

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

## Supplementary Materials

Kaen Kogashi  
Mitsubishi Electric  
Japan

Anoop Cherian  
Mitsubishi Electric Research Labs  
US.

Meng-Yu Jennifer Kuo  
Nara Women's University  
Japan

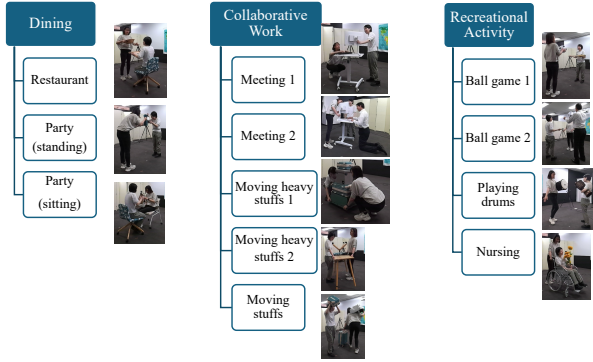


Figure 1. The MMHOI dataset is categorized into three main interaction types: dining, collaborative work, and recreational activities. These interaction types span 12 daily-life scenarios.

This supplementary material provides details of MMHOI dataset construction introduced in Section 3, details of the method in Section 4, additional results of MMHOI-Net in Section 5, also failure cases of our proposed method.

### 1. Details of MMHOI Dataset Construction

**Dataset Statistic.** As shown in Fig. 1, MMHOI dataset is categorized into three main interaction types – dining (3 scenarios), collaborative work (5 scenarios), and recreational activities (4 scenarios) – spanning 12 daily-living scenarios to ensure a balanced distribution of interactions. Fig. 2 and Fig. 14 present the statistical distributions of actors, objects, and interactions. Fig. 2(a)-(d) depict actor height, weight, object dimensions (length, width, height), sample distribution of interactive body part classes, while Fig. 2(e) illustrates the 14 predefined body parts. Fig. 14 shows the sample distribution of 78 action classes, with 13 rare classes containing fewer than 35 test samples.

**Action Annotation.** The action annotation interface is shown in Fig. 3, facilitates the labeling of human-object interactions within the dataset. Each action annotation is independently verified by at least two annotators to maintain

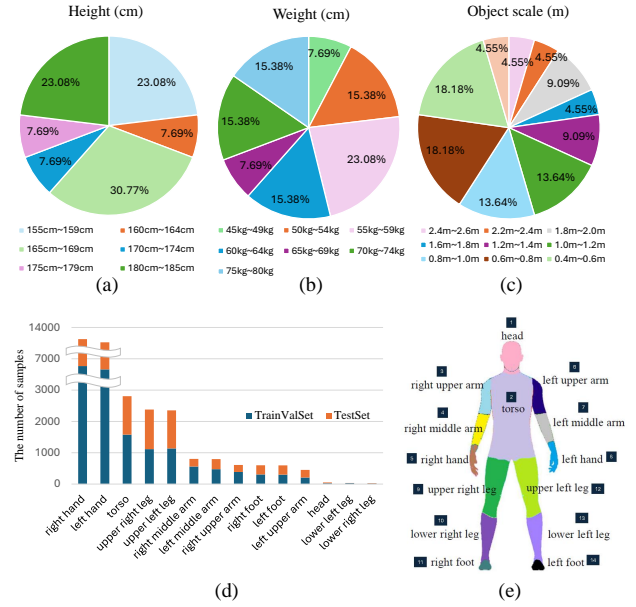


Figure 2. Statistics of humans and objects in the MMHOI dataset. (a) actor height, (b) actor weight, (c) object scales, (d) sample size distribution of interactive body part classes, and (e) 14 predefined body parts.

accuracy. Annotators are responsible for assigning appropriate verb labels to each human-object pair, capturing the nature of their interaction. The MMHOI dataset contains 78 human-object action types, with 21 unique action verbs. In contrast, CORE4D includes only 5 verb categories: “move (together)”, “raise (together)”, “rotate (together)”, “pass”, and “others”, making MMHOI significantly more diverse in terms of action representation.

