focus on object manipulation [20, 41, 57, 61, 62, 103, 107] and sometimes conditioned on text descriptions [80, 108, 109] without getting hand pose. Several methods focus on getting plausible hands by exploring to integrate hand pose into rigid object manipulation by generating grasp poses or motions given a target object trajectory [13, 16, 96, 127], dynamic manipulation with hands and object trajectories [115], and text-based hand manipulation with rigid objects [14]. Considering both body and hand, some approaches address body motion generation for grasping rigid objects [69, 93, 94, 125]. Moving beyond simple grasping, some methods generate full-body motion

and object interactions conditioned with action label [29], while others use correlations to generate object positions or scenes [76, 82, 111]. More recently, methods have synthesized hand motions conditioned on object states, including articulated motions [116, 120, 124], or generate hand-object interactions based on text input [10].

3. 3D Parametric Home Capture System

3.1. HOI Data Parameterization

To effectively model human-object interactions, we consider a parameterized 3D space that captures the nuanced relationship between human motions and object movements. For the human parts, we focus on the motion of the whole body including dexterous hand movements as the essential components. Rigid objects are represented through 6 DoF rigid motions, and we also incorporate object-specific dynamics such as opening the laptop or turning the knobs of a gas stove. Formally, we represent the status of a human and the environment at time t as $\mathbf{S}(t) = \{\mathbf{S}_p(t), \mathbf{S}_e(t), \mathcal{T}\}$, where $\mathbf{S}_p(t)$ is the current status of a human subject at time t, the environment status $\mathbf{S}_e(t)$ represents the current status of surrounding objects, and $\mathcal{T}$ denote a corresponding language description. The human status $\mathbf{S}_p(t) = \{\mathbf{S}_b(t), \mathbf{S}_{lh}(t), \mathbf{S}_{rh}(t)\}$ is composed of the body $\mathbf{S}_b(t)$, left-hand $\mathbf{S}_{lh}(t)$, and right-hand parameters $\mathbf{S}_{rh}(t)$, where each of them can be represented as mocap outputs via a global location and local joint orientations. The environment status is represented as $\mathbf{S}_e(t) = \{\mathbf{S}_e^j(t)\}_{j=1}^N$, where $\mathbf{S}_e^j(t)$ is the status of the j-th object, assuming we consider N different object instances. The object status $\mathbf{S}_e^j(t) = \{\mathbf{l}_e^j(t), \boldsymbol{\theta}_e^j(t), \boldsymbol{\phi}_e^j(t)\}$ is represented by 3D translation $\mathbf{l}_e^j(t) \in \mathbb{R}^3$, 3D orientation $\boldsymbol{\theta}_e^j(t) \in SO(3)$, and object specific parameters $\boldsymbol{\phi}_e^j(t)$. The length of $\boldsymbol{\phi}_e^j(t)$ varies based on the type and the dimension of movable parts of the objects (e.g. $\boldsymbol{\phi}_e^{\text{laptop}}(t)$ and $\boldsymbol{\phi}_e^{\text{drawer}}(t)$ contain one and two part parameters each).

3.2. Hardware System and Architecture

Our ParaHome system encompasses an area of $12.4m^2$, as shown in Fig. 2. To capture full body motion, subtle hand motions, and 3D object movements across the room setting, our system integrates a multi-camera system and an IMU-based wearable motion capture suit and gloves, leveraging their complementary strengths. To cover the entire volume of the room and reduce occlusion issues, we install 70 RGB industrial cameras. We use Xsens motion suit [75] and Manus hand gloves [71] for wearable motion capture solutions. For further system details, see supp. mat..

3.3. Modeling and Tracking 3D Objects

3D ArUco Marker Design and Placement. To reliably track objects and their articulated motions, we attach ArUco markers [27] to all faces of the 3D cube spanning $6cm$ inspired by [130] as shown in Fig. 2. The major motivation of our 3D marker solution lies in its robust tracking advantage during complicated manipulation scenarios, where often objects are severely occluded by hand grasping. In particular, we find attaching markers on object surface as in previous approaches [23] is sub-optimal, often requiring laborious manual post-processing. We quantitatively demonstrate the strength and robustness of our 3D marker solution in Sec. 4.

Computing Transforms between Object and 3D Markers. We obtain high-quality 3D mesh of all objects placed in our system via off-the-shelf scanner and manual alignment. Our scanned objects are shown in Fig 3. For better visibility, we attached one or multiple 3D ArUco markers and flat-style markers to each object part. Then, objects are tracked by detecting the ArUco markers in every frame. The position and orientation of each object are obtained as follows:

$$\mathbf{T}_{obj}(t) = \mathbf{T}_{mar \to obj} \mathbf{T}_{mar}(t) \tag{1}$$

where $\mathbf{T}_{mar \to obj}$ is a pre-computed fixed transformation from the marker to the object, and $\mathbf{T}_{mar}(t)$ is a transformation for the markers from the object canonical space to the current pose in the camera system space at capture time t. $\mathbf{T}_{mar \to obj}$ is computed by selecting several corresponding points between the object scan and the one in camera space.

Modeling Object Articulations. Since the whole parts of an object are scanned as a single chunk, we manually separate each rigidly moving part and build a parameterized structure for articulated models. We assume either a revolute or sliding joint. Both types of joints require articulation axis $\mathbf{a}_e \in \mathbb{R}^3$ and in case it is revolute, pivot point $\mathbf{p}_e \in \mathbb{R}^3$ additionally. Thus, we define the joint status of j-th object as $\boldsymbol{\phi}_e^j(t) = \{\boldsymbol{\tau}_i^j\}_{i=1}^n$ where $\boldsymbol{\tau}_i^j = \{\mathbf{a}_{e,i}^j, \mathbf{p}_{e,i}^j, s_{e,i}^j(t)\}$. Here, n represents the number of parts, and $s_{e,i}^j(t) \in \mathbb{R}$ denotes a relative part state (either radian for revolute joint or meter for sliding joint) from part status in the object canonical space. Examples are shown in Fig. 3. See supp. mat. regarding the process of computing articulation axis $\mathbf{a}_e^j$ and pivot point $\mathbf{p}_e^j$.

Capturing 3D Object Motions. We track the status of each object at time t, $\mathbf{S}_e(t) = \{\mathbf{l}_e(t), \boldsymbol{\theta}_e(t), \boldsymbol{\phi}_e(t)\}$, via our multi-view system. The 3D rigid transformation (translation $\mathbf{l}_e(t)$ and 3D orientation $\boldsymbol{\theta}_e(t)$) is computed by specifying the attached marker corners at each time t, triangulating N number of 2D marker points $\{\mathbf{m}_i(t)\}_{i=0}^N$ into 3D marker points $\{\mathbf{M}_i(t)\}_{i=0}^N$. Given the 3D marker corners $\{\mathbf{M}_i(t)\}_{i=0}^N$ and the corresponding corners $\{\hat{\mathbf{M}}_i\}_{i=0}^N$ in the object canonical space, we compute $\mathbf{T}_{mar}(t)$, via the Kabsch algorithm. Then, the object transformation $\mathbf{T}_{obj}(t)$ can be computed as Eq. 1. The object-specific dynamic status $\boldsymbol{\phi}_e(t)$ is similarly acquired via transformations from the 3D markers attached to the base part to those on the movable parts.

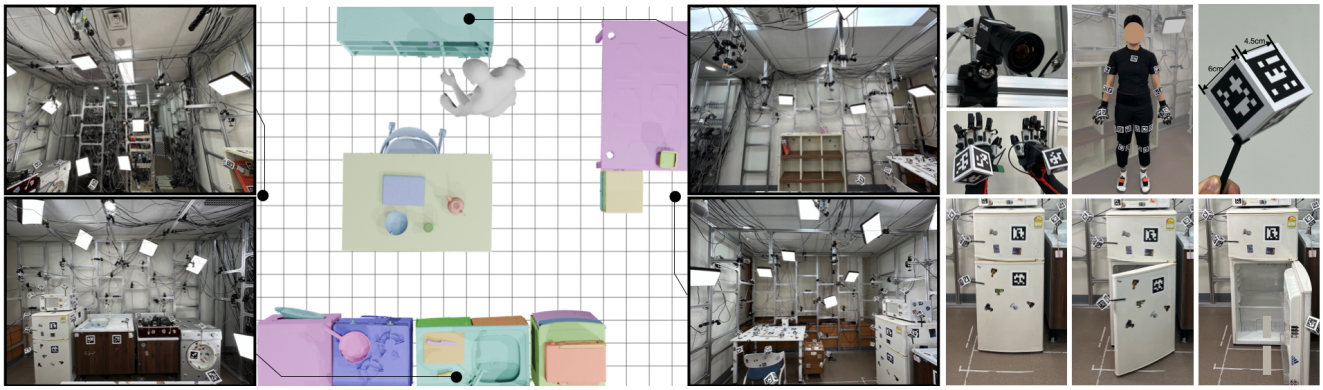

Figure 2. (Center) Reconstructed scene of ParaHome from top view. Pictures adjacent to the rendering were taken from the center of the room, headed towards the corresponding black dots in the scene. (Right) Pictures of RGB camera, IMU based motion capture devices with attached body markers and the 3D marker solution on an articulated object.

