TACO: Benchmarking Generalizable Bimanual Tool-ACtion-Object Understanding

Yun Liu[1,2,3], Haolin Yang[4], Xu Si[1], Ling Liu[5], Zipeng Li[1], Yuxiang Zhang[1], Yebin Liu[1], Li Yi[†,1,2,3]
[1]Tsinghua University [2]Shanghai Artificial Intelligence Laboratory [3]Shanghai Qi Zhi Institute
[4]Beijing University of Posts and Telecommunications [5]Beijing Institute of Technology

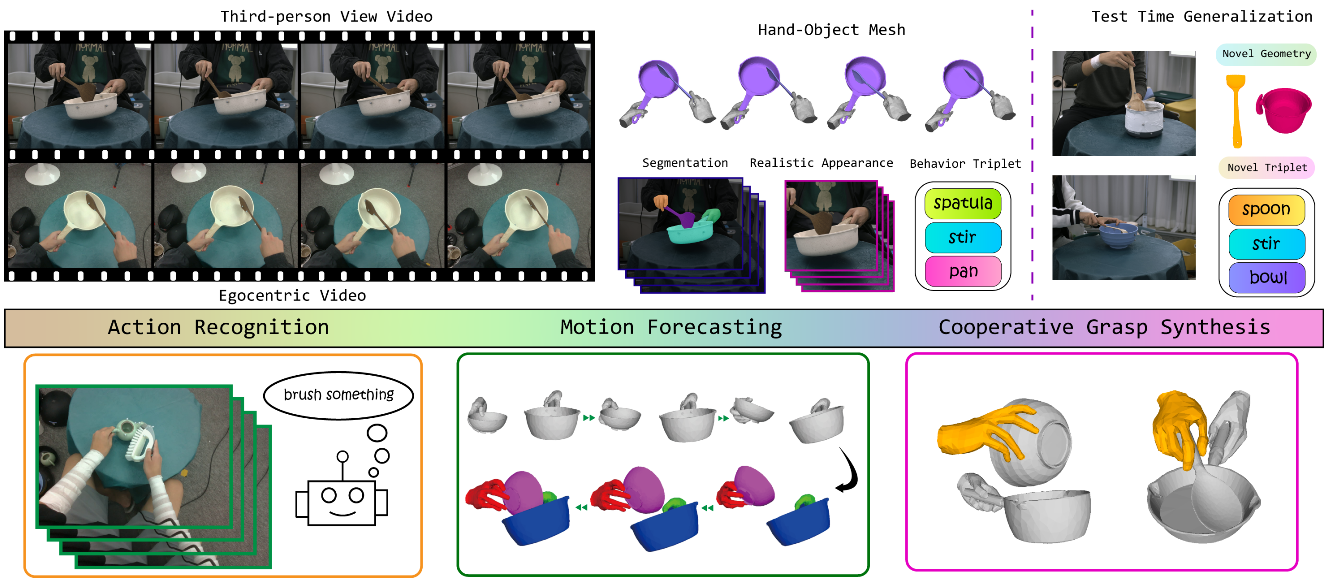

Figure 1. TACO is a large-scale bimanual hand-object manipulation dataset covering extensive tool-action-object combinations in real-world scenarios. It involves videos from both 12 third-person views and one egocentric view, together with precise hand-object meshes, 2D segmentation, realistic hand-object appearances, and behavior triplet annotations. With rich diversities of object shapes and interaction behaviors, TACO supports test-time generalization to unseen object geometries and novel behavior triplets and benchmarks various generalizable research topics, *e.g.*, action recognition, motion forecasting, and cooperative grasp synthesis.

Abstract

Humans commonly work with multiple objects in daily life and can intuitively transfer manipulation skills to novel objects by understanding object functional regularities. However, existing technical approaches for analyzing and synthesizing hand-object manipulation are mostly limited to handling a single hand and object due to the lack of data support. To address this, we construct TACO, an extensive bimanual hand-object-interaction dataset spanning a large variety of tool-action-object compositions for daily human activities. TACO contains 2.5K motion sequences paired with third-person and egocentric views, precise hand-object 3D meshes, and action labels. To rapidly expand the data scale, we present a fully automatic data acquisition pipeline combining multi-view sensing with an optical motion capture system. With the vast research fields provided by TACO, we benchmark three generalizable hand-object-interaction tasks: compositional action recognition, generalizable hand-object motion forecasting, and cooperative grasp synthesis. Extensive experiments reveal new insights, challenges, and opportunities for advancing the studies of generalizable hand-object motion analysis and synthesis. Our data and code are available at `https://taco2024.github.io`.