**3D Annotation.** As shown in Fig. 4, 3D ground truth (GT) annotations are obtained via a two-step process. Each object’s location is initialized at the center of the four cameras. We first apply Iterative Closest Point (ICP) alignment with depth data for initial registration, followed by manual refinement in FreeCAD [4]. This refinement process

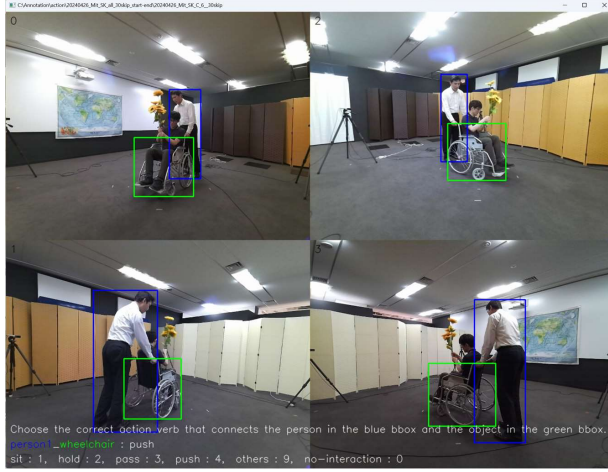


Figure 3. Action annotation UI screen capture. Annotator is asked to choose the correct action verb’s number from the presented candidate verbs.

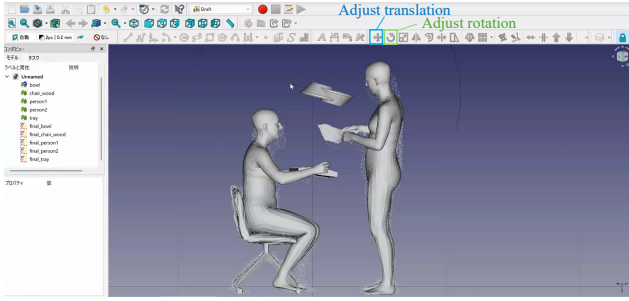


Figure 4. CAD annotation screen capture. CAD operator adjust rotation and translation to finalize the 3D GT.

involves adjusting human and object positions using translation and rotation tools (blue and green bounding boxes in Fig. 4). Each annotation then undergoes verification by three CAD operators before being finalized. The GT annotation records each object’s displacement and orientation relative to its initialized position.

## 2. Details of Experiments

**Implementation Details.** We initialize the MMHOI-Net weights using the multiHMR\_672\_S pre-trained model [1], which processes images at a resolution of  $672 \times 672$ . However, human meshes for baseline methods are inferred using the multiHMR\_896.L model from the Multi-HMR GitHub repository [1]. MMHOI-Net is trained on the MMHOI dataset with a batch size of 32 and a patch size of 24. During optimization, the loss weights are set as follows:  $\lambda_h = 1$ ,  $\lambda_{param} = 1$ ,  $\lambda_{det} = 1$ ,  $\lambda_o = 10$ ,  $\lambda_p = 10$ ,  $\lambda_{main} = 10$ ,  $\lambda_{sub} = 10$ ,  $\lambda_{act} = 100$ ,  $\lambda_{bp} = 10$ , and  $\lambda_{cons} = 100$ . Since



Figure 5. Patch mask region per object. The blue bounding box indicates the whole L-SAM detected region while red bounding box indicates mask region used for the main and sub patches offset.

we initialize human-related parameters using Multi-HMR [1], the corresponding loss weights ( $\lambda_h$ ,  $\lambda_{param}$ ,  $\lambda_{det}$ ) are set lower relative to the others. We train MMHOI-Net using automatic mixed precision [3] for 500k iterations. Following [1], we adopt SMPL-X with 10 shape components during both training and inference.