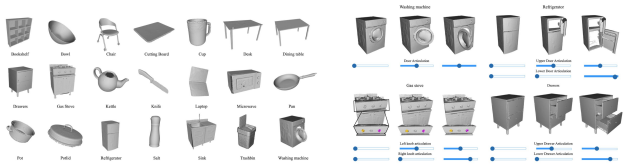

Figure 3. (Left) Scanned 3D models in ParaHome system. (Right) Articulation state of 3D models. Blue bars show the object-specific parameters $s^j_{e,i}(t)$ for each object part i. As $s^j_{e,i}(t)$ changes, corresponding parts of the objects show different articulation states

3.4. Capturing 3D Human Motions

It is challenging to leverage two heterogeneous systems, a multi-camera system and wearable motion suits and gloves, since they do not share a common spatial world coordinate. Specifically, wearable motion capture systems suffer from the drift issue in localizing the global root position, and more critically they employ an imperfect assumption of the body and hand skeleton scale that differs from the actual measurement. As such, directly transferring the output of wearable captures into the camera system space cannot fulfill our goal of capturing precise hand-object interactions. In this section, we present a method to spatially align two systems.

Aligning Wearable Mocap in Multi-Camera System. To spatially align the body motion capture with our multi-view camera system, we need to provide correspondences between the two systems. For this purpose, we attach 3 or 4 ArUco markers to each of 11 near-rigid body parts during body alignment capture(torso, hands, upper arms, lower arms, upper legs, lower legs), as shown in Fig. 2. We denote the four corner points of the j-th marker attached to the i-th body part as $\mathcal{M}^b_{i,j} \in \mathbb{R}^{4\times 3}$, where the positions are defined in the local joint coordinate w.r.t the corresponding body part. Specifically, the goal of our alignment process is: (1) to obtain authentic body skeleton configuration $\mathcal{B} = \{\mathcal{O}\}$, which represents the offsets of the child joints from the parent joints, and (2) body marker locations $\mathcal{M}^b$. Given these parameters and the joint angle measurements provided by the mocap suit $\theta^t \in \mathbb{R}^{23\times 3}$ at time t, we can transform the body-attached markers into the person-centric coordinate denoted as $\mathbf{M}^b_{mocap}(t)$ via forward kinematics function $\mathcal{K}_b$ as: $\mathbf{M}^b_{mocap}(t) = \mathcal{K}_b(\mathcal{M}^b, \theta^t, \mathcal{B})$. At the same time, the body-attached markers can be reconstructed via our multi-view system, denoted as $\mathbf{M}^b_{cam}(t)$ defined in the camera system space. Then, the rigid transformation to transfer the mocap data into the camera system space can be computed with the marker correspondence: $T^{cam}_b(t) = T(\mathbf{M}^b_{mocap}(t), \mathbf{M}^b_{cam}(t))$, where only visible markers in $\mathbf{M}^b_{cam}(t)$ are considered for the computation.

Note that we can compute the rigid transformation as long as at least one body-attached marker is visible in $\mathbf{M}^b_{cam}(t)$, providing robustness to the marker occlusions in computing the global location of the actor. We perform such transformation for the entire frames of body alignment capture which consists of motions that encompass various body poses. Our objective function for $\mathcal{B}$ and $\mathcal{M}^b$ is defined as:

$$\min_{\mathcal{M}^b, \mathcal{B}} \sum_{t=1}^{T} \lambda_b \mathcal{L}^t_{body} + \lambda_f \mathcal{L}^t_{foot}. \tag{2}$$

$\mathcal{L}_{body}$ is the mean-squared error between corresponding markers. See supp. mat. for further details. We use a constraint term $\mathcal{L}_{foot}$ to enforce the foot parts to be close to the ground without penetration or floating. The result of our body alignment process is shown in Fig. 4, showing the aligned markers after the optimization.

Calibrating Hand Mocap. The hand mocap outputs from the gloves also suffer from similar issues which are unknown skeleton lengths of the actual hand and localizing into the camera system space. Since a subtle error may cause a large deviation in the hand interaction scene, such issues are critical in obtaining high-quality data. As a solution, we present a new hand alignment protocol using a calibration structure. We specify the ordered 3D corner vertices of the structure,

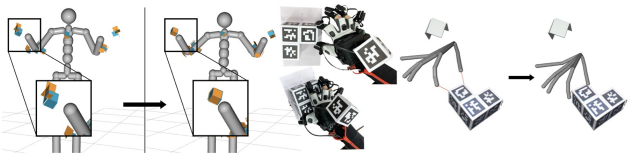

Figure 4. (Left) Before/After Body Calibration, Orange: forward kinematic output, Blue: RGB Triangulated Result (Right) Hand Calibration Protocol and Before/After Calibration Protocol

$C = \{\mathbf{c}_i \in \mathbb{R}^3\}_{i=1}^6$, which we localize in camera system space via triangulation. During the alignment protocol, the participant touches the calibration structure's known corner locations with their fingertips, from which we can approximate the desired locations of fingertips. Examples of hand calibration protocols and the hand calibration structure are shown in Fig. 4 and our supplementary video as well.

Similar to the body calibration, the goal of hand calibration is to obtain the authentic hand skeleton configuration $\mathcal{H} = \{\mathcal{S}^h \in \mathbb{R}^{20\times 1}, \mathcal{O}^h \in \mathbb{R}^{20\times 3}\}$ and positions of 3D markers in the local hand-centric coordinate $\mathcal{M}^h \in \mathbb{R}^{3\times 4\times 3}$, where $\mathcal{S}^h$ and $\mathcal{O}^h$ each denotes scales of hand skeleton and per-joint skeletal offsets. We perform an optimization:

$$\min_{\mathcal{M}^h, \mathcal{H}} \lambda_t L_{tip} + \lambda_w \mathcal{L}_{wrist} + \lambda_p \mathcal{L}_{pen}. \tag{3}$$

$\mathcal{L}_{tip} = \sum_{k\in\mathbf{o}} \left\| T_h^k \mathcal{K}_h^k(\mathcal{H})_i - \mathbf{c}_j^k \right\|$ penalizes the difference between a fingertip and the corresponding corner of the calibration cube c_j^k ($\mathbf{o}$ denotes the order of correspondences), where T_h^k is the hand-to-camera transformation at k-th time and $\mathcal{K}_h$ is the forward kinematics function to transform the i-th fingertip. We add $\mathcal{L}_{wrist}$ to penalize the distance between the wrist positions from the body mocap and from the hand marker to enforce hands are rigidly connected to the body mocap. We also add $\mathcal{L}_{pen}$ to prevent fingertips from penetrating the calibration cube. See supp. mat. for details.

3.5. Post-Processing to Enhance Qualities

As an advantage of the heterogeneous nature of our systems, we can leverage cues from both systems to reduce jitters and handle tracking failures.

Enhancing Hand Tracking. The global hand location estimated by the attached 3D marker may suffer from motion jitter or tracking failures due to occlusions or low camera visibility. Given two measurements, regarding the wrists from hand-attached markers and the body mocap, we perform a weighted interpolation to reduce motion jitters. In the case of tracking failure of the hand-attached 3D marker, we rely on the body motion from mocap since it's occlusion-free. We demonstrate the robustness of method in our experiments. See our supp. mat. for details.

Enhancing Object Tracking. Although we use multiple cameras with our 3D marker system, object tracking failure may still happen due to severe occlusions. As a solution, we leverage the body motion measurement from our system to enhance object tracking qualities. Specifically, we first infer whether an object is currently close to the human actor by checking the distances between human joint positions and the object's surface. If an object has been moved by the actor, we assume the object is near-rigidly connected to the close body joints, and apply the temporal transformation of the body joints to interpolate the missing object trajectories from the neighboring times. In practice, we apply this method for both reducing jitters and recovering the tracking failures.

Figure 5. An example of SMPL-X shape parameter fitting. (Left) Projected keypoints, mask and rendered SMPL-X with the optimized shape parameter. (Right) Rendered SDF within $5cm$ to visualize an affordance information using optimized SMPL-X.

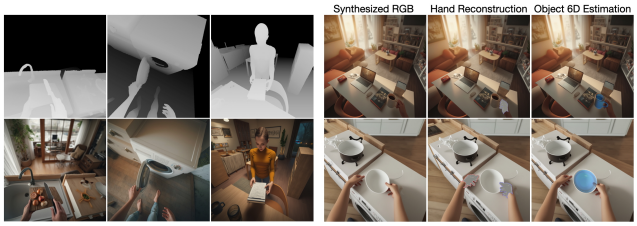

Figure 6. (Left) Examples of synthesized RGB images using ParaHome data. (upper) Rendered depth images of ParaHome data. (lower) Synthesized RGB image using text annotation, depth, 2D keypoints. (Right) HOI Reconstruction using synthesized RGB

Fitting Human Body Model. It is crucial to have 3D human mesh model aligned to each capture participant to reason about accurate contact between humans and the objects. We start by fitting SMPL-X [78] shape parameter to each participant. Using HQ-SAM [55], we initialize body segmentation masks and refine inaccurate pixels manually from 3 to 4 camera views. Then we optimize the parameters by aligning the rendered SMPL-X with the masks as well as projected joints. A result sample is shown in Fig. 5. For the pose parameter, directly computing SMPL-X pose parameter to fit into Xsens skeleton causes unnatural or inaccurate body motions even with optimized shape parameters due to different skeletal configurations between Xsens [75] and SMPL-X. To enhance the fitting quality, we design a simple autoencoder to learn natural body motion manifold space using both AMASS [70] and our Xsens body pose data. Using our trained model, we get an initial latent vector from Xsens body motion input, followed by further latent code optimization so that the decoded SMPL-X body pose fits to Xsens wrist, ankle, foots and figer tips. An example of fitting result is shown in Fig. 5. See the supp. mat. for details.

