MMHOI: Modeling Complex 3D Multi-Human Multi-Object Interactions

Kaen Kogashi
Mitsubishi Electric
Japan

Anoop Cherian
Mitsubishi Electric Research Labs
United States

Meng-Yu Jennifer Kuo
Nara Women's University
Japan

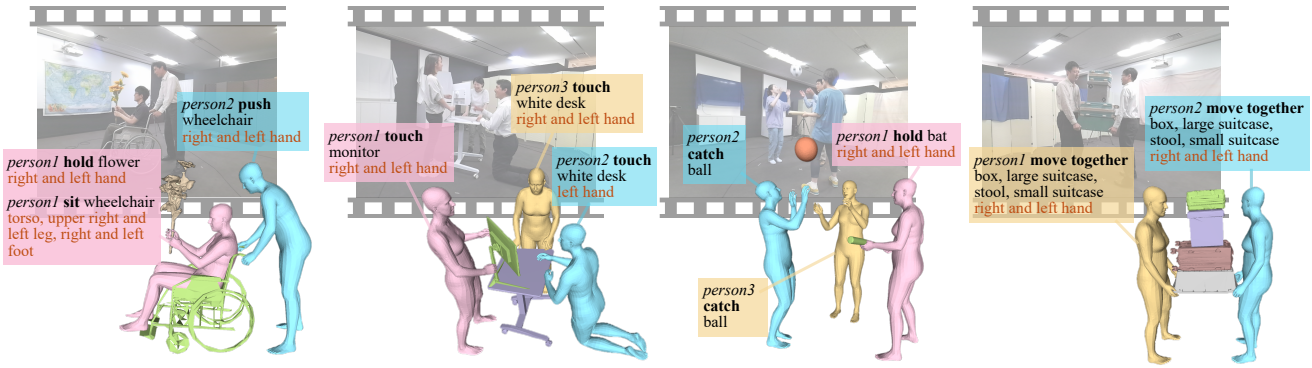

Figure 1. Example scenes from our MMHOI dataset – a new, large-scale dataset with high-quality annotations of multiple 3D humans, objects, actions, and interaction body parts, enabling holistic reasoning in complex interaction scenarios. Person IDs are shown in *italic*, action annotations in **bold**, object names in regular font, and interacting body parts are highlighted in orange text.

Abstract

Real-world scenes often feature multiple humans interacting with multiple objects in ways that are causal, goal-oriented, or cooperative. Yet existing 3D human-object interaction (HOI) benchmarks consider only a fraction of these complex interactions. To close this gap, we present MMHOI – a large-scale, Multi-human Multi-object Interaction dataset consisting of images from 12 everyday scenarios. MMHOI offers complete 3D shape and pose annotations for every person and object, along with labels for 78 action categories and 14 interaction-specific body parts, providing a comprehensive testbed for next-generation HOI research.

Building on MMHOI, we present MMHOI-Net, an end-to-end transformer-based neural network for jointly estimating human–object 3D geometries, their interactions, and associated actions. A key innovation in our framework is a structured dual-patch representation for modeling objects and their interactions, combined with action recognition to enhance the interaction prediction. Experiments on MMHOI and the recently proposed CORE4D datasets demonstrate that our approach achieves state-of-the-art performance in multi-HOI modeling, excelling in both accuracy and reconstruction quality. The MMHOI dataset is available at https://zenodo.org/records/17711786.

1. Introduction

Perceiving 3D human-object interactions (HOIs) is central to understanding human behavior and is crucial for applications like action recognition, 3D scene reconstruction, and human-centric AI. While progress has been made in modeling 3D interactions between a single human and a single object [2, 3, 5, 7, 8, 18, 31, 36], real-world scenarios often involve multiple humans interacting with multiple objects simultaneously. These settings introduce greater complexity due to increased interaction patterns, diverse object categories, and ambiguity in recognizing multi-entity actions. Addressing these challenges is essential for developing robust models that generalize to the dynamic, collaborative nature of real-world environments.

In response, recent efforts have begun exploring more complex interaction settings beyond single-human scenarios. One such effort is HOI-M^3 [37], which expands the scope to multiple humans and objects. However, it treats humans and objects as independent entities, overlooking the collaborative dynamics of real-world interactions. Recently, CORE4D [14] complements this by capturing 3D interactions with cooperative intent, but focuses only on two-person-single-object scenarios with limited action diversity. Moreover, the lack of varied action and detailed body part annotations limits these datasets' ability to provide high-level task-specific guidance for low-level interaction model-

ing, leaving the synergy between 3D reconstruction and action recognition in multi-HOI scenarios largely unexplored.

To address the above challenges, we introduce **MMHOI**, a novel large-scale dataset that provides comprehensive 3D annotations for multiple humans and objects, coupled with action labels that capture diverse real-world interactions. See Fig. 1, 2 for a few example scenes from our dataset. MMHOI offers a combination of spatial and semantic data, consisting of ∼600k frames captured across 12 daily-living scenarios, featuring 13 participants (7 males and 6 females) interacting with 22 commonly used objects. Each interaction is annotated with 78 distinct action classes, enabling the study of both cooperative and non-cooperative activities. Unlike previous datasets, MMHOI incorporates 14 interactive body part labels that reveal the specific body regions involved in object interactions, facilitating a deeper understanding of human-object interaction dynamics at both the spatial and semantic levels.

Building upon the MMHOI dataset, we propose a novel single-shot framework that simultaneously estimates the pose, shape, and 3D location of all humans and objects in camera space from a single RGB image. Human meshes are parameterized using the SMPL-X model [23], while object poses are represented with 6DoF parameters. Our framework leverages a Vision Transformer (ViT)-based architecture [1, 4] to extract patch-level features for both human and object modeling. To capture object-specific spatial and interaction cues, we introduce a structured dual-patch representation: each object is described by a main patch covering its primary region and a sub-patch encoding interaction-relevant features. This representation enables the model to reason about both object properties and fine-grained human-object interactions more effectively.