**Object Patch Selection and Mask Refinement.** To ensure precise object representation, we refine the L-SAM masks used for defining main and sub patches, particularly for large objects such as tables, chairs, monitors, and flowers. As shown in Fig. 5, we apply mask shrinking strategies to 7 out of the 22 object categories to focus on the most interaction-relevant regions. For chair objects, the mask is

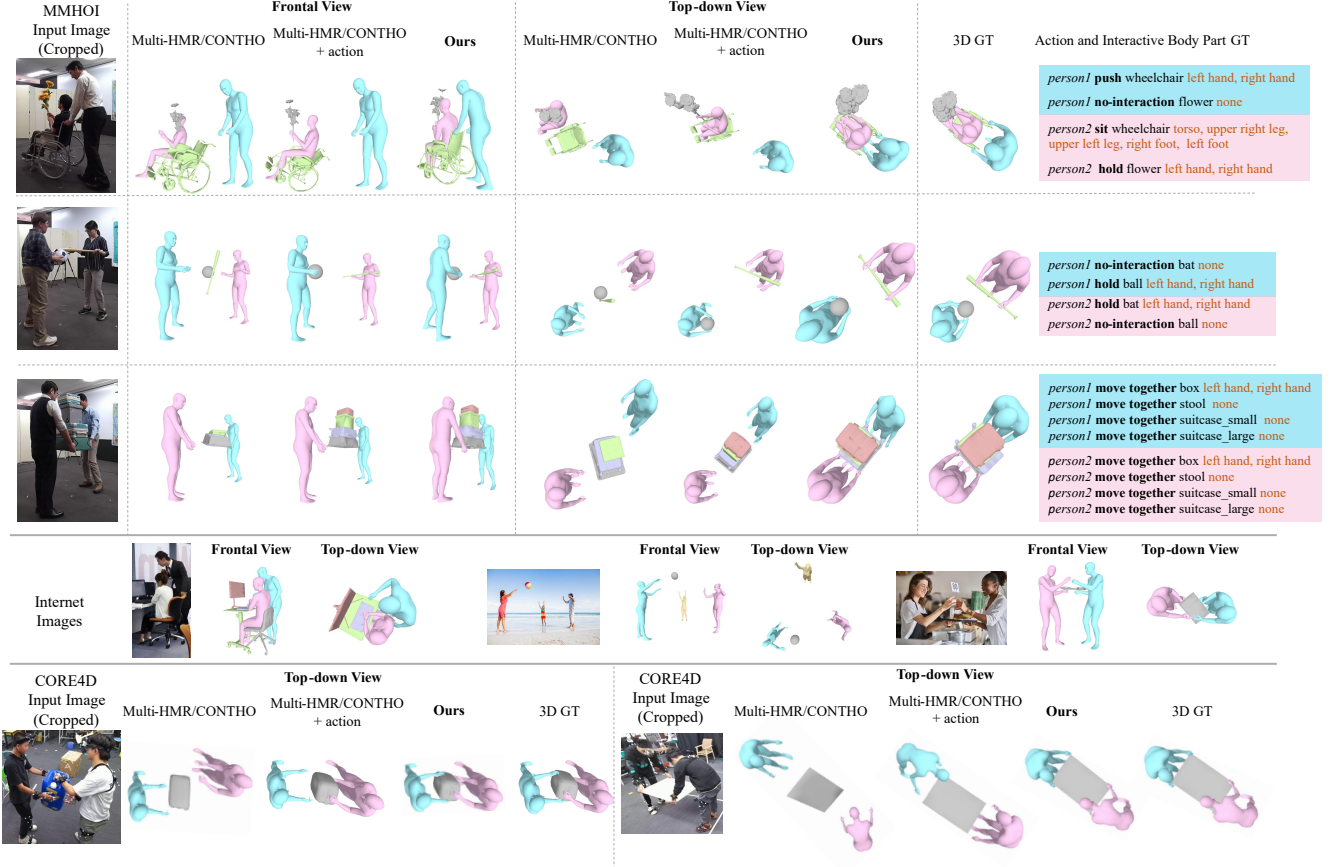


Figure 6. Additional results comparing MMHOI-Net with Multi-HMR/CONTHO, with and without action supervision, on MMHOI and CORE4D (S1) test sets. Our method outperforms prior approaches in inferring reasonable multi-HOI reconstructions.

restricted to the bounding box height range of 20%-55% from the top. Similarly, for tables, monitors, and flowers, the mask is limited to 0% – 15%, 0% – 60%, 40% – 100% of the bounding box height from the top, respectively.