Synthesizing realistic RGB. Inspired by recent approaches Obman [42] and InterTrack [106] which demonstrate the

	Average Tracked	During Manipulation
Mean Cam. Num. (Visibility)	8.94	8.59
Reproj. Error [pixel]	1.021	1.016

IR Marker Num.	**Ours**	**4**	**7**	**10**	**20**	**40**
Tracking Success Ratio	1.0	0.76	0.79	0.86	0.90	0.93

Table 1. Evaluation on system settings (Up) Average Reprojection error detected in the scene and during manipulation by humans. (Down) Average number of tracked object ratio on multiple sampled windows. Numbers in the upper row represent number of virtual passive markers attached to the surface of the target object.

usefulness of synthetically produced RGB data paired with 3D GTs, we also provide corresponding synthetic RGB images for our ParaHome dataset by leveraging diffusion-based image synthesis model [3]. We render our 3D scenes from various viewpoints, including egocentric views, to produce depth maps and 2D body joint cues, which are used as the control inputs for synthesizing realistic RGB images with a ControlNet model [3]. We also find that incorporating our corresponding text annotations as an additional input for the ControlNet can further enhance the quality of the synthesis outputs. Examples are shown in Fig. 6.

To assess the realism and efficacy of the generated images, we perform a quantitative evaluation by applying off-the-shelf 3D estimation methods on our synthesized RGB as inputs to two tasks: RGBD-based 6D object pose estimation [122] and RGB-based 3D hand pose estimation [79]. We compare these results to those from existing benchmarks (ROPE [122], FreiHAND [129], and HO3D [36]), and demonstrate that recent state-of-the-art methods achieve similar patterns in performance. Notably, performance on our dataset is lower than on previous benchmarks, presumably due to our dataset's more complex human-object interaction scenarios with mutual occlusion. Qualitative results are shown in Fig. 6 and quantitative results are in supp. mat..

4. Validating Capture System and Data

4.1. System and Capture Evaluation

Impact of Camera Number. We demonstrate the advantage of using all 70 cameras against the alternative solutions of using fewer cameras. To quantify the impact of camera number, we count the number of detected marker corners by simulating systems with varying numbers of cameras where the subsets are derived from the complete set of cameras. As seen in Fig. 7, the detected ratio tends to increase linearly with the addition of cameras to the subsets without saturation. This observation implies that we take full advantage of our system without experiencing redundancy.

Advantage of 3D Marker Cube. We compare our 3D cube marker solution against an alternative solution of attaching markers on the surface of the targets (e.g., IR markers as in [23]). To simulate such IR marker system, we sample a set of points on the object mesh surface, assuming the sampled points as virtual markers. We choose a 4-minute long sequence, where our 3D marker-based object tracking is fully successful. Then, we assess the tracking ratio of virtual surface markers, considering varying numbers of markers attached as shown in Tab. 1. We simulate an occulsion to markers using human mesh, and assume tracking fails if they are visible from less than 3 cameras. Setting window length as 300, an object tracking fails if there exists any frame that less than 4 attached markers are tracked. Finally, we compute the average tracking success ratio in each marker setup, as shown in Tab. 1. As seen, attaching markers on the object surface suffers from tracking failures mainly because the actors tend to touch the object during interactions and the existence of multiple objects in the scene.

3D Marker Triangulation Quality. To measure the quality of tracking objects via ArUco markers, we report the average reprojection error of triangulated ArUco corners for all tracked scenes and for the interaction target parts. We also report the average number of cameras (visibility) used in corner triangulation for each case, when in contact and averaged in all scenes. Check Tab. 1 for the details.

Tracking Quality via Assessing Temporal Jitters. We compare the quality of tracking of our two heterogeneous systems, a multi-view system and the IMU-based wearable mocap suits. As a way to verify the quality, we assess the temporal motion jitters of both system, by plotting the derivative of acceleration for a certain interval of time. For this test, we capture a dedicated sequence, interacting with small and movable objects such as a cup and a kettle by moving them into various poses in multiple areas. The result is shown in Fig. 7, where both systems show similar jitter levels. Given that we use the commercial expert-level wearable motion capture system equipped with high-quality IMUs, this particular result demonstrates that our object tracking quality from cameras is comparable to the wearable capture system, enabling subtle interaction captures.

Hand Alignment Evaluation. To quantify the quality of our hand alignment procedure shown in Sec. 3.4, we perform a validation capture, where a participant touches random marker corners attached to the objects in random places inside the studio. Then, we measure the average distance between the fingertip and the target marker corner at the contact. The Average Position Error (APE) is 11 mm with 86 finger touches with various fingertips. Considering a possible bias due to the finger width(15-20 mm), our result demonstrates high precision in spatial alignment quality.

4.2. Post Processing Validation

We validate the performance of our post-processing method. As a way to quantify the performance, we simulate the tracking failure cases by intentionally dropping certain intervals of frames from successfully tracked sequences, and applying our post-processing method to recover the dropped

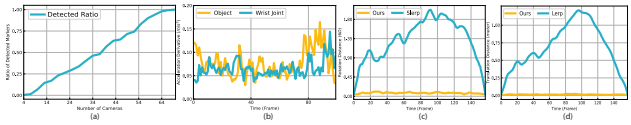

Figure 7. (a) Ratio of markers being detected with respect to 70 cameras to sampled number of cameras (b) Acceleration derivative of object and interacting hand wrist joint (c) Comparison of difference between original translation and hole-filled versions (d) Comparison of difference between original rotation (in 6D) and hole-filled versions

frames. By comparing the originally tracked cues with the recovered cues from the post-processing, we can assess the quality of our post-processing outputs.

Evaluating Object Tracking Enhancement. We perform our object-tracking enhancement process using the cues from hand tracking and compare its quality by a naive interpolation method based on linear functions (e.g. *lerp* for translation and *slerp* for rotation). The quantitative results are shown in Fig. 7, where our method powered with mocap device observations demonstrates much better performance in recovering the tracking failures.

Evaluating Hand Tracking Enhancement. To test the robustness of our system, we recover undetected markers of the dropped frames with our postprocessing algorithm. The error between hand joints computed using recovered marker positions and joint positions of original output is 9 mm.

5. Modeling Parameterized Episodic HOI

To learn the characteristic correlations between the environment $\mathbf{S}_e(t)$ changes and the human actions $\mathbf{S}_p(t)$, parameterizing both in a common spatio-temporal space is crucial, which motivates us to build our ParaHome system. One direction towards generative modeling for human-object interaction is to probabilistically model the distributions of possible configurations of humans and objects in the parametric 3D space: $P(\mathbf{S}(0:t))$, where $\mathbf{S}(t) = \{\mathbf{S}_e(t), \mathbf{S}_p(t)\}$ and t being a sequence length from 0 to T. Intuitively, this formulation captures the likelihood of plausible human and object configurations and dynamics. We can alternatively formulate it as a prediction problem: $\mathbf{S}_o(0:T) = \mathcal{F}(\mathbf{S}_i(0:T))$, where the input $\mathbf{S}_i(0:T)$ is a subset of full states $\mathbf{S}(0:T)$ and $\mathbf{S}_o(0:T)$ being the reconstructed unseen cues predicted by the model. For instance, we can build the model to infer 3D object movements from the human body motions, finger motions from desired object movements or formulate the HOI understanding as a future status forecasting task: $\mathbf{S}_o(t+1:t+w) = \mathcal{F}(\mathbf{S}_i(t-w:t))$. Furthermore, such formulation can also include the text descriptions $\mathcal{T}$ as an additional input condition. All these formulations can be interesting future research directions as a way to understand and learn spatio-temporal relations of human-object interactions in casual and natural activities, to which our ParaHome system and collected dataset can contribute. This paper explores two possible example problems: text-conditioned motion synthesis and object-guided motion synthesis.

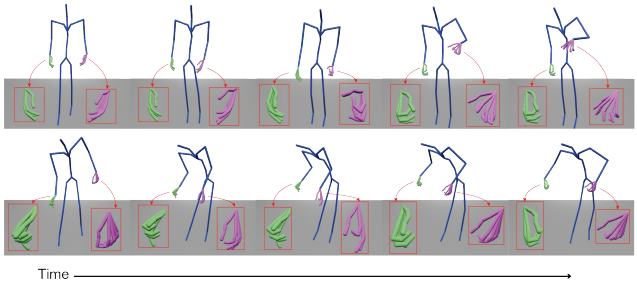

Figure 8. A sampled example of (Up) 'A person walks shortly then lift kettle with left hand and pours toward cup holding right' (Down) 'A person opens the cabinet'

5.1. Text Conditioned Motion Synthesis

We consider a task of generating human motions both body and hands given text descriptions. In such context, we aim to train a model that covers both diverse human actions (e.g. running, jumping) and sophisticated finger-level manipulations(e.g. turning on gasstove, cutting ingredients). Since our dataset mainly focuses on capturing manipulations, we utilize current large-scale text-motion paired dataset HumanML3D [33] to accomodate a diverse range of human motions. Specifically, we downsample HumanML3D dataset to match the size of our dataset to balance the occurrence during training. For evaluation, we compare the models in three variations : model trained with ParaHome soley, model trained with HumanML3D soley and the model trained with evenly mixed dataset to see which model closely approximates the distribution of the entire combined dataset. We also qualitatively demonstrate that generating an intermediate action that contains both features from ParaHome and HumanML3D is possible.