[†]Corresponding author.

1. Introduction

In our everyday lives, humans effortlessly synchronize the movements of both hands to manipulate a pair of objects, such as using a spatula to stir in a pan while cooking. Usually, these objects are not symmetric in their roles, with one acting as a tool to enhance the bimanual action performed on the other object. This allows us to characterize the bimanual behaviors with tool-action-object triplets. Various triplet combinations usually incorporate intricate and differing temporal and spatial coordination among the tool, the object, and the two hands. Understanding versatile bimanual behaviors could immediately benefit numerous applications in VR/AR, human-robot interaction [8, 44], and dexterous manipulation [7, 49, 74, 78], which pose significant challenges for today's computer vision systems.

Tackling the above challenges requires the support of large-scale annotation-rich datasets. Existing dataset efforts [2, 4, 15, 18, 26, 39, 72] on hand-object interactions (HOI) primarily focus on unimanual actions. However, simply aggregating two unimanual actions falls short of encompassing bimanual coordination behaviors. Furthermore, some works [13, 19, 34, 66, 75] examine bimanual manipulation of a single object which can hardly extend to interpreting tool-action-object triplets. The most relevant dataset [33] currently available covers only 12 objects and fewer than 20 distinct tool-action-object triplets, severely limiting its potential to facilitate and benchmark the understanding of bimanual tool-action-object interactions that can generalize to novel object geometries or previously unseen triplets. This limitation partially stems from the challenging nature of jointly capturing two dynamic hands interacting with two objects in the real world, but further underscores the urgency of developing methodologies for generalizable bimanual HOI understanding. Therefore, we put generalization as our primary focus while exploring the curation of a large-scale bimanual HOI dataset that can both facilitate and benchmark generalizable understanding.

In particular, we present **TACO**, a large-scale bimanual manipulation dataset encompassing a diverse array of tool-action-object compositions in real-world settings. Serving as a knowledge base of multi-object cooperation, TACO focuses on daily scenarios involving the use of tools to interact with target objects and collects 131 types of <tool category, action label, target object category> triplets across 20 object categories, 196 object instances, and 15 daily actions. We organize the dataset based on the triplets and make sure triplets have different levels of overlaps, this naturally defines the semantic distance between different motion trajectories and supports studying generalization with different extents. To expedite the expansion of our dataset, we combine the advantages of marker-based and markerless motion capture systems and present a fully automatic data acquisition pipeline that can guarantee motion quality and visual data quality at the same time. For each time step, the pipeline automatically provides labels including the precise recovery of hand-object mesh and segmentation on markerless vision data. As a result, TACO comprises a total of 2.5K motion sequences and 5.2M video frames incorporating 3rd-person and egocentric views.

Benefiting from the rich and diverse motion sequences within our dataset, we carefully benchmark three tasks aiming at generalizing common human manipulation knowledge to unseen object geometries and novel tool-action-object combinations: **1)** compositional action recognition, **2)** generalizable hand-object motion forecasting, and **3)** cooperative grasp synthesis. Extensive studies are conducted on these benchmarks: First, diverse hand-object interactive motions with action annotations pave the way to learning to recognize hand-object action under novel object combinations in real-world scenarios. Additionally, the availability of 4D dynamic hand-object pose sequences allows for fine-grained hand-object motion tendency forecasting. Finally, we provide contact-rich hand-object meshes to facilitate studying hand grasp synthesis for unseen object geometries and categories under interaction scenarios. Extensive experiments have been conducted on these benchmarks, revealing generalization challenges faced by existing methodologies when dealing with novel object geometries, object categories, and interaction triplets. We anticipate that TACO will offer more opportunities for researching and developing universal generalization strategies in the understanding and creation of hand-object interactions.