**Effective Dataset Size** Our dataset has 4–14 unique action classes per scenario. Further, our training and test sets are balanced across 12 scenarios, each with  $\sim 50K$  frames. During training, we used a validation set to avoid overfitting.

### 3. Additional Results and Details

**Additional Visual Examples** Fig. 6 presents additional qualitative comparisons between our method and Multi-HMR/CONTHO, including results with and without action supervision, as well as in-the-wild internet images on MMHOI and CORE4D (S1) test sets. Fig. 7 shows samples of comparing our method with Multi-HMR/PHOSA, Multi-HMR/CHORE and Multi-HMR/CONTHO (with and without action supervision) on MMHOI and CORE4D (S1) test

sets. Our model, trained on MMHOI, demonstrates superior reconstruction of multi-human-object interactions (multi-HOIs) compared to prior methods. Notably, our approach achieves more plausible and coherent reconstructions, effectively capturing human-object affordances and reducing physically implausible configurations.

**Additional Ablation Studies.** Fig. 8 demonstrates the impact of our dual-patch representation approach and losses, showing its effectiveness in enhancing 3D reconstruction accuracy and enforcing HOI consistency. Table 1 reports action recognition and interactive body part detection performance across various human-object interactions on MMHOI and CORE4D (S1). CORE4D does not provide body part annotations, so we generate them as follows. For each body part, we compute the Chamfer Distance  $\Psi$  between its point set  $\mathcal{P}_{bp}^i$  and the corresponding object point set  $\mathcal{P}_o$ . If  $\Psi < \delta$  (with threshold  $\delta = 5mm$ ), we consider the body part to be interacting with the object; otherwise, no interaction is assigned.