Formulation: We define our problem as learning a data distribution using a conditional variational model $\mathcal{F}$, which generates a sequence $\mathbf{S}_p(t:t+w)$ from noise $z \sim \mathcal{N}(0, \mathbf{I})$ and condition c , by training the model to reconstruct the data itself. Thus $\mathbf{S}_p(t:t+w) = \mathcal{F}(z, c)$ where $z \sim \mathcal{N}(0, \mathbf{I})$ and c being text description. For each time t, we define $\mathbf{S}_p(t)$ as a concatenation of root height $r_y \in \mathbb{R}^1$, root velocity $r_v \in \mathbb{R}^2$, root rotation $r_w \in \mathbb{R}^1$, localized joint position $j_p \in \mathbb{R}^{61\times3}$, joint relative rotation $j_w \in \mathbb{R}^{61\times6}$, joint local velocity $j_v \in \mathbb{R}^{61\times3}$, and foot contact $t_f \in \mathbb{R}^4$, thus $\mathbf{S}_p(t) = \{r_y, r_v, r_w, j_p, j_w, j_v, t_f\} \in \mathbb{R}^{743}$ following Text2Motion [33]. We use MDM [95] framework which utilizes diffusion model p_θ to generate samples from noise $x_N \sim \mathcal{N}(0, \mathbf{I})$ via reverse process

$$p_\theta(x_{n-1}|x_n, c) = \mathcal{N}(x_{n-1}; \mu_\theta(x_n, c), \sigma_n^2\mathbf{I})$$

where $x_n = \mathbf{S}_p^n$. Also instead of training model μ_θ to esti-

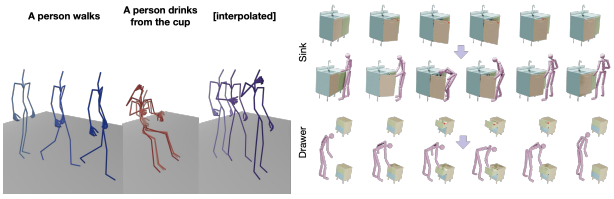

Figure 9. (Left) A sampled example of latent interpolation result. Right most is a sequence generated from the interpolated noise. (Right) Generated body motion(lower) given object motion trajectory from 0 to T(upper).

Method	R Precision (top3)↑	Mutimodal. dist↓	FID↓	Diversity→	Multimodality↑
Real	0.84	2.75	0.07	15.50	–
ParaHome	0.40	12.46	148.31	10.13	**3.52**
Humanml3D	0.46	8.26	74.66	8.21	2.23
Mixed	**0.73**	**3.75**	**0.85**	**15.06**	2.37

Table 2. Evaluation of models trained with different set of data

mate noise, we directly train model to get cleaned sample $\mathbf{S}_p^0$. Then optimization target in training the model is

$$\mathcal{L} = \mathbb{E}_{\mathbf{S}_p^0, n\sim[1,N]} \left\| \mathbf{S}_p^0 - \mathcal{F}(\mathbf{S}_p^n, n, c) \right\|_2^2$$

After training, generating samples given text condition c, we can invert the samples using DDIM and acquire an approximate latent vector [90]. In order to generate an intermediate action between generated samples $S_{p,1}^0$ from ParaHome and $S_{p,2}^0$ from HumanML3D, we invert the samples into $\tilde{S}_{p,1}^N$ and $\tilde{S}_{p,2}^N$ via the above process. Then we interpolate the latents $\tilde{S}_{p,\text{interp}}^N = u\tilde{S}_{p,1}^N + (1-u)\tilde{S}_{p,2}^N$ and run reverse diffusion process to get the motion $S_{p,\text{interp}}^0$.

Results: Following MDM, we measure *R-Precision*, *Multimodal Distance*, *FID*, *Diversity*, and *Multimodality* to compare each trained model's expressiveness and distance to the distribution of unseen test set of the mixed dataset. As in Tab. 2, the results on *R-Precision*, *FID* and *Diversity* show that the training with evenly mixed data describes both dataset without degrading both motions. This proves that our dataset and HumanML3D can be well blended and qualitative samples are shown in Fig. 8. Also, we qualitatively demonstrate a denoised result of the interpolated latent noise, where the original latents are inverted from samples generated with text descriptions from each dataset. The model trained with evenly mixed dataset and current algorithm have capability of mixing both distinct styles of data (i.e. variety of human movement motions in HumanML3D and manipulating actions in our dataset) thus can be used in generating intermediate actions as in Fig. 9 (Left) where a person walks and drinks simultaneously.

5.2. Object-guided Body Motion Synthesis

In this section, we consider the scenario of synthesizing a sequence of full-body motions given a sequence of object states. This is to investigate the viability of learning a spatio-temporal correlation between human motions and object state changes.

Object	condition	MPE↓(cm) rc-joints	rc-wrists	glb-root	glb-joints	MOE↓	Foot skating↓
Refrigerator	w/o $S_{p\to tp}$	1.05	2.66	7.79	9.26	0.85	**1.43**
	w/ $S_{p\to tp}$	**0.97**	**2.09**	**5.60**	**6.57**	**0.75**	1.74
Drawer	w/o $S_{p\to tp}$	3.81	5.25	5.58	9.92	0.66	**1.01**
	w/ $S_{p\to tp}$	**2.61**	**2.90**	**3.92**	**5.88**	0.66	1.58
Sink	w/o $S_{p\to tp}$	1.32	2.86	6.94	8.32	0.79	**1.08**
	w/ $S_{p\to tp}$	**1.21**	**1.93**	**3.83**	**4.84**	**0.76**	1.67
Washing machine	w/o $S_{p\to tp}$	1.33	3.14	11.29	14.43	1.00	**2.00**
	w/ $S_{p\to tp}$	**1.29**	**2.62**	**9.03**	**11.46**	**0.92**	2.60

Table 3. Quantitative results of object guided motion synthesis task.

Formulation: We formulate the problem as, $\mathbf{S}_p(t:t+w) = \mathcal{F}_{o\to p}\left(\mathbf{S}_{to}(t:t+w), S_{p\to o}(t)\right)$, where $\mathbf{S}_{to}(t:t+w) \in \mathbf{S}_e(t:t+w)$ is the desired status changes of a target object including rigid body motion and object-specific articulated parts representation. $S_{p\to o}(t)$ is the relative location and orientation of the person's root towards the object at the initial time. The output $\mathbf{S}_p(t:t+w)$ is the synthesized human motion in the corresponding time window. For representing human motion, we use both person root-centric coordinates as in [45] and global coordinates. As a model for achieving this task, we extend a transformer-based model for text to 3D human motion synthesis [83], by introducing Object State Encoder $\mathcal{O}_{enc}$ as a replacement of the text encoder module. We provide more details on data representation, model architecture and experimental results for alternative representation options in our supp. mat..

Results: To assess the quality of generated human body motions, we quantitatively measure mean positional errors from the ground truth and the output in two different coordinate systems : (1) root-centered coordinate, (2) global coordinate. For case (1), we compute entire joints error (*rc-joints*) and wrist error (*rc-wrist*). For case (2), we compute root-position error (*glb-root*) and global-joints error (*glb-joints*). Additionally we compute foot-skating as in [44]. The quantitative results are shown in Tab. 3 including qualitative results in Fig 9. The qualitative results demonstrate convincing performance in synthesizing plausible human motions for the target tasks. Also, using relative spatial cues as input (*with* $S_{p\to tp}(0)$) contributes to more accurate motion synthesis. We add further analysis regarding the results in supp. mat..

6. Discussion

We present a ParaHome system, specifically designed to capture human motions, finger movements, and object dynamics in 3D during natural human-object interactions in a studio apartment setting. Leveraging our system, we collect a new HOI dataset, which offers key improvements over existing datasets, including (1) capturing 3D body and hand motion alongside 3D object movement within a contextual home environment; (2) encompassing human interaction with multiple objects and concurrent usage of objects; (3)

including articulated objects with multiple parts and diverse scale objects. We explore the opportunities in leveraging our dataset for human motion synthesis via generative modeling.

We endeavor to improve our system further to handle limitations. Firstly, we plan to evolve our system towards a markerless approach by eliminating artificial markers and constructing motion priors from the current database. Secondly, we aim to replicate our system in different room settings, with more diverse objects ensuring better generalizability.

References

[1] Andreas Aristidou, Yiorgos Chrysanthou, and Joan Lasenby. Extending FABRIK with model constraints. *Comput. Animat. Virtual Worlds*, 2016. 17

[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. *CVPR*, 2022. 1, 2, 15

[3] Black Forest Lab, 2024. https://huggingface.co/black-forest-labs. 6, 18

[4] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. *ECCV*, 2020. 1

[5] Christoph Bregler, Jitendra Malik, and Katherine Pullen. Twist based acquisition and tracking of animal and human kinematics. *IJCV*, 2004. 2

[6] Thomas Brox, Bodo Rosenhahn, Juergen Gall, and Daniel Cremers. Combined region and motion-based 3D tracking of rigid and articulated objects. *TPAMI*, 2010. 2

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. 2017. 18

[8] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. *ECCV*, 2020. 2

[9] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, 2021. 2

[10] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. pages 1577–1585, 2024. 3

[11] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. *CVPR*, 2021. 1, 2

[12] Kong Man Cheung, Simon Baker, and Takeo Kanade. Shape-from-silhouette across time part i: Theory and algorithms. *IJCV*, 2005. 2

[13] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. pages 20577–20586, 2022. 2

[14] Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. *arXiv preprint arXiv:2403.17827*, 2024. 2

[15] Stefano Corazza, Lars Mündermann, Emiliano Gambaretto, Giancarlo Ferrigno, and Thomas P. Andriacchi. Markerless Motion Capture through Visual Hull, Articulated ICP and Subject Specific Model Generation. *IJCV*, 2010. 2