Another key aspect of our framework is the use of action recognition as a signal to enhance 3D reconstruction. Specifically, we use the features from the human and object perception modules in our transformer network to predict action labels and the interaction body parts. Next, body-part interaction consistency loss constraints are enforced between interactive body parts' ground truth classes and corresponding object regions, improving spatial alignment and preventing physically implausible configurations.

We present experiments demonstrating the benefits of our method in multi-human multi-object interactions modeling on our proposed MMHOI and the recent CORE4D datasets. We benchmark our model against prior approaches that are re-purposed to our task setting. Our results show that while standard methods falter in inferring cross-interactions, our method excels in modeling complex interactions, achieving state-of-the-art results and superior reconstruction of humans, objects, and their interactions. Our key contributions are summarized below.

- We introduce **MMHOI**, a novel large-scale dataset featuring detailed multi-HOI 3D annotations, action labels, and interactive body part labels. To the best of our knowledge, MMHOI is the only dataset comprising simultaneous multi-human and object-object interactions.
- We propose MMHOI-Net, a novel method for multi-human and multi-object interaction reconstruction in 3D, incorporating a structured dual-patch object representation and leveraging action guidance.
- The effectiveness of MMHOI-Net is demonstrated through extensive experiments on MMHOI, and also on the recent CORE4D [14] datasets.

2. Related Works

3D Single Human and Object Interaction. Several works that focusing on full-body interactions have been proposed to advance the HOI modeling [2, 3, 8, 18, 31, 36]. For instance, the BEHAVE dataset [2] benchmarks full-body human interactions with a single movable object, facilitating progress in modeling isolated HOI tasks [21, 34]. However, these datasets predominantly focus on interactions involving a single human and a single object, overlooking the more complex scenarios where multi-human interact with multi-object in cooperative environments.

3D Multiple Humans and Objects Interaction. Recently, the HOI-M^3 dataset [37] advanced the field by capturing interactions among multi-humans and multi-objects in shared contextual environments. However, it lacks scenarios in which multiple humans simultaneously interact with one or more objects, leaving a significant gap for the explicit studying of cooperative interactions. More recently, CORE4D [14] focuses on two-person interactions involving a single object, capturing aspects of cooperation however has limited action annotations and do not consider multi-object scenarios. The absence of detailed action and interactive body part annotations in both datasets limits their scope in modeling interaction consistency and physical plausibility. Instead, our MMHOI dataset captures rich, complex interactions involving co-existing multi-human and multi-object-object interactions, with diverse action labels.

Action Recognition and 3D HOI. Action recognition and 3D human pose estimation are deeply interconnected. Early works primarily relied on 2D keypoints for action recognition [12, 33], however recent research has demonstrated that 3D pose estimation provides rich geometric information and context for complex interaction reasoning [16]. Although sparse 3D body joint representations often fail to capture fine-grained details, detailed 3D poses [13, 24] are seen to improve action recognition and single HOI modeling. Leveraging these insights, we propose to use action recognition to help enforce consistency in multi-HOI 3D estimation, ensuring that the reconstructed body poses remain faithful to their interaction objects.

3. MMHOI Dataset

The MMHOI dataset is designed to capture multi-HOIs among cooperative interactions, providing high-quality data for advancing 3D reconstruction and action understanding. In Tab. 1, we show a comparison between MMHOI and prior datasets. Below we outline the key statistics, data acquisition, and processing pipeline **(see supp. for details)**.

3.1. Dataset Statistics

Our dataset captures realistic and diverse human-object interactions in 12 daily scenes, such as dining, collaborative work, and recreational activities (see Fig. 2). It comprises ∼600k frames featuring interactions performed by 13 participants (7 males and 6 females) with 22 commonly used objects spanning various categories. To complement the 3D and object annotations, MMHOI includes action annotations covering 78 distinct action classes and 14 interacting body parts. Notably, MMHOI is the first dataset to combine 3D annotations of multi-human and multi-object-object with detailed action annotations, capturing their interactions in diverse and realistic scenarios.[1]

3.2. Dataset Acquisition

The MMHOI data was captured using a multi-camera setup consisting of four Microsoft Azure Kinect sensors. Each sensor recorded scenes at a resolution of 2048 × 1536 pixels and 30 fps. Calibration across the four sensors was performed using AprilTag markers [30] to ensure precise spatial alignment. Although the Azure Kinect cameras provide built-in synchronization, occasional data drift due to hardware limitations necessitated manual adjustments to achieve accurate temporal alignment.

3.3. Data Processing

To achieve high-quality annotations, a rigorous pipeline was developed to process humans, objects, and actions within the scenes.

Human Annotations. Using synchronized and calibrated four-view videos, human segmentation was performed with the L-SAM [15]. We employ SMPL-X [23] as the 3D parametric human model to represent the statistical variations in human shape and pose. Following [1], the model parameters include pose parameters $\theta \in \mathbb{R}^{53\times3}$ and shape parameters $\beta \in \mathbb{R}^{10}$. From these parameters, SMPL instantiates a human body mesh $M \in R^{N\times3}$ with $N = 10,475$ vertices. To initialize 3D human body representations from each camera view, we employed Multi-HMR [1], a single-view model that estimates SMPL-X parameters independently for each viewpoint. Professional annotators reviewed the results, selecting the view with minimal occlusion for refinement. These models were further refined using Iterative Closest Point (ICP) alignment with depth data (recorded from Kinect) in 3D space, followed by manual adjustments with FreeCAD [26] to finalize their positions.

[1]Each subject in our dataset provided their written consent to be included.