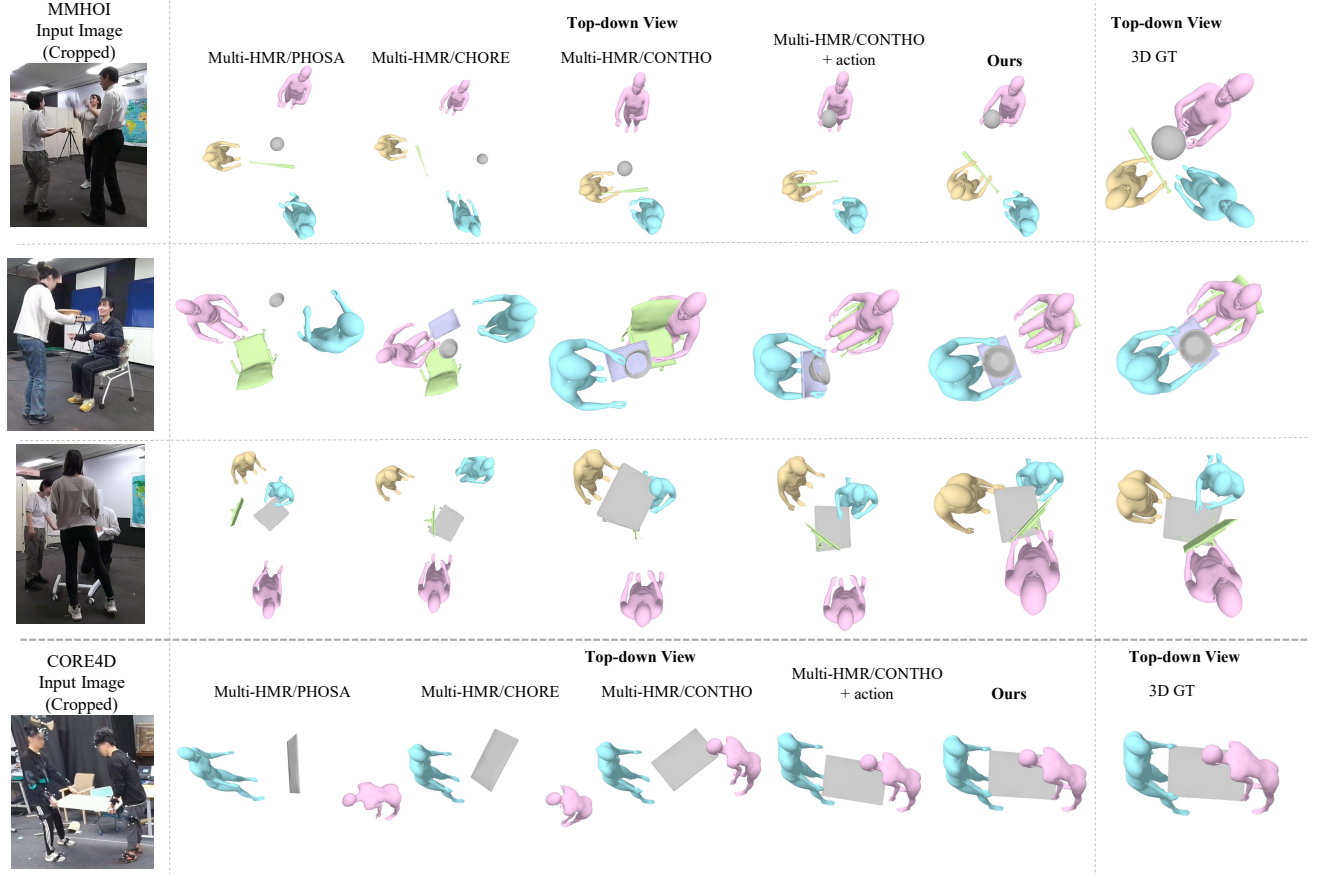


Figure 7. Additional results comparing MMHOI-Net with Multi-HMR/CONTHO (with and without action supervision), Multi-HMR/CHORE and Multi-HMR/PHOSA on MMHOI and CORE4D (S1) test sets. Our method outperforms past methods in multi-HOI 3D reconstructions.

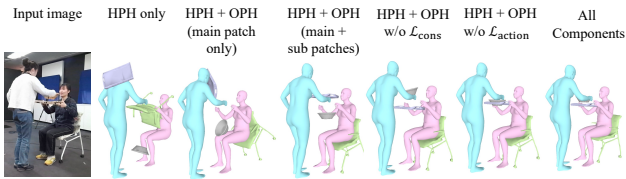


Figure 8. Additional ablation studies. Our proposed dual-patch representation approach and losses improve reconstruction accuracy and interaction consistency.

**Generalization to New Objects** To analyze generalization of our method to new objects, we re-trained our MMHOI-Net without 5 object categories (plastic chair, mug, drum brown, backpack, small suitcase) on MMHOI dataset and tested it to predict the poses of these objects (Fig. 10). We also evaluate MMHOI-Net on CORE4D’s unseen object test set (S2). CORE4D provides two test sets: seen objects (S1) and unseen objects (S2). Results for S1 are presented in Table 2 and Figure 6 of the main paper. As

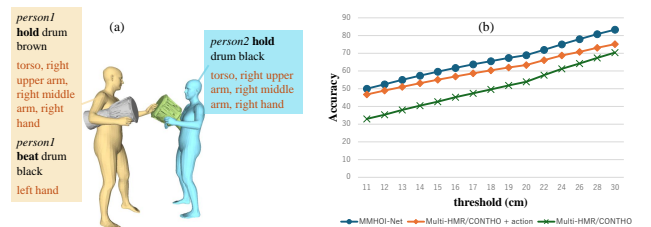


Figure 9. (a) Example to illustrate the evaluation of interactions described in Fig. 5(a) of the main paper, and (b) of this figure. (b) shows evaluation of Single-HOI after Procrustes alignment.

shown in Tab. 2, our method outperforms prior approach on the S2 split. We believe the action cues allow our model to generalize and predict plausible HOI dynamics for unseen objects (of the same affordance).



Table 1. Action recognition and interactive body part detection accuracy (%) on MMHOI-Net. For MMHOI dataset, by “Rare”, we mean 13 action classes that have fewer than 35 test samples, while “Non-Rare” includes the remaining action categories. “All” represents the overall accuracy across all action categories on MMHOI and CORE4D (S1) test set.