In summary, our main contributions are threefold:

- We present TACO, the first large-scale real-world 4D bimanual hand-object-interaction dataset covering diverse tool-action-object compositions and object geometries.
- We develop an automatic data acquisition pipeline that provides precise recovery of hand-object mesh and segmentation together with markerless vision data.
- We benchmark three tasks toward generalizable hand-object motion analysis and synthesis. We provide a comprehensive discussion and highlight the new challenges posed by TACO.

2. Related Work

2.1. Hand-object Interaction Datasets

Understanding hand-object interaction is an emerging research topic that is supported by a large number of datasets. Many widely-used video datasets [11, 12, 17, 55] capture 2D real-world hand-object interaction videos to facilitate traditional visual perception studies such as action recognition and HOI detection. With the rapid development of 3D computer vision, recent advances have incorporated 3D hand-object mesh annotations into data-capturing processes and presented various 3D [20, 26] or

4D [2, 4, 12, 13, 15, 18, 19, 34, 39, 61, 66, 72, 75] datasets with different focuses. Table 1 briefly summarizes the characteristics of these datasets. A majority of existing datasets focus on single-hand grasping, a basic manipulation skill that can benefit existing dexterous robot learning [7, 48, 49, 78]. Beyond grasping, a line of datasets captures diverse and complex manipulation skills requiring humans to perform complicated, multi-step actions like using tools [13, 33, 39], deforming soft bodies [66], and multi-person handovers [75]. From the perspective of object behaviors, KIT [33], HOI4D [39], and AffordPose[26] capture functional manipulations aiming at showing actual functions and usages of objects. KIT further collects multi-object interaction behaviors, however, it lacks object diversities and is hard to support generalizable studies. Our dataset captures diverse bimanual functional manipulation behaviors for multi-object cooperation, serving a wide range of research directions.

2.2. Generalizable Hand-object Motion Analysis

Generalizable hand-object-interaction motion analysis aims to understand the attributes of hand-object motion when manipulating novel object instances. To precisely recognize hand actions during the manipulation of unseen objects, a series of compositional action recognition approaches [21, 31, 43, 50, 51, 59, 69, 76] represent objects through their bounding boxes to prevent the network from overfitting to intricate object geometries. To detect and infer action-object pairs under few-shot or zero-shot settings, a prevailing methodology used in recent generalizable HOI detection methods [22–24, 64, 65, 80] decomposes actions and objects into distinct features. This decomposition allows leveraging abundant training data with either similar actions or similar objects to handle rare action-object pairs that emerge at test time. It also provides the potential to discover novel and reasonable action-object pairs [25]. To estimate hand and object poses without object geometries, Zhu *et al.* [82] propose learning hand-object interaction priors for each object category and transferring them to unseen objects within the same category during test time, enhancing the method adaptability to diverse unseen object instances.

2.3. Generalizable Hand-object Motion Synthesis

Synthesizing realistic hand-object-interaction motion on novel object geometries and categories is an emerging research topic that includes two main challenges: how to model generic, realistic, and diverse hand motion patterns from existing HOI data, and how to apply them to novel object instances. To address the first challenge, a branch of data-driven approaches decomposes complicated hand-object contact into spatial relations between each *hand joint* [30, 37, 38, 79] or *hand surface point* [71, 77, 81] and the object's relevant closest local regions, and such relations can be formulated as a combination of whether contact occurs [37, 38, 71, 77, 79, 81], hand-object distances [37, 71, 77, 81], positions [37, 38, 77, 79, 81] and directions [37, 38, 79] of contact points, and whether contact points should be closer [71]. Another line of work [28–30, 35] encodes hand-object interaction integrally to an implicit global feature by a conditional variational autoencoder (CVAE) [57] structure. As solutions to the second challenge, generalizable methods [28, 30, 35, 38, 79] decompose the object's geometry from HOI and encode it as a condition for generative models. CAMS [79] further maps objects into a canonicalized object space to decrease the gap in object geometry. One limitation of existing approaches is that they do not focus on multi-hand and multi-object cooperation due to the lack of data support.

3. Constructing TACO

In this section, we respectively describe the data capturing system (Section 3.1), the data annotating pipeline (Section 3.2), and dataset statistics (Section 3.3).