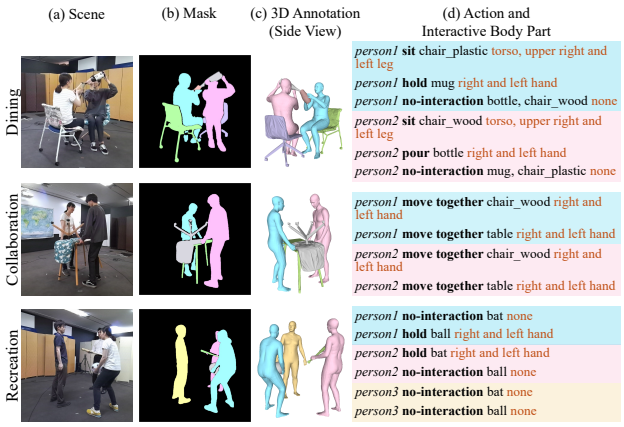

Figure 2. Overview of MMHOI. The dataset is categorized into three main interaction types: dining, collaborative work, and recreational activities, each type belongs to 12 scenarios. MMHOI consists of (a) RGB images, (b) segmentation masks, (c) 3D tracking of multiple humans and objects, and (d) action and interactive body part labels.

Object Annotations. Object segmentation was also performed using L-SAM [15]. All objects in MMHOI are prescanned with the FreeScan UE Pro2 to obtain high-fidelity CAD models. These are initially aligned to depth maps (recorded from Kinect) via ICP to estimate 6D poses, then manually refined by annotators for accurate placement. We set the center of four Kinects as the world origin, and the CAD model's center as the object's origin. The object's final pose (R, T) is computed via ICP between the original CAD template and the manually annotated placement.

Action Annotations. In addition to 3D annotations of humans and objects, MMHOI provides detailed action annotations to contextualize interactions. Each frame is labeled by trained annotators, who assign predefined action classes and specify which of the 14 body parts are involved in the interaction. The annotations are inspired by daily scenarios in video-based action datasets [6, 9, 19, 20, 22, 25, 27, 29, 32, 35, 39], and reflect the concept of *synergy*—how human actions and object affordances complement each other. Actions include both cooperative and individual tasks, such as "passing an object," "assembling," and "organizing." Annotators cross-referenced multi-view RGB-D recordings to ensure accurate alignment between actions and 3D spatial configurations. Action labels are formatted as time-aligned sequences, enabling seamless integration with 3D and semantic data.

Table 1. Comparisons between MMHOI and 3D HOI reconstruction datasets. MMHOI includes inter- and intra- human-and-object interactions, along with action and interaction body part annotations. We include the maximum # of humans per scene, the # of (conc)urrent interactions, and the # of interacting parts. * indicates captions are provided instead of action labels.

Datasets	Dyn. Obj.	# Conc. Interact.	# Interact. Parts	Interact. Action	max. # Humans	Obj-Obj Interact.
GRAB [31]	✓	✗	✗	✗	✗	✗
BEHAVE [2]	✓	✗	✗	✗	✗	✗
CHAIRS [8]	✓	✗	✗	✗	✗	✗
NeuralDome [36]	✓	✗	✗	✗	✗	✗
InterCap [7]	✓	✗	✗	✗	✗	✗
HIMO [17]	✓	✗	✗	✗	✗	✓
ParaHome [10]	✓	✗	✗	✓*	✗	✓
HOI-M^3 [37]	✓	✗	✗	✓*	$\geq$3	✗
CORE4D [14]	✓	1	✗	5	2	✗
MMHOI (Ours)	✓	4	14	21	3	✓

4. MMHOI-Net

This section introduces our MMHOI-Net, a vision-transformer based single-shot multi-task model for jointly learning the 3D human-object configurations, inferring the actions, and refining the body-part interactions. In Fig. 3, we illustrates the overall architecture of MMHOI-Net. The input to our model is a single RGB image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ of resolution $H \times W$ depicting a multi-HOIs scene and a mask of this scene produced using L-SAM [15]. Our model outputs the SMPL-X parameters of all individual humans in $\mathbf{I}$, all the object meshes in the scene, their corresponding root 3D locations $t \in \mathbb{R}^3$ in camera coordinates, and associated action classes between all humans and object pairs.

4.1. Vision Transformer-Based Feature Encoding

To effectively process the input image, we employ a Vision Transformer (ViT) backbone [4] to extract high-level image features. The image $\mathbf{I}$ is first subdivided into non-overlapping patches of size $P \times P$. Each patch is embedded into a feature token, and these tokens are processed with self-attention blocks into $\mathbf{E} \in \mathbb{R}^{H/P\times W/P\times D}$ with dimension D features. Similar to Multi-HMR [1], each output token maps spatially to a patch of the input RGB image. We employ the Human Perception Head (HPH) [1] for modeling the humans in the scene, and propose an Object Perception Head (OPH) for modeling the mesh parameters of the objects. Both these heads use the same ViT backbone, which is initialized with weights from a pre-trained MultiHMR model [1] during training. Below, we provide more details of HPH and our novel OPH. As depicted in Fig. 3, we further concatenate features from HPH and OPH for the action and body part classification.

4.2. Human Perception Head

Following Multi-HMR [1], this head identifies all major human keypoints and predicts their SMPL-X parameters and depth. For N humans in an image, a set of queries $\{\mathbf{q}_n\}_{n\in[N]}$ is initialized from the ViT feature tensor and processed in parallel through a stack of L blocks ($L = 2$ in the experiments), where alternating cross-attention and self-attention layers to progressively refine the queries. The final output, $\mathbf{Q}^L \in \mathbb{R}^{(D+D')\times N}$, is a set of N refined human features used to regress the final human parameters (see [1] for further details).

4.3. Object Perception Head

To jointly regress human and object geometries, we introduce an *Object Perception Head (OPH)* in our model (Fig. 3). Similar to the HPH branch for humans, OPH for objects follow a similar structure, however incorporates two patches instead of one – a main patch and a sub-patch – to model objects at multiple scales, which we call a *structured dual-patch representation*. Why do we need two patches? Recall that in [1], the human head is used as the primary keypoint to place the reconstructed human mesh. Objects in our dataset, however, vary in shape and symmetry, and scale. This is further exacerbated across categories, and thus treating them as monolithic entities is inadequate. To overcome this, we propose to use two patches overlapping the object mask to more robustly capture the object's location, pose, and interaction points.