Dataset	Rare	Non-Rare	All	Body Part
MMHOI	12.1	70.3	61.0	67.0
CORE4D (S1)	-	-	76.5	88.7

#### 4. Details of Interaction Evaluation (Fig. 5 main paper)

In this section, we detail the evaluation scheme used to compute the accuracy metrics presented in Fig. 5(a) and 5(b) of the main paper. For clarity, we illustrate the process using the setting shown in Fig. 9(a). Specifically, we evaluate interaction prediction in the Single HOI setting (Fig. 9(b)), Multi-HOI setting (Fig. 5(a) main paper), and object-object interactions (Fig. 5(b) main paper).

**Single Body Part-Object Interaction Evaluation:** We compute the Chamfer Distance between the SMPL body part and object meshes after Procrustes alignment. In this case, our goal is to predict the alignment of our prediction (source) with the transformed target (GT after Procrustes alignment) for each potential interaction separately. To this end, we compute the individual pairwise interactions, compute the alignment with GT, and average over the matches. For example, with respect to Fig. 9 (a), as seen there is interaction between the SMPL model of the yellow person and the drum brown. We will consider predicted regions for the torso, right upper arm, right middle arm, and right hand for the yellow person and compute their distances to the drum brown mesh, and check if each of these distance are less than the given threshold against the corresponding aligned GT meshes. We repeat this for all the interaction regions for the entire scene and average over them to compute the accuracy at a given threshold. We vary this threshold to produce the left plot in Fig. 9 (b).

**Multiple Body Part-Object Interaction Evaluation:** We compute the result in (a) of Fig. 5 as follows. We compute all points of contact body parts between a person and an object, and instead of averaging over each HOI separately, we compute the entire set of interactions as a whole, by adding up the misalignments with the GT meshes. If the total misalignment is more than a given threshold, we assume the interaction prediction is a failure. We repeat this process for varied thresholds to produce the plot in Fig. 5 (a). With respect to the Fig. 9 (a), in this case, we assume

Table 2. Generalization performance using 5 unseen objects on MMHOI test subset, unseen objects test set on CORE4D (S2).

Dataset	Method	S Human Chamfer ↓	S Object Chamfer ↓
MMHOI	Multi-HMR + CONTHO MMHOI-Net	7.47 <b>6.59</b>	87.16 <b>64.51</b>
CORE4D (S2)	Multi-HMR + CONTHO MMHOI-Net	23.93 <b>19.31</b>	198.49 <b>151.82</b>

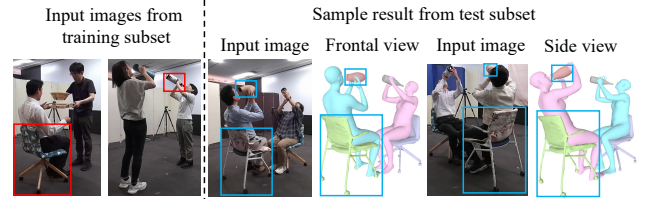


Figure 10. Generalization results to objects unseen during training on MMHOI dataset. Red and blue boxes indicate representative seen and unseen objects respectively.

the model predicts all contact body part points of the person in yellow (torso, right upper arm, right middle arm, right hand, and left hand). We also compute all the contact body part points of the person in blue (torso, right upper arm, right middle arm, right hand). We next compute the objects in contact with each person (drum brown and drum black). We compute the Procrustes alignment of the GT points with all the objects to compute the chamfer distances, which are added together to find the full error.

**Object-Object Interaction Evaluation:** Fig. 5(b) in the main paper presents performance on object-object interaction prediction. To evaluate this, we extract a test subset from MMHOI containing all samples with object-object interactions. We compute the result in Fig. 5 (b) as follows. We apply Procrustes alignment for global matching in Multi-HOI meshes before compute the chamfer distance. We compute the alignment of the GT points from the object-object pair to compute the chamfer distance. By adding up the misalignments with the GT object meshes, if the total misalignment is more than a given threshold, we assume the object-object interaction prediction is a failure. We repeat this process for varied thresholds to produce the plot in Fig. 5 (b). As shown in Fig. 5 (b), our MMHOI-Net outperforms a past method which highlighting our method implicitly capture multi-object-object poses from the novel object dual-patch representation.