3.1. Data Capturing

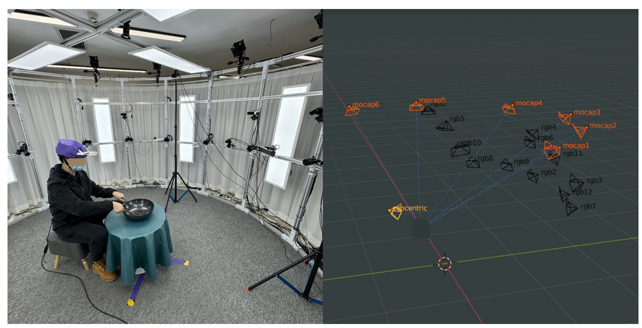

Figure 2. Data capturing system and camera views.

We obtain hand motion from multi-view RGB videos while capturing the object motion by attaching four markers on the object surface and tracking them with a mocap system. To this end, as shown in Figure 2, our data capturing system incorporates 12 synchronized industrial FLIR cameras and a NOKOV optical motion capture suite with 6 infrared Mars4H cameras. To capture egocentric RGBD videos, a helmet equipped with a Realsense L515 camera is worn by the actor. Similar to objects, the motion of the L515 camera is tracked by the mocap system. The frequency of our cameras and mocap system is 30 Hz. The resolutions of our allocentric and egocentric images are 4096x3000 and 1920x1080, respectively.

Object model acquisition. To capture contacts between hands and objects and support relevant studies, we obtain an accurate 3D mesh for each object with an industrial EinScan 3D scanner. Each object mesh is represented by up to 100K triangular faces for fine-grained geometries.

dataset	data characteristics: bimanual	multi-object cooperation	functional manipulation	category-level	egocentric	multi-view	markerless	mocap	action label	# number of: sequence	frame
FHPA [15]					✓			✓	✓	1.2K	105K
ObMan [20]										-	154K
ContactPose[2]						✓		✓		2.3K	3.0M
GRAB [61]	✓							✓		1.3K	1.6M
HO3D [18]						✓	✓			27	78K
KIT Binamual Manipulation [33]	✓	✓	✓		✓	✓		✓	✓	588	1.6M
H2O [34]	✓				✓	✓	✓			191	571K
DexYCB [4]						✓	✓			1.0K	582K
H2O-3D [19]	✓					✓	✓			17	76K
H2O for handover [75]	✓					✓		✓		6.0K	5.0M
OakInk [72]						✓		✓		778	314K
HOI4D [39]			✓	✓	✓		✓		✓	4.0K	1.2M
HMDO [66]	✓					✓	✓			12	21K
ARCTIC [13]	✓				✓	✓		✓		339	2.1M
AffordPose [26]			✓	✓					✓	-	27K
TACO (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	2.5K	5.2M

Table 1. Comparison of TACO with existing 3D hand-object interaction datasets.

3.2. Data Annotating

Figure 3 depicts our automatic data acquisition pipeline. Given color images, 3D marker positions, and object models as inputs, the pipeline successively performs object pose optimization, hand keypoint localization, hand pose optimization, hand-object segmentation, and marker removal.

Object pose optimization. Since our objects are rigid bodies, their motions are 6-dimensional and can be tracked by capturing marker positions using the mocap system. To accurately associate markers with the object, we draw inspiration from Mosh++ [41] and optimize marker-to-surface correspondence to encourage contact between markers and object surfaces while avoiding penetration. The object poses are obtained by combining marker positions relative to the object mesh and the captured marker motions.

3D hand keypoint localization. Leveraging multi-view sensing, we first estimate 2D hand keypoints from each camera view, respectively, and then fuse them into 3D. For 2D hand keypoint estimation, directly applying existing dual-hand approaches [19, 27, 36] could fail when severe occlusion occurs or two hands are far apart. We thus follow [3, 40] and design a method to separately detect two hands using pre-trained YOLOv3 [52] and then utilize MMPose [9] to obtain single-hand 2D keypoints. To fuse multi-view 2D keypoints to 3D under severe occlusion, we use RANSAC [14] to filter out imprecise 2D keypoints and localize 3D keypoints by triangulation.

Contact-aware hand pose optimization. We adopt MANO [53] to formulate a 3D hand mesh as $\Theta_h = \{\theta, \beta, t\}$, where $\theta \in \mathbb{R}^{48}$, $\beta \in \mathbb{