Our dual patch representation is obtained as follows. Given an object mask detected using L-SAM [15], the main patch for the object is defined as a patch covering the largest region of the mask containing the object mask center. The sub-patch is selected from one of the 8 surrounding patches to the main patch, corresponding to the second-largest region of the object mask and containing human-object interaction masks, i.e., patches containing mask parts belonging to both humans and the object (see Fig. 4).[2]

In order for our model to be aware of the pose of the object, we propose to regress local patch offsets that implicitly guide the MMHOI-Net towards inferring the holistic object pose using the dual-patch representation. Let the main patch coordinates, *i.e.*, the object center (x_{main}, y_{main}), be defined as the center of the smallest bounding box encapsulating the L-SAM produced object mask. Suppose for a main patch at location (i, j), if (u^i_{main}, v^j_{main}) is the patch center, then we define the offsets $(\delta^i_{main_u}, \delta^j_{main_v}) := (x^i_{main} - u^i_{main}, y^j_{main} - v^j_{main})$. As noted by the red direction vectors in the main patches (yellow and purple in Fig. 4), this offset is pointed from the patch center to the object mask centroid within the main patch. The sub-

[2]If no such interaction patches are found, we use the second largest patch containing the object mask.

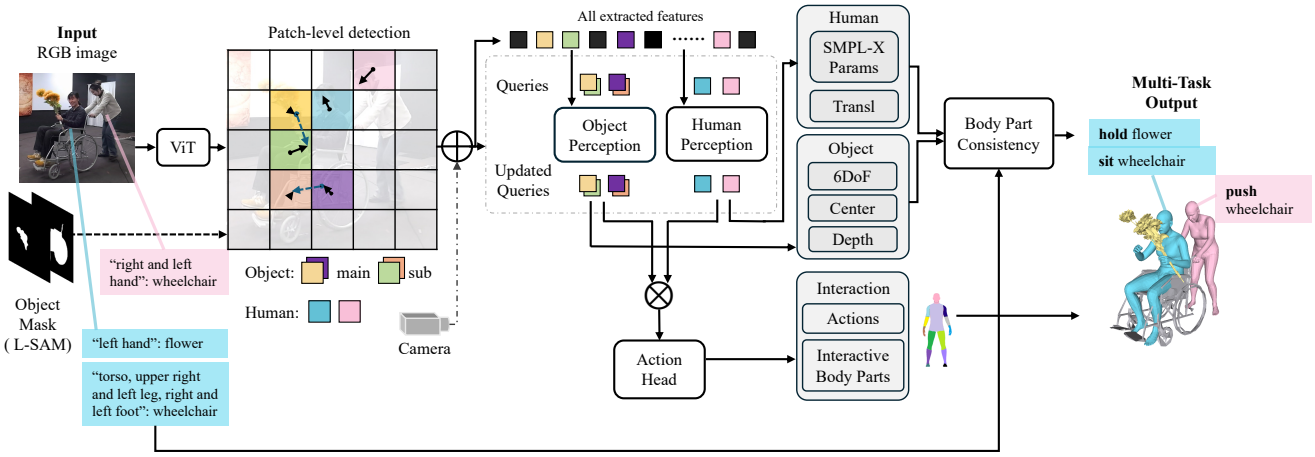

Figure 3. MMHOI-Net model architecture. Given a single RGB image, our model jointly estimates the 3D geometry of multiple humans and objects while incorporating action recognition as a supervisory signal. A ViT backbone extracts patch-level features. For human perception, detected keypoints serve as queries in a Human Perception Head [1], regressing SMPL-X pose, shape, and translation. Object perception head uses a structured dual-patch representation to regress object 6DoF, center, and depth. An action MLP predicts action and interaction body part classes.

patch center coordinates (x_{sub}, y_{sub}) are computed in a similar manner to the main patch (except that this point is the centroid of the object mask within the sub-patch) using the second largest mask region per patch (pink and green patches in Fig. 4). For the sub-patch center (u^i_{sub}, v^j_{sub}), the sub-patch offset coordinates $(\delta^i_{sub_u}, \delta^j_{sub_v})$ are computed as $(\delta^i_{sub_u}, \delta^j_{sub_v}) := (x^i_{sub} - u^i_{sub}, y^j_{sub} - v^j_{sub})$. As is clear, the sub-patch offsets point from the sub-patch center to the sub-patch mask center (depicted by the blue arrows in Fig. 4). The main and the sub-patch offsets jointly capture a ray pointed from the object center in the direction of the sub-patch, characterizing an approximate orientation of the object (black arrows). Thus, our key insight is that by having the network to regress these offsets during training (through a suitable loss that will be discussed shortly), it implicitly learns the pose of the object mesh.

Similar to the processing pipeline in the HPH head, given M objects in an image (detected by L-SAM), the final output of the transformer cross-attention layers is given by $\mathbf{Q}^{2\times L} \in \mathbb{R}^{2\times(D+D'\times M)}$ and viewed as a set of $2 \times M$ output object features. Next, the updated main and sub-patch queries from OPH are concatenated, and regressed using an MLP to produce the 6D object pose, center, and depth in 3D. To further enhance localization accuracy, available camera parameters are embedded into each patch as in [1].