#### 5. Limitations in Details.

**Single-view and Occlusions** Our model mitigates single-view ambiguity using our dual-patch representation and action cues. In Fig. 11 (a), we report RMSE in depth prediction (m) using pelvis as the root for human and the ob-

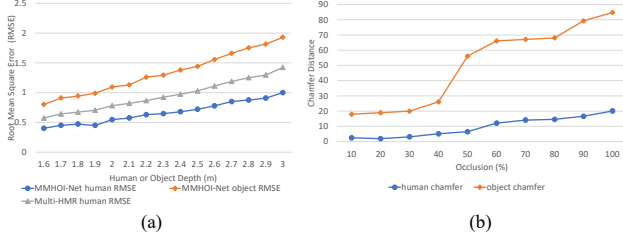


Figure 11. In (a), we report the RMSE (m) per depth on MMHOI-Net and Multi-HMR. In (b), we show Chamfer distance (cm) per occlusion rate (%) on MMHOI-Net.

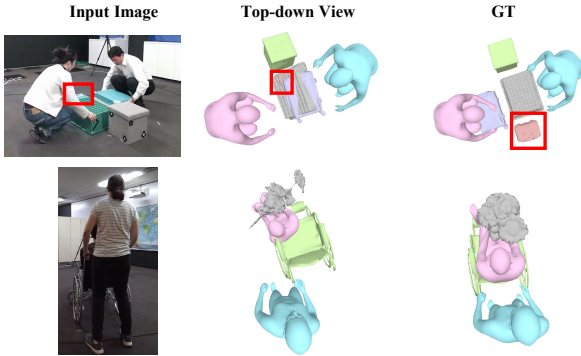


Figure 12. Failure cases of MMHOI-Net predictions. Severe occlusions and ambiguous “no-interaction” cases lead to inaccurate reconstructions.

ject center as root for the object. As past methods, e.g., CHORE, do not infer the absolute depth, we could compare ours only to Multi-HMR. Fig. 11 (b) shows our method’s effectiveness in handling partial occlusions. Severe occlusions of humans or objects remain challenging and often lead to inaccurate action predictions and reconstruction errors. Fig. 12 shows typical failure cases of MMHOI-Net. A potential improvement is leveraging temporal consistency [5] through multi-frame information.

**Group Interactions** Our  $N \times M$  pairwise interaction structure focuses on HOI pairs and may not capture the complexity of true group interactions. While we do consider cooperative tasks (e.g., “object-move together”), extending to full group interactions will require obtaining 3D ground truth of all interacting individuals, which we plan to explore in future work.

## 6. Analysis of CORE4D Dataset

In Tab. 2 of the main paper, Chamfer Distance and V2V errors on the CORE4D test set are higher than those on the MMHOI dataset for both prior methods and MMHOI-Net. This is primarily due to the lower quality of 3D ground truth in CORE4D [2]. As illustrated in Fig. 13 (left), the human

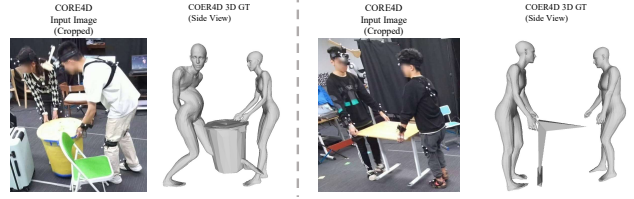


Figure 13. CORE4D dataset have quality problems in SMPL-X model (left sample) and object mesh (right sample).

mesh exhibits physically implausible SMPL-X configurations, while the object mesh (right) also appears unrealistic. Additionally, some sequences in CORE4D are not synchronized.