[16] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. *CVPR*, 2020. 2

[17] Yudi Dai, Yitai Lin, Chenglu Wen, Siqi Shen, Lan Xu, Jingyi Yu, Yuexin Ma, and Cheng Wang. Hsc4d: Human-centered 4d scene capture in large-scale indoor-outdoor space using wearable imus and lidar. 2022. 2

[18] Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. *SIGGRAPH*, 2008. 2

[19] Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. *NeurIPS*, 35:13800–13813, 2022. 2

[20] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. *arXiv preprint arXiv:2311.16097*, 2023. 2

[21] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *CVPR*, 2020. 2

[22] Einstar, 2023. https://www.einstar.com/. 14

[23] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. 2023. 2, 3, 6, 14, 15

[24] Y. Furukawa and J. Ponce. Dense 3d motion capture from synchronized video streams. *CVPR*, 2008. 2

[25] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. *CVPR*, 2009. 2

[26] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. *CVPR*, 2018. 1, 2, 15

[27] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6): 2280–2292, 2014. 3

[28] D. Gavrila and LS Davis. Tracking of humans in action: A 3-D model-based approach. *ARPA Image Understanding Workshop*, 1996. 2

[29] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, pages 1–12. Wiley Online Library, 2023. 3

[30] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *CVPR*, 2021. 2

[31] Markus Gross, Stephan Würmlin, Martin Naef, Edouard Lamboray, Christian Spagno, Andreas Kunz, Esther Koller-Meier, Tomas Svoboda, Luc Van Gool, Silke Lang, Kai Strehlke, Andrew Vande Moere, and Oliver Staadt. Blue-c: A spatially immersive display and 3d video portal for telepresence. *SIGGRAPH*, 2003. 2

[32] Ralph Gross and Jianbo Shi. The cmu motion of body (mobo) database. 2001. 1

[33] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7

[34] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. *CVPR*, 2022. 2

[35] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. *CVPR*, 2021. 2

[36] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. *CVPR*, 2020. 1, 2, 6, 17, 18

[37] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. *CVPR*, 2022. 2, 15

[38] Shangchen Han, Beibei Liu, Robert Wang, Yuting Ye, Christopher D Twigg, and Kenrick Kin. Online optical marker-based hand tracking with deep labels. *ACM TOG*, 37(4):1–10, 2018. 2

[39] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, 2019. 1

[40] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *ICCV*, 2021. 2

[41] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. *arXiv preprint arXiv:2302.00883*, 2023. 2

[42] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 5

[43] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 1, 2

[44] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. *NeurIPS*, 35:4244–4256, 2022. 8

[45] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 2016. 8, 18

[46] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. *GCPR*, 2022. 2, 15

[47] 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, 2013. 1

[48] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. pages 14713–14724, 2023. 2

[49] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *ICCV*, 2023. 2, 14, 15

[50] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. 2024. 2, 15

[51] Payam Jome Yazdian, Eric Liu, Li Cheng, and Angelica Lim. Motionscript: Natural language descriptions for expressive 3d human motions. *arXiv e-prints*, pages arXiv–2312, 2023. 2

[52] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. *CVPR*, 2015. 1

[53] Takeo Kanade, Peter Rander, and P.J. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multimedia*, 1997. 2

[54] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In *ICCV*, pages 14632–14643, 2023. 2

[55] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. *arXiv:2306.01567*, 2023. 5

[56] Roland Kehl and Luc Van Gool. Markerless tracking of complex human motions from multiple views. *CVIU*, 2006. 2

[57] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. *arXiv preprint arXiv:2307.07511*, 2023. 2

[58] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. *ICCV*, 2021. 15

[59] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. 2017. 17

[60] Jiye Lee and Hanbyul Joo. Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. *CVPR*, 2024. 2

[61] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 2023. 1, 2

[62] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. *ECCV*, 2024. 2

[63] Han Liang, Yannan He, Chengfeng Zhao, Mutian Li, Jingya Wang, Jingyi Yu, and Lan Xu. Hybridcap: Inertia-aid monocular capture of challenging human motions. In *AAAI*, 2023. 2

[64] Jianhui Liu, Yukang Chen, Xiaoqing Ye, and Xiaojuan Qi. Prior-free category-level pose estimation with implicit space transformation. 2023. 17

[65] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*, 2021. 2

[66] Xinpeng Liu, Haowen Hou, Yanchao Yang, Yong-Lu Li, and Cewu Lu. Revisit human-scene interaction via space occupancy. *arXiv preprint arXiv:2312.02700*, 2023. 2

[67] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. *CVPR*, 2022. 2, 15

[68] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. pages 21740–21751, 2024. 2, 15

[69] Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Grasping diverse objects with simulated humanoids. *arXiv preprint arXiv:2407.11385*, 2024. 2

[70] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. *ICCV*, 2019. 1, 5, 17

[71] Manus, 2023. https://www.manus-meta.com/. 3, 15

[72] T. Matsuyama and T. Takai. Generation, visualization, and editing of 3d video. *3DPVT*, 2002. 2

[73] Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J. Gortler, and Leonard McMillan. Image-based visual hulls. *SIGGRAPH*, 2000. 2

[74] Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. Generating continual human motion in diverse 3d scenes. *arXiv preprint arXiv:2304.02061*, 2023. 2

[75] Movella, 2023. https://base.xsens.com/. 3, 5, 15

[76] Yinyu Nie, Angela Dai, Xiaoguang Han, and Matthias Nießner. Pose2room: Understanding 3d scenes from human activities, 2022. 3

[77] Shaohua Pan, Qi Ma, Xinyu Yi, Weifeng Hu, Xiong Wang, Xingkang Zhou, Jijunnan Li, and Feng Xu. Fusing monocular images and sparse imu signals for real-time human motion capture. In *SIGGRAPH Asia*, pages 1–11, 2023. 2

[78] 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 *CVPR*, 2019. 2, 5, 18

[79] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. 2024. 6, 17

[80] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023. 2

[81] Benjamin Petit, Jean-Denis Lesage, Edmond Boyer, and Bruno Raffin. Virtualization Gate. *SIGGRAPH Emerging Technologies*, 2009. 2

[82] Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? *CVPR*, 2023. 3

[83] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. *ECCV*, 2022. 2, 8

[84] Ralf Plankers and Pascal Fua. Articulated Soft Objects for Multi-View Shape and Motion Capture. *TPAMI*, 2003. 2

[85] Gerard Pons-Moll, Vladimir Guzov, Julian Chibane, Riccardo Marin, Yannan He, and Torsten Sattler. Interaction replica: Tracking human-object interaction and scene changes from human motion. 2023. 2

[86] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 15

[87] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. *CVPR*, 2022. 2

[88] Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. Hulc: 3d human motion capture with pose manifold sampling and dense contact guidance. In *ECCV*. Springer, 2022. 1

[89] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2):4, 2010. 1

[90] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 8

[91] Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Fast articulated motion tracking using a sums of gaussians body model. *ICCV*, 2011. 2

[92] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. *ECCV*, 2020. 1, 2, 15

[93] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *CVPR*, 2022. 2

[94] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. *CVPR*, 2023. 2

[95] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 7

[96] Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. *ECCV*, 2022. 2

[97] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. *TOG*, 2008. 2

[98] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. *SIGGRAPH*, 2008. 2

[99] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *CVPR*, 2021. 2

[100] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *CVPR*, 2021.

[101] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *CVPR*, 2022. 2

[102] Xi Wang, Gen Li, Yen-Ling Kuo, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In *3DV*. IEEE, 2022. 2

[103] Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023. 1, 2

[104] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *NeurIPS*, 2022. 1, 2

[105] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *CVPR*, 2023. 2

[106] Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. Intertrack: Tracking human object interaction without object templates. *arXiv preprint arXiv:2408.13953*, 2024. 2, 5

[107] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. *ICCV*, 2023. 2

[108] Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *arXiv preprint arXiv:2403.19652*, 2024. 2

[109] Jie Yang, Xuesong Niu, Nan Jiang, Ruimao Zhang, and Siyuan Huang. F-hoi: Toward fine-grained semantic-aligned 3d human-object interactions. *arXiv preprint arXiv:2407.12435*, 2024. 2

[110] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *ICCV*, 2021. 2

[111] Hongwei Yi, Chun-Hao P Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J Black. Mime: Human-aware 3d scene generation. In *CVPR*, pages 12965–12976, 2023. 3

[112] Hongwei Yi, Justus Thies, Michael J. Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. *ECCV*, 2024. 2

[113] Marsil Zakour, Partha Pratim Nath, Ludwig Lohmer, Emre Faik Gökçe, Martin Piccolrovazzi, Constantin Patsch, Yuankai Wu, Rahul Chaudhari, and Eckehard Steinbach. Adl4d: Towards a contextually rich dataset for 4d activities of daily living. *arXiv preprint arXiv:2402.17758*, 2024. 2

[114] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. 2024. 2, 15

[115] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: Neural manipulation synthesis with a hand-object spatial representation. *ACM TOG*, 40(4), 2021. 2

[116] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. Artigrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. *arXiv preprint arXiv:2309.03891*, 2023. 3

[117] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*, 2023. 1, 2, 15

[118] Jiajun Zhang, Yuxiang Zhang, Hongwen Zhang, Xiao Zhou, Boyao Zhou, Ruizhi Shao, Zonghai Hu, and Yebin Liu. Ins-hoi: Instance aware human-object interactions recovery. *arXiv preprint arXiv:2312.09641*, 2023. 2

[119] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. 2024. 2

[120] Jiajun Zhang, Yuxiang Zhang, Liang An, Mengcheng Li, Hongwen Zhang, Zonghai Hu, and Yebin Liu. Manidext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion. *arXiv preprint arXiv:2409.09300*, 2024. 3

[121] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 2

[122] Mengchen Zhang, Tong Wu, Tai Wang, Tengfei Wang, Ziwei Liu, and Dahua Lin. Omni6d: Large-vocabulary 3d object dataset for category-level 6d object pose estimation. 2025. 6, 17

[123] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *ECCV*, 2022. 2

[124] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. *CVPR*, 2023. 3

[125] Yanzhao Zheng, Yunzhou Shi, Yuhao Cui, Zhongzhou Zhao, Zhiling Luo, and Wei Zhou. Coop: Decoupling and coupling of whole-body grasping pose generation. pages 2163–2173, 2023. 2

[126] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. *ECCV*, 2022. 2

[127] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Gears: Local geometry-aware hand-object interaction synthesis. pages 20634–20643, 2024. 2

[128] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 17

[129] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. *ICCV*, 2019. 2, 6, 17, 18

[130] Gaspard Zoss, Derek Bradley, Pascal Bérard, and Thabo Beeler. An empirical rig for jaw animation. *ACM TOG*, 37 (4):1–12, 2018. 3

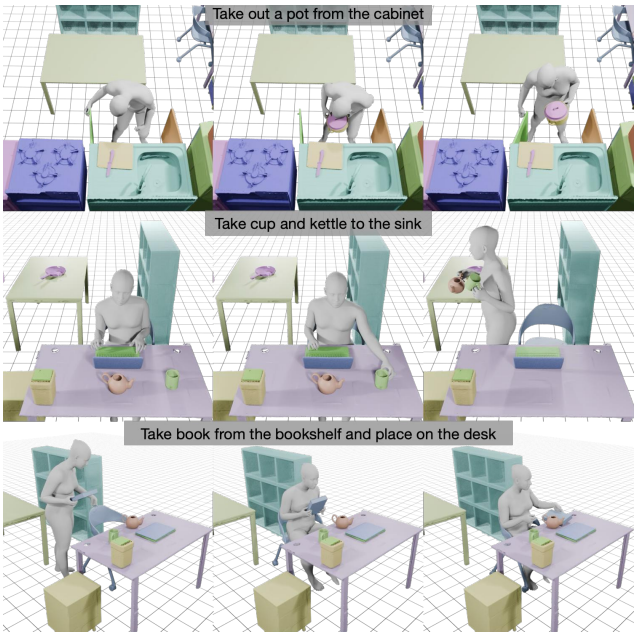

Figure 10. Rendered scenes and text annotation for each scene from example scenario.

A. Dataset Details

A.1. Dataset Contents

Scanned Object Mesh. We obtain high-quality 3D mesh scans of all objects placed in our system via an Einstar3D scanner [22]. We scan each object at least twice to reduce the unscanned areas or holes by changing the orientations of the objects (e.g., up-side-down), and fuse the scanned meshes via manual alignments. They are zero-centered and scaled to a metric scale.

Object Articulation Information. Objects with articulation contain axis a_e. If the part has a revolute joint, we include pivot point p_e additionally. These are defined in the object canonical space and are utilized in getting each object part-transformation toward the camera space.

Object Position and Orientation in the Camera Space. Each object's spatial information is computed using the per-frame transformation of markers attached to each object.

Relative Orientation of Hand/Body Joints. Orientation of each hand and body joints with respect to their parent joints is recorded and processed via a motion capture system.

3D Hand/Body Joint Positions in the Camera Space. With the positions of markers attached to the body in the mocap space acquired via body alignment protocol, translation and orientation of body to camera space are obtained using the positions of corresponding markers in the camera spaces. We compute the positions of two hands and body using the obtained translation and orientation.

Text Annotation for Each Action. For each capture, participants receive verbal instructions detailing the actions they will perform. These instructions specify which objects to interact with and how to interact with them, as illustrated in Fig. 10. The instructions are recorded and synchronized with the motion data. Additionally, we manually inspect the instrument to create more accurate text annotations, ensuring they are reliably mapped to each action.

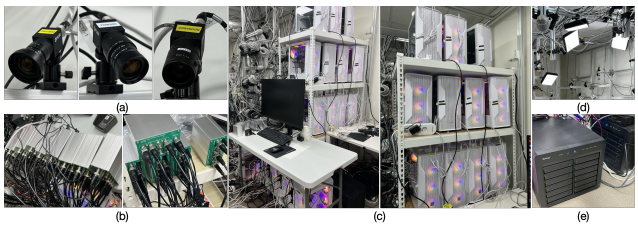

Figure 11. System Devices. (a) RGB cameras with 3 types of lenses. (b) Signal distributers (c) Desktop machines (d) LED Lights (e) NAS storage systems

Per-frame Contact Information. At each frame where contact between Left/Right/Body and object occurs, the corresponding frame and object category/body part information is recorded.

A.2. Dataset Comparison

As shown, our ParaHome dataset is the comprehensive dataset which capture all authentic and dynamic human-object interaction scenarios in a natural room environment. Our dataset includes dexterous body motion and movement of all objects in the scene and encompasses natural manipulation motions involving articulated objects and multiple objects even in concurrent usage scenarios. Our capture scenarios feature natural and sequential manipulations like cooking, as shown in our supplementary video. Furthermore, we collected data from 38 participants, capturing a wide range of motion styles across individuals.

B. System Details

B.1. Using ArUco Markers

Even though several works [23, 49] utilized IR markers for motion tracking, we find using ArUco markers to be more suitable in our capture system. We aim to capture in a broader spatial spectrum(i.e. entire room setting filled with objects) involving multiple interactions in a single capture time. Such environments filled with multiple furniture/objects and hand-object interactions involving multiple direct contacts, cause a significant occlusion as simulated in Sec.4.1 of our main manuscript. Even though ArUco markers have its downsides in corrupting RGB data and influencing natural human motions, using RGB data is not within our interest as mentioned Sec 6. of main manuscript and we empirically placed markers to minimize such interruptions.

Dataset	hours #	subject #	object #	body	hand #	contact	obj. 6d.	obj. artic.	multi obj.	setup
GRAB [92]	3.8	10	51	✓	2	✓	✓	✗	✗	standing
BEHAVE [2]	4.2	8	20	✓	-	✓	✓	✗	✗	portable
InterCap [46]	0.6	10	10	✓	2	✓	✓	✗	✗	portable
FHPA [26]	0.9	6	26	✗	1	✗	✓	✗	✗	room
H2O [58]	1.1	4	8	✗	2	✗	✓	✗	✗	table
H2O-3D [37]	-	5	10	✗	2	✗	✓	✗	✗	table
HOI4D [67]	22.2	9	800(16)	✗	1	✗	✓	✓	✗	room
Chairs [49]	17.3	46	81	✓	2	-	✓	✓	✗	standing
ARCTIC [23]	1.2	10	11	✓	2	✓	✓	✓	✗	standing
NeuralDome [117]	4.6	10	23	✓	2	✓	✓	✓	✗	standing
OAKINK2 [114]	12.38	9	75	✓	2	✓	✓	✓	✓	table
TACO [68]	2.53	14	196(20)	✗	2	✓	✓	✗	✓	table
TRUMANS [50]	15	7	20(placeholders)	✓	2	✓	✓	✓	✓	room
Ours	8.1	38	22	✓	2	✓	✓	✓	✓	room

Table 4. Comparison of existing human-object interaction datasets

B.2. Hardware Details

In order to cover the entire volume of the room and to reduce occlusion issues, we install 70 RGB industrial cameras, BFLY-31S4C-C. The cameras capture videos at 30Hz in 2048×1536 resolution. We set the exposure time at $3msec$, which shows a good balance between low-motion blur and sufficient brightness. We use three types of lenses (thirty $3mm$ lenses, twenty $5mm$ lenses, and twenty $6mm$ lenses), where the wide-angle lens ($3mm$) is helpful in capturing wide area. We calibrate the cameras using Structure-from-Motion via COLMAP [86] with multiple randomly patterned fabrics placed in our system. We provide pre-calibrated initial intrinsic parameters for the three types of lenses derived from 2 or 3 samples of lenses for better convergence in camera pose estimation. We scale the calibrated 3D space into a real-world metric (in meters) by locating checkerboards with known sizes during camera calibration.

All cameras, the motion capture suit, and gloves are synced and gen-locked via a common square wave signal that comes from the motion capture device to synchronize two heterogeneous systems, which is crucial to precise HOI captures. To deliver the sync signals to a large number of cameras, we utilized 11 signal distributors in a hierarchical manner, each of which can be connected to 8 cameras via GPIO cables. We use 1 master and 18 slave desktop machines to control the cameras and process captured records. Each slave machine is connected to 3 or 4 cameras via Ethernet cables and equipped with a 4-port 1G ethernet board, and 2 SSDs with a capacity of 500GB and 1TB each. 15 LED lights ($4500lm$) are installed to provide sufficient illumination. Pictures of our system devices are shown in Fig. 11.

To capture both body motion and subtle hand motions, we use IMU-based motion capture equipments, Xsens motion suit [75] and Manus hand gloves [71]. The body motion system captures the motions at 60Hz.

Object	Part1	Part2
Sink	revolute	revolute
Laptop	revolute	-
Drawer	sliding	sliding
Gas stove	revolute	revolute
Microwave	revolute	-
Trashbin	revolute	-
Washing machine	revolute	-
Refrigerator	revolute	revolute

Table 5. Part information of articulated objects

C. Data Acquisition

C.1. Modeling Object Articulations.

To capture the movement of articulated objects, we model each object as a parametric 3D model by defining the object-specific articulated motion parameters. This modeling requires scanning individual parts separately and compositing them in a canonical space by defining axis direction, pivot points, revolute joints, and so on, based on the object types. During HOI captures, we track the motion of each part via our marker system (e.g. monitor of a laptop and the base), from which we compute the articulated motion parameters. In this subsection, we describe the process of modeling articulated objects as parametric 3D models. Articulation information of each object with multiple parts is shown in Tab. 5.

To find axis $\mathbf{a}_e$ and pivot point $\mathbf{p}_e$ of the articulated objects, we capture markers attached to each object part at different part states separately and acquire each marker corners in the ParaHome space as $\{m_i(t)\}_{t=1}^n$. Prior to applying algorithm, we transform marker corners $\{m_i(t)\}_{t=1}^n$ back

to object canonical space with $T^{-1}_{mar \to obj}$ and utilize transformed marker corners in the canonical space $\{m'_i(t)\}^n_{t=1}$. For the sliding joint, axis a_e can easily be calculated using marker corners at time t and t' as:

$$\mathbf{a}_e = \frac{m'_i(t) - m'_i(t')}{\|m'_i(t) - m'_i(t')\|}$$

In case object part has a revolute joint, we start initializing an axis $\mathbf{a}_e$ and each relative state $\Delta s_e(t, t') = |s_e(t) - s_e(t')|$ between time t and t' (for the target articulated object captured at different n number of states, time t and time t' satisfies $t \neq t'$ and $t, t' \in \{1, 2, \cdots, n\}$). Then we apply optimization with marker corners toward all possible pairs of times t and t'. Let f be a map defining rotation transformation with respect to pivot and given axis-angle and denote as $\mathbf{T}_{t' \to t} = f(a_e, \Delta s_e(t, t'), p_e)$. Then for a set of all possible time pairs $\mathbf{P}$, the optimization target for axis a_e, relative state Δs_e and pivot p_e is defined as:

$$\underset{a_e, \Delta s_e(t,t'), p_e}{\arg\min} \sum_{(t,t') \in \mathbf{P}} \|m'_i(t) - \mathbf{T}_{t' \to t} m'_i(t')\|^2$$

Since initial axis $\mathbf{a}_e$ and pivot $\mathbf{p}_e$ are initialized in the object canonical space, we directly utilize acquired information to derive transformations using detected markers for each capture data.

C.2. Body Alignment Detail (Sec 3.4 in Main Paper)

In this subsection, we provide additional details of our spatial alignment process between a multiview camera system and wearable motion capture systems, described in Sec. 3.2 in our main manuscript.

To resolve the issue of imperfect body and hand skeleton scale from the wearable motion capture system, we attach 3 or 4 ArUco markers to each near-rigid body part (torso, hands, upper arms, lower arms, upper legs, lower legs) to assign correspondences. During alignment capture, participants perform the range-of-motion movement by rotating their arms and legs while pinned or bent, particularly twisting their wrists to locate each hand wrist. With the captured data, we optimize body skeleton configuration $\mathcal{B} = \{\mathcal{O}\}$ and body markers locations $\mathcal{M}^b$ via gradient decent with a learning rate of 0.008 for 50 epochs. Specifically for weights of body and foot, $\lambda_b = 100, \lambda_f = 5000$ are used. In the case that alignment is not well optimized, we additionally penalize excessive length change in spines and difference in skeleton lengths between the left and right sides of the body by adding an extra regularization term. Once the alignment procedure is finished, we remove markers (all from the upper legs, one for each upper arm, lower arm, and lower leg) to minimize interference with the movements of the participant. The selection of remaining markers is determined based on their importance during captures, where we assess

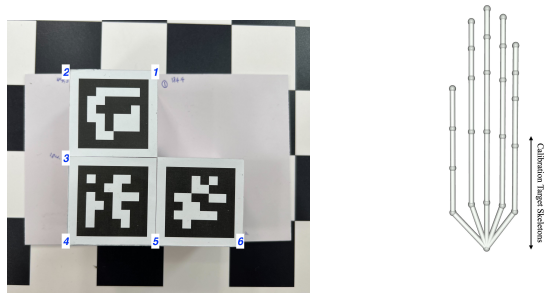

Figure 12. (Left) The hand calibration structure to precisely measure hand skeleton configuration and to find the relative locations of hand markers attached to the wrist in hand-centric coordinate (Right) Hand skeleton and Calibration targets

Corner #	Hand Side	Seq1	Seq2	Seq3	Seq4
1,2	Right	1, 2	1, 3	1, 4	1, 5
1,3	Right	1, 2	1, 3	1, 4	1, 5
2,4	Right	1, 2	1, 3	1, 4	1, 5
5,2	Right	1, 2	1, 3	1, 4	1, 5
6,2	Right	1, 2	1, 3	1, 4	1, 5
6,3,2	Right	1, 2, 3	1, 3, 4	1, 4, 5	-
2,1	Left	1, 2	1, 3	1, 4	1, 5
3,1	Left	1, 2	1, 3	1, 4	1, 5
4,2	Left	1, 2	1, 3	1, 4	1, 5
5,2	Left	1, 2	1, 3	1, 4	1, 5
6,2	Left	1, 2	1, 3	1, 4	1, 5
2,5,6	Left	1, 2, 3	1, 3, 4	1, 4, 5	-

Table 6. Hand Calibration Protocol

their importance by evaluating whether their absence would compromise the accuracy of body positioning in the camera space. Check our supplementary video for an example of body alignment motion.

C.3. Hand Calibration Structure and Protocol (Sec 3.4 in Main Paper)

As human usually handle objects with their fingers, fingertips play an important role during interaction. We made the calibration structure to better locate fingertips and find the hand skeletons and relative locations between the attached hand markers to each wrist. The hand calibration structure is composed of three cubes with ArUco markers and the ordered 3D corner vertices of the structure are defined as $C = \{\mathbf{c}_i \in \mathbb{R}^3\}^6_{i=1}$ as shown in Fig. 12. During the hand calibration procedure, we request each participant to touch the calibration structure's corners with their fingertips. We instruct them to touch specified multiple corners at each step using two or three fingertips. A participant undergoes 23 steps of such touching processes per-hand. The Tab. 6 comprises hand calibration instructions for subjects to follow. Corner # is a set of two or three target corner numbers of the calibration structure which the subject should contact with their fingertips. Also, the orders of fingers to touch the target corners are specified with numbers corresponding to each finger, which are (1:Thumb, 2:Index, 3:Middle, 4:Ring,

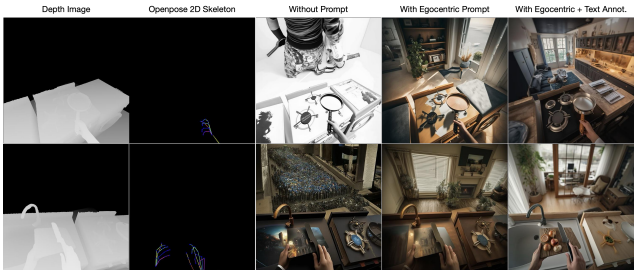

Figure 13. Synthesized RGBs and Comparison between with or without text annotation.

5:Little). An example of the hand calibration procedure is shown in our supplementary video.

C.4. Implementation Details on Hand Calibration

Here we describe the details of the hand calibration method. As described in section 3.4, optimization parameters are hand skeleton configuration $\mathcal{H} = \{\mathcal{S}^h, \mathcal{O}^h\}$ and positions of 3D markers in the local hand-centric coordinate $\mathcal{M}^h$. We empirically decide the general target range of the optimization skeleton to the palm area shown in Fig. 12 and add constraints that limit the skeleton scales(s_i) for each skeleton segment i between $0.8 \leq s_i \leq 1.2$, and additional skeleton offset value(δ_j) for target joints j with $|\delta| \leq 0.01$ in meter scale to avoid unnatural deformation of hand skeleton. The location of hand markers $\mathcal{M}^h$ is optimized through a total of 150 iterations. The skeleton scale and additional offset are optimized starting from 50 and 100 iterations each. We use three losses, L_{tip} to measure the Euclidean distance from the hand tips to paired corners, L_{wrist} to measure the distance from the wrist location from body motion capture device and the wrist position computed by hand marker position and L_{pen} to measure penetration of hand to the calibration structure. The penetration loss is computed by a cosine similarity between the calibration structure's normal vector and the target corner-to-hand tip vector. In summed loss $\lambda_t L_{tip} + \lambda_w \mathcal{L}_{wrist} + \lambda_p \mathcal{L}_{pen}$, losses are weighted equally by $\lambda_t = 1, \lambda_w = 1, \lambda_p = 1$. But they are manually adjusted based on the touch accuracy and body calibration accuracy per participant. After the alignment process, the average Euclidean distance between the corner and the target fingertip results in 0.83 (in centimeters).

C.5. Fitting Human Body Model

We illustrate details on fitting SMPL-X [70] pose parameters to Xsens skeletons using the optimized shape parameters. For SMPL-X hand pose, we devise an optimization scheme which runs FABRIK solver [1] initially to get SMPL-X hand skeleton retargetted to Xsens hand and optimize each hand joint pose directly to fit into each retargetted joint positions. For body pose, we do not consider global orientation and translation in acquiring body pose for simplicity. As an input data representation, we split each sequence with a window size of 60, and reform body joint rotations except for hands in 6D representations [128], thus $x_{in} \in \mathbb{R}^{60\times21\times6}$ per batch. For the model, we use a variation of Temporal Convolution Network [59] for the encoder and decoder. During training, we define the default reconstruction loss, $\mathcal{L}_{\text{recon}}$, for joint rotation and further incorporate the end effector loss, $\mathcal{L}_{\text{end}}$. This additional loss includes the SMPL-X vertices of the hands and legs, as well as the wrist, foot, and hand tip joints. To regulate any present noise in the motion, we also add joint velocity loss as an regularization. Then the loss sums up to:

$$\mathcal{L} = \lambda_{recon}\mathcal{L}_{recon} + \lambda_{end}\mathcal{L}_{end} + \lambda_{vel}\mathcal{L}_{vel}$$

After training, we extract windows of all sequences with step size of 30, and initialize a latent code z_{pose} by feeding the encoder with the Xsens joint rotation data by roughly matching joint category between two different skeletons(i.e. jLeftT4Shoulder of Xsens to left collar of SMPL-X). Then we optimize z_{pose} by feeding into the trained decoder to fit with Xsens skeleton wrist, hand tip, ankle, and foot joints. Thus denoting a set of paired Xsens and SMPL-X target joints as $\mathcal{J}$, we formulate the optimization problem as:

$$z^*_{\text{pose}} = \arg\min_{z_{\text{pose}}} \sum_{j_{\text{xsens}}, j_{\text{smplx}} \in \mathcal{J}} \|j_{\text{xsens}} - j_{\text{smplx}}\|^2$$

After optimization, we use decoded output body pose using z^*_{pose}. Since we sample sequences as 60-length windows with step size of 30, there exists discrepancies in body poses where contiguous windows overlap. We use *slerp* to compensate for such discontinuities for each joint pose parameters.

C.6. Synthesizing Realistic RGB

Utilizing the ParaHome dataset, which provides diverse and rich 3D motion data, we generate RGB images from various viewpoints, all aligned with 3D annotations. We employ

Dataset	Method	AUC@IoU$_{25}$↑	AUC@IoU$_{50}$↑	AUC@IoU$_{75}$↑
ROPE [122]	IST-Net [64]	28.7	10.6	0.5
ROPE [122]	GenPose++ [122]	**39.9**	**19.1**	**2.0**
ParaHome$_{all}$	GenPose++ [122]	26.4	10.3	0.6
ParaHome$_{rigid}$	GenPose++ [122]	29.7	12.2	0.9

Table 7. Quantitative comparison of category-level object pose estimation on ROPE [122] and ParaHome synthetic data. Since the two datasets differ in the presence of articulation, we divide ParaHome data into two subsets, all including articulation objects and $rigid$ with only rigid objects.

Dataset	Method	PA-MPJPE↓	PA-MPVPE↓	F@5↑	F@15↑
FreiHAND [129]	HaMeR [79]	6.0	5.7	0.785	0.990
HO3D [36]	Pose2Mesh [79]	12.5	12.7	0.441	0.909
HO3D [36]	HaMeR [79]	7.7	7.9	0.635	0.980
ParaHome(Ours)	HaMeR [79]	9.47	9.46	0.25	0.85

Table 8. Quantitative comparison of 3D hand pose reconstruction on FreiHand, HO3D and ParaHome synthetic data. PA-MPVPE and PA-MPJPE are meausred in mm.

Figure 14. Rendered results of 6d reconstruction model on synthesized data. (Left) Successful cases. (2) Failure cases due to the occlusion.

Figure 15. Rendered results of the 3D hand reconstruction model on synthesized data. (Left) Successful cases. (Right) Cases with large 3D keypoints loss due to occlusion.

a diffusion-based image synthesis model [3] combined with ControlNet to create 2D RGB images consistent with the 3D data. Human-object interaction scenes are rendered from multiple perspectives, including egocentric, high-angle, and front-facing views. Rendered depth maps and OpenPose [7] joint information are integrated, along with text prompts to enhance image quality and alignment with the original data. The impact of including text information is shown in Fig. 13, implying improved alignment and realism. Quantitative results on off-the-shelf 3D estimation models are presented in Table 7 and Table 8. For object 6D estimation, our synthetic data achieves accuracy comparable to the ROPE dataset, demonstrating the realism of the generated images for model to detect. Specifically, occlusions from hand interactions and complex object articulation in the ParaHome dataset result in lower accuracy, as shown in Fig.14, suggesting future potential improvement. For 3D hand pose estimation, the synthetic data performs competitively with other datasets quantitatively, though occlusions during manipulation lead to slightly reduced accuracy compared to HO3D[36] and FreiHAND [129], as illustrated in Fig. 15.

D. Sequence Example Visualization

D.1. Sequence Visualization

Sampled data from our collected datasets are shown in Fig. 18. Corresponding text annotations for actions are provided under the caption.

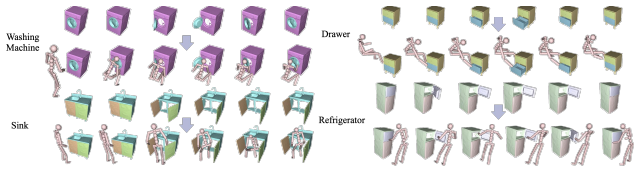

Figure 16. Synthesized body motions conditioned by sequences of object state

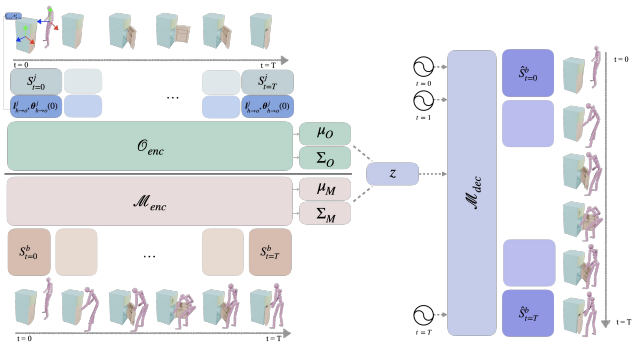

Figure 17. Model Architecture estimating human motion based on object states

Object	body repr.	window	MPE↓(cm) rc-joints	rc-wrists	glb-root	glb-joints	MOE↓	Foot skating↓
Refrigerator	root centric	30	0.58	1.36	0.68	1.15	0.22	0.98
		60	0.76	1.73	2.47	3.22	0.43	1.05
		90	0.97	2.09	5.60	6.57	0.75	1.74
	SMPL-X	30	1.67	2.69	0.68	2.32	0.12	1.58
		60	1.93	2.94	1.83	3.95	0.24	1.50
		90	3.49	5.26	5.32	8.87	0.53	2.67
Drawer	root centric	30	1.51	2.10	0.81	1.86	0.13	1.04
		60	1.52	1.74	1.02	2.26	0.24	0.80
		90	2.61	2.90	3.92	5.88	0.66	1.58
	SMPL-X	30	3.43	4.04	0.40	3.87	0.11	1.38
		60	3.27	3.83	0.96	4.17	0.12	1.04
		90	2.78	3.39	2.73	5.20	0.18	1.10
Washing Machine	root centric	30	0.50	0.89	0.35	0.77	0.15	0.94
		60	0.55	1.06	1.44	2.14	0.48	1.35
		90	1.29	2.62	9.03	11.46	0.92	2.60
	SMPL-X	30	1.94	2.53	0.42	2.24	0.09	1.04
		60	3.68	4.29	2.53	6.59	0.44	2.72
		90	6.03	7.55	6.01	13.29	0.60	2.68
Sink	root centric	30	0.58	0.98	0.42	0.82	0.16	0.79
		60	0.61	0.94	1.02	1.47	0.29	0.81
		90	1.21	1.93	3.83	4.84	0.76	1.67
	SMPL-X	30	2.27	2.91	0.52	2.64	0.10	0.90
		60	2.64	3.12	1.26	3.79	0.18	0.87
		90	2.60	3.12	2.16	4.80	0.27	1.01

Table 9. Quantitative results of the ParaHome task.

E. Experiments

E.1. Synthesizing Body Motion for Desired Object Manipulation

Train Details: The goal of our model shown in Fig. 17 is to synthesize a plausible 3D human motion conditioned with sequences of object state at a range of times. We represent target object status at each time t as $\mathbf{S}_{to}(t) = \{\phi^j(t)\}$ using joints state $\phi^j(t) \in \mathbb{R}^2$. We represent body pose at time t as $\mathbf{S}_p(t) = \{X_t, \Delta p, \Delta r\}$ using body pose X_t, root's linear velocity Δp and angular velocityΔr. We test with two types of body pose representation: the person root-centered skeleton representation [45], and the SMPL-X [78] body pose. In training, we use AdamW optimizer and LR=$1e-4$ with 1500 epochs and batch size 32.

Additional Results: We train our baseline model for four objects including a refrigerator, drawer, washing machine, and

sink with window sizes 30, 60, and 90. The quantification results are shown in Table 9. As shown in the result table, as the window size decreased, the accuracy increased in most items and root-centered body skeleton representation results in better accuracy for pose-dependent attributes (rc-joints, rc-wrists, glb-joints) but SMPL-X notation results in better global orientation and root position. Additional examples of visualization are shown in Fig. 16 and our supplementary video.

Figure 18. Example scenes of ParaHome dataset and aligned text annotation (Column1) Move kettle and cup from desk to the sink. (Column2) Take laptop from the desk and move to the table. (Column3) Take pan from the gas stove to the table. (Column4) Put laundry in the washing machine. (Column 5) Throw away STH into trash can.