4.4. Human-Object Interaction (Action) Head

We observe that action cues offer valuable semantic context, revealing the underlying structure of 3D human-object interactions. For instance, if an action module predicts "picking up a box", the corresponding body parts (*e.g.*, hands) must be touching the box. Inspired by prior works [21, 34] that use interactive body parts for 3D reconstruction in single-HOI, we extend this to multi-HOI by introducing an HOI head for action detection (see Fig. 3). To this end, we do a pairwise concatenation of the set of N human features output from HPH and a set of $2 \times M$ object features output from OPH, thus producing features $G = \{H_{main}\}_N \otimes \{(O_{main} \oplus O_{sub})\}_M$, where $\{H_{main}\}_N$ denotes the set of N main patch (head) output features from the HPH branch and $\{(O_{main} \oplus O_{sub})\}_M$ indicate the concatenation of the output object main O_{main} and sub-patch O_{sub} features from OPH, respectively. After this step, we obtain $|G| = N \times M$ human-object pairs from the direct product, which is passed to a 2-layer MLP to jointly predict both action and interactive body part classes.

4.5. Training Losses

Our model is optimized through a weighted combination of human and object reconstruction, interaction prediction, body-part detection, and 3D HOI reconstruction losses.

Human Reconstruction Loss. Similar to [1], the updated queries of the HPH branch are viewed as a set of N output features, which are used to regress the N human parameters using a shared MLP. The reconstruction loss aims to ensure accurate recovery of 3D human geometries in still images. The overall loss is defined as:

$$\mathcal{L}_{hum} = \lambda_h(\mathcal{L}_{hproj} + \mathcal{L}_{hmesh}) + \lambda_{param}\mathcal{L}_{param} + \lambda_{det}\mathcal{L}_{det},$$

where $\mathcal{L}_{hproj}$ and $\mathcal{L}_{hmesh}$ minimizes the re-projection error and output meshes for humans. $\mathcal{L}_{param}$ regularizes the SMPL pose, shape parameters, human depth, and human patch offset regression. The $\mathcal{L}_{det}$ captures the human detection loss. $\mathcal{L}_{hproj}$, $\mathcal{L}_{hmesh}$, and $\mathcal{L}_{param}$ minimize with

L_1 regression losses. $\mathcal{L}_{det}$ is minimized using binary cross-entropy loss. $\lambda_h, \lambda_{param}, \lambda_{hdet}$ are the loss weights. Refer to supp. materials for details.

Object Reconstruction Loss. The final output of the OPH branch are viewed as a set of $2\times M$ output features and is used to regress the M object parameters using a shared MLP. The object reconstruction loss aims to ensure accurate recovery of the 3D object geometries in still images. The overall loss is defined as:

$$\mathcal{L}_{obj} = \lambda_o(\mathcal{L}_{oproj}+\mathcal{L}_{omesh})+\lambda_p\mathcal{L}_p+\lambda_{main}\mathcal{L}_{main}+\lambda_{sub}\mathcal{L}_{sub},$$

where the object related loss $\mathcal{L}_p$ refers to inferring the 6D object pose, i.e., predicting the translation $T \in \mathbb{R}^3$, rotation $R \in \mathbb{R}^{3\times3}$, center $C \in \mathbb{R}^3$ (as in [34]) and depth (m) in 3D space. The input mesh to $\mathcal{L}_{omesh}$ input is computed by transforming the object mesh using the predicted R and T from the original CAD coordinates. $\mathcal{L}_{oproj}$ computes the 2D re-projection error of the 3D object mesh. $\mathcal{L}_p$ minimizes errors in object rotation, translation, center, and depth w.r.t. the ground truth. The main and sub-patches offset regression losses captured in $\mathcal{L}_{main}$ and $\mathcal{L}_{sub}$ minimize the predicted main and sub-patch offset coordinates $(\delta^i_{main_u}, \delta^j_{main_v})$ and $(\delta^i_{sub_u}, \delta^j_{sub_v})$ to the GT offset. All of $\mathcal{L}_{oproj}$, $\mathcal{L}_{omesh}$, $\mathcal{L}_p$, $\mathcal{L}_{main}$, and $\mathcal{L}_{sub}$ are minimized with L_1 losses. $\lambda_o, \lambda_p, \lambda_{main}$, and λ_{sub} are the loss weights.

Interaction Loss. The interaction loss enforces semantic and geometric consistency in human-object interactions by refining 3D reconstruction based on action and interactive body part detections. It is defined as:

$$\mathcal{L}_{interact} = \lambda_{act}\mathcal{L}_{act} + \lambda_{bp}\mathcal{L}_{bp} + \lambda_{cons}\mathcal{L}_{cons}.$$

Inputs to $\mathcal{L}_{act}$ and $\mathcal{L}_{bp}$ losses are the predicted action and body part classes from the HOI (action) head. Input to the $\mathcal{L}_{cons}$ is the human body part and object 3D meshes from HPH and OPH, respectively, as well as the body part GT labels. We use the cross-entropy losses in $\mathcal{L}_{act}$ and $\mathcal{L}_{bp}$ for action and interactive body part classification, respectively. The consistency loss $\mathcal{L}_{cons}$ aligns predicted interactive body parts with their corresponding object regions to prevent physically implausible configurations. Specifically, to enforce alignment, $\mathcal{L}_{cons}$ minimizes the distance between the ground-truth interactive body part points set $\mathcal{P}^i_{\text{bp}}$, where $i \in \mathcal{I} \subseteq \{1, \ldots, N\}$ indexes body parts and the predicted object points set $\mathcal{P}_{\text{o}}$:

$$\mathcal{L}_{\text{cons}} = \sum_{i\in\mathcal{I}} \max\left(\Psi(\mathcal{P}^i_{\text{bp}}, \mathcal{P}_{\text{o}}) - \delta,\ 0\right),$$

where $\Psi(\mathcal{P}^i_{\text{bp}}, \mathcal{P}_{\text{o}})$ is the Chamfer Distance between $\mathcal{P}^i_{\text{bp}}$ and $\mathcal{P}_{\text{o}}$, with a threshold $\delta = 5\text{mm}$. If Ψ exceeds δ, the loss encourages proximity. $\lambda_{act}, \lambda_{bp}, \lambda_{cons}$ are the loss weights. Our final training objective is as follows:

$$\mathcal{L} = \mathcal{L}_{hum} + \mathcal{L}_{obj} + \mathcal{L}_{interact}.$$