## 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
- [2] 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. 6
- [3] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018. 2
- [4] Juergen Riegel, Werner Mayer, and Yorik van Havre. Freecad. *Freecadspec2002. pdf*, 2016. 1
- [5] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

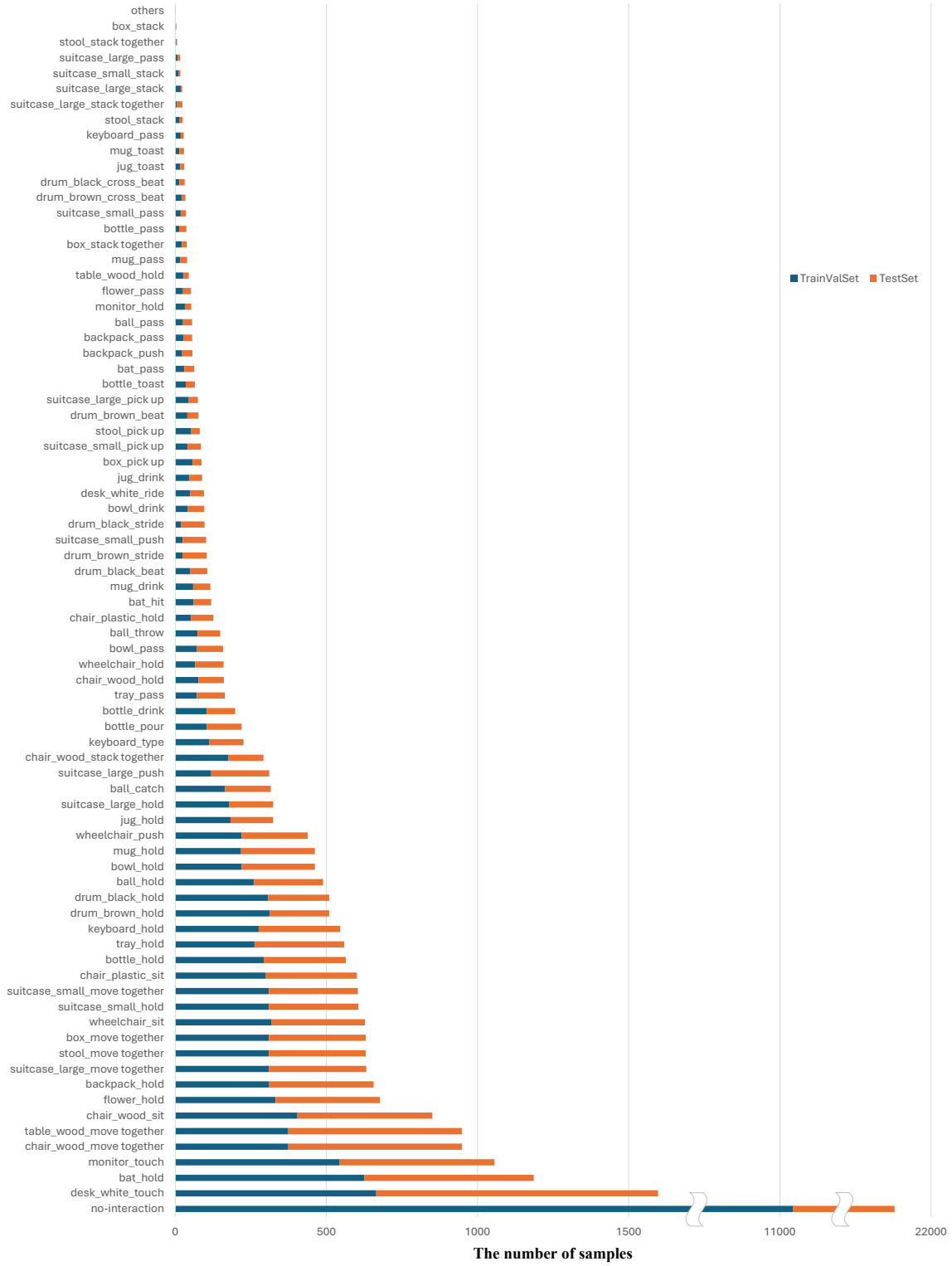


Figure 14. Sample distribution of action classes in the MMHOI dataset.