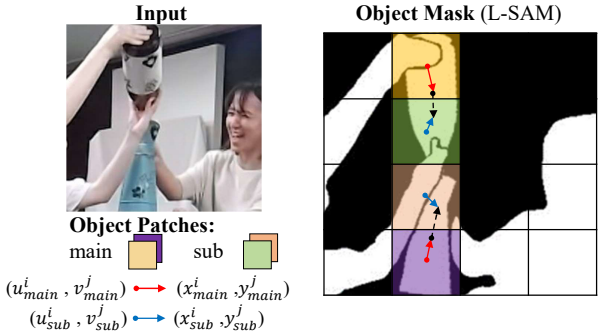

Figure 4. Our structured dual-patch object representation for inferring the object mesh parameters. The black arrows indicate approximate orientations of the objects.

5. Experiments and Results

Evaluation Datasets. We present experiments using our proposed MMHOI dataset as well as the CORE4D dataset. For MMHOI, we use 16.2k images for training and 4.1k for testing. In order to evaluate the generalization of our approach to other datasets, we provide experiments using the recently introduced CORE4D dataset, that proposes the task of collaborative human-object motion forecasting and interaction synthesis in cases where two humans interact with a single object. We present comparisons on CORE4D using repurposed prior approaches. We used 482 of their sequences for training and remaining for testing our approach.

Evaluation Metrics. We evaluate our framework on 3D reconstruction, action recognition, and interactive body part detection. For 3D reconstruction, we compute Chamfer Distance (CD) and Vertex-to-Vertex (V2V) distance to measure the accuracy of human and object mesh predictions. For action recognition and interactive body part detection, we report the average classification accuracy across all action and body part classes. For fair comparisons in 3D reconstruction, we apply Procrustes alignment before computing these metrics. In our evaluation, the 'S'-variant applies Procrustes alignment independently to *each* human-object pair, leading to large errors due to mismatches in multi-HOI scenes. In contrast, the 'M'-variant uses global alignment, yielding more accurate spatial assignment of humans to their interacting objects.

Implementation Details. Our framework is trained on the MMHOI dataset using a ViT-L backbone, with pre-trained weights from Multi-HMR [1] for the Human Perception Head. The training process takes three days on a single NVIDIA RTX A6000 (48GB) GPU. While MMHOI provides ground truth annotations in both SMPL-H and SMPL-X formats, we use SMPL-X in our experiments.

State-of-the-art Comparisons. We compare our approach against state-of-the-art single-HOI reconstruction

Table 2. State-of-the-art comparisons on MMHOI and CORE4D (S1) test sets. Chamfer Distance (CD) and Vertex-to-Vertex (V2V) distance are reported in cm for both human and object reconstructions. By "Action", we denote results obtained by integrating the method with actions. By "S"/"M", we mean to apply Procrustes alignment at the single/multi HOI level, i.e., aligning the predicted SMPL-X human meshes and object meshes at a single-HOI level or all at once, respectively.

Dataset	Method	S Hum CD ↓	S Obj CD ↓	M Hum CD ↓	M Hum V2V ↓	M Obj CD ↓	M Obj V2V ↓
MMHOI	Multi-HMR [1]+PHOSA[38]	7.50	82.07	-	-	-	-
	Multi-HMR [1]+CHORE[34]	7.41	79.46	-	-	-	-
	Multi-HMR [1]+CONTHO[21]	7.40	78.06	-	-	-	-
	Multi-HMR [1]+PHOSA[38] + Action	7.25	73.48	9.41	28.81	27.97	56.64
	Multi-HMR [1]+CHORE[34]+ Action	7.15	69.94	9.35	28.73	26.87	53.74
	Multi-HMR [1]+CONTHO[21]+ Action	7.03	65.42	9.33	28.71	26.64	53.63
	Multi-HMR [1]	-	-	9.30	28.74	-	-
	MMHOI-Net (ours)	**6.57**	**55.06**	**6.47**	**23.18**	**21.02**	**49.85**
CORE4D (S1)	Multi-HMR [1]+PHOSA[38] + Action	21.87	173.64	29.74	83.69	81.70	140.59
	Multi-HMR [1]+CHORE[34]+ Action	20.32	161.97	28.74	81.85	79.32	134.62
	Multi-HMR [1]+CONTHO[21]+ Action	19.71	153.18	27.15	79.78	77.26	126.63
	Multi-HMR [1]	-	-	26.85	76.74	-	-
	MMHOI-Net (ours)	**18.40**	**144.17**	**18.11**	**57.95**	**58.82**	**119.61**

Table 3. Ablation studies assessing the component contributions in reconstruction and action recognition.

Reconstruction	M Hum CD ↓	M Obj CD ↓
HPH only	9.35	30.25
OPH (main patch only)	8.92	26.75
OPH (full: main + sub-patches)	8.10	23.74
OPH (full) + $\mathcal{L}_{act}$	7.31	22.97
OPH (full) + $\mathcal{L}_{act}$ + $\mathcal{L}_{bp}$	7.29	22.81
OPH (full) + $\mathcal{L}_{act}$ + $\mathcal{L}_{bp}$ + $\mathcal{L}_{cons}$	**6.47**	**21.02**

Action Recognition	Accuracy (%) ↑
w/o body part detection	59.99
w/ body part detection	**60.96**

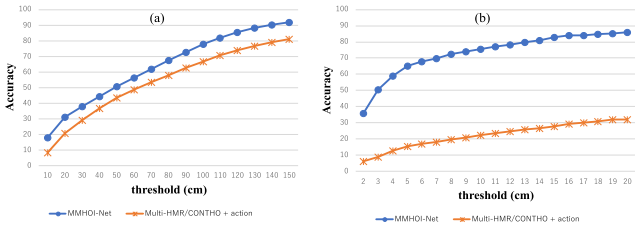

Figure 5. Evaluation of interaction prediction on MMHOI. (a) plots the % of scenes where the predicted interaction body parts are close to the objects within a threshold (x-axis, in cm). (b) shows object-object interaction prediction comparison. Both plots evaluate multi-HOI accuracies after Procrustes alignment. The plots highlight the benefit of explicitly modeling multi-HOIs.

methods, including PHOSA [38], CHORE [34], and CONTHO [21].[3] Since MMHOI focuses on multi-HOI scenarios where hand-object interactions are critical, we adopt SMPL-X-based Multi-HMR [1] for improved representation. To ensure a fair comparison, we adapt PHOSA, CHORE, and CONTHO to work with Multi-HMR outputs by treating multi-HOI as independent single-HOI instances and aggregating their results. Specifically, human meshes are estimated using Multi-HMR's SMPL-X outputs, while object poses are inferred separately for each individual in multi-human scenes. Since these methods are designed for single-HOI, we compute object meshes for shared interactions and obtain the final pose by averaging their predicted rotations [28] and translations. Table 2 shows that our method consistently outperforms these baselines, demonstrating improved multi-HOI reconstruction accuracy. Specifically, for Multi-HMR/PHOSA, Multi-HMR/CHORE, and Multi-HMR/CONTHO, we apply pairwise Procrustes alignment between humans and objects (1–3 rows in Table 2), as these methods were originally designed for single-HOI scenarios. We also evaluate these methods with action supervision by integrating our action head (Sec. 4.4), which incorporates multi-HOIs information. In this case, we perform evaluations on both single-HOI and multi-HOI settings (4–6 rows in Table 2). Although action supervision enhances the performance of past methods, our approach consistently surpasses them.

In the lower part of Table 2, we compare MMHOI-Net with recent reconstruction methods on the CORE4D [14] seen-object (S1) test set, where our method achieves superior performance. We also assess interaction accuracy by measuring the alignment between predicted interactive body parts and object regions in Fig.5(a), and show object-object interaction results in Fig.5(b). These findings high-

[3] We do not include PICO [3] in our evaluation as it heavily relies on its own PICO-db dataset, which does not share object categories with ours.

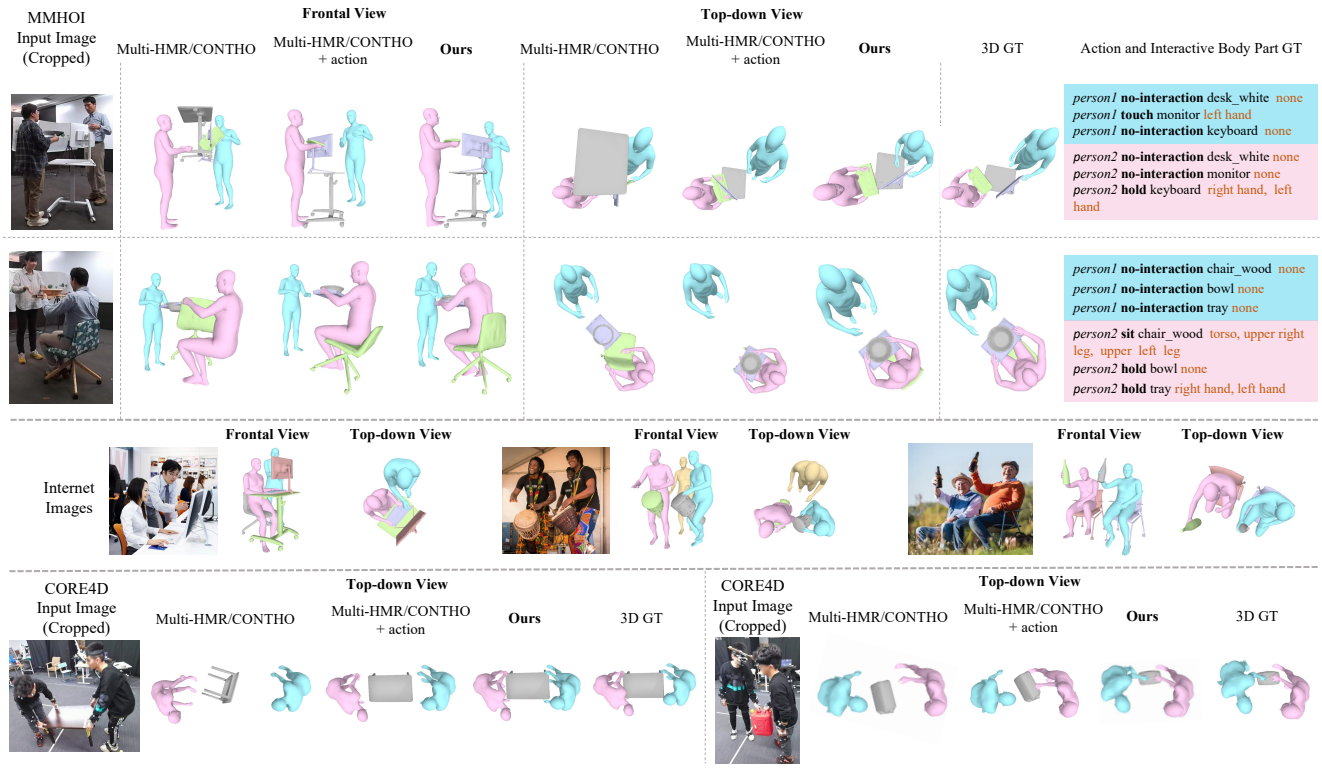

Figure 6. Qualitative comparisons of our method against Multi-HMR/CONTHO without and with action supervision on MMHOI and CORE4D (S1) test sets. Our approach shows better quality in human-object interaction prediction, occlusion handling, and spatial consistency. The third row shows the zero-shot inference on Internet images demonstrating the generalization capabilities of our model.

light the limitations of single-HOI methods, which fail to capture multi-HOIs dependencies, leading to suboptimal reconstructions in dense human-object interaction scenarios. We do not include HOI-M^3 [37] in our evaluation as their full dataset and pre-trained models are unavailable.

Qualitative Evaluations. Fig. 6 presents qualitative comparisons between our method and Multi-HMR/CONTHO (with and without action supervision) across various object types, occlusions, and challenging viewpoints on MMHOI and CORE4D datasets. Our approach improves interaction reasoning and better preserves object affordance constraints, resulting in more coherent and semantically meaningful reconstructions. Additionally, we demonstrate the generalizability of our method by evaluating it on internet images without additional fine-tuning.

Ablation Studies. Table 3 shows that our full model performs best, while removing key components degrades accuracy, highlighting their importance in multi-HOI modeling. Specifically, we see that our dual-patch object representation leads to a 3 point drop in multi-object chamfer distance against a 'main patch only' representation. We also analyze the changes in action recognition performance when the body-parts are detected with +1% improvement. See [11] for more details, including results showing generalization to new objects.

6. Conclusions

We introduced MMHOI, the first large-scale dataset capturing the complexity of real-world multi-human and multi-object-object interactions, complete with 3D annotations, action labels, and interaction body parts. Building on MMHOI, we proposed MMHOI-Net, a transformer framework that reconstructs human-object geometries while reasoning about their interactions and actions. Experiments show MMHOI-Net outperforms prior methods in reconstruction accuracy and interaction consistency on MMHOI and CORE4D datasets.

Yet, challenges remain – severe occlusions disrupt predictions. Also, while we consider cooperative tasks, extending to full group interactions will require obtaining 3D ground truth of all interacting individuals, which we plan to explore in future work. Albeit these shortcomings, we believe our work paves the way for future research on 3D multi-HOI, providing a foundation for advancing HOI understanding. See the supplementary materials and [11] for dataset samples, training details, and qualitative results.

References

[1] Fabien Baradel*, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas*. Multi-HMR: Multi-person whole-body human mesh recovery in a single shot. In *ECCV*, 2024. 2, 3, 4, 5, 6, 7

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 1, 2, 4

[3] Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun Lakshmipathy, Agniv Chatterjee, Michael J. Black, and Dimitrios Tzionas. PICO: Reconstructing 3D people in contact with objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2, 7

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 4

[5] 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. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[6] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. pages 6047–6056, 2018. 3

[7] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction from multi-view RGB-D images. *International Journal of Computer Vision (IJCV)*, 2024. 1, 4

[8] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *ICCV*, 2023. 1, 2, 4

[9] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3

[10] Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions, 2024. 4

[11] Kaen Kogashi, Anoop Cherian, and Meng-Yu Jennifer Kuo. MMHOI: Modeling complex 3D multi-human multi-object interactions. arXiv:2510.07828, 2025. 8

[12] Y. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H. Fang, Y. Wang, and C. Lu. Transferable interactiveness knowledge for human-object interaction detection. pages 3580–3589, 2019. 2

[13] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020. 2

[14] Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, and Li Yi. Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1769–1782, 2025. 1, 2, 4, 7

[15] Medeiros Luca. Language segment-anything. *GitHub repository. https://github.com/luca-medeiros/lang-segment-anything (accessed Feb. 15, 2025.)*. 3, 4

[16] Diogo C Luvizon, David Picard, and Hedi Tabia. Multi-task deep learning for real-time 3d human pose estimation and action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2752–2764, 2020. 2

[17] Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, and Xiaokang Yang. Himo: A new benchmark for full-body human interacting with multiple objects, 2024. 4

[18] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. Unifying representations and large-scale whole-body motion databases for studying human motion. *IEEE Transactions on Robotics*, 32(4):796–809, 2016. 1, 2

[19] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. 2009. 3

[20] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8, 2019. 3

[21] Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3d human and object via contact-based refinement transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10218–10227, 2024. 2, 5, 7

[22] Alonso Patron, Marcin Marszalek, Andrew Zisserman, and Ian Reid. High five: Recognising human interactions in tv shows. In *BMVC*, pages 50.1–50.11, 2010. doi:10.5244/C.24.50. 3

[23] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 3

[24] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Christoph Feichtenhofer, and Jitendra Malik. On the benefits of 3d pose and tracking for human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 640–649, 2023. 2

[25] Elisa Ricci, Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulò, Narendra Ahuja, and Oswald Lanz. Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos. In *2015 ICCV*, pages 4660–4668, 2015. 3

[26] Juergen Riegel, Werner Mayer, and Yorik van Havre. Freecad. *Freecadspec2002. pdf*, 2016. 3

[27] M. S. Ryoo and J. K. Aggarwal. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA), 2010. 3

[28] coders scipy. Get the mean of the rotations. 7

[29] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. arXiv:2010.10864, 2020. 3

[30] T Song. Extrinsic calibration for multiple azure kinect cameras. *GitHub repository. https://github. com/stytim/k4a-calibration (accessed Oct. 7, 2022.)*. 3

[31] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 4

[32] Coert van Gemeren, Robby T. Tan, Ronald Poppe, and Remco C. Veltkamp. Dyadic interaction detection from pose and flow. In *European Conference on Computer Vision (ECCV)*, 2014. 3

[33] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He. Pose-aware multi-level feature network for human object interaction detection. pages 9468–9477, 2019. 2

[34] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision*, pages 125–145. Springer, 2022. 2, 5, 6, 7

[35] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012. 3

[36] 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, 4

[37] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-m3: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 516–526, 2024. 1, 2, 4, 8

[38] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020. 7

[39] Hang Zhao, Zhicheng Yan, Lorenzo Torresani, and Antonio Torralba. Hacs: Human action clips and segments dataset for recognition and temporal localization. arXiv:1712.09374, 2019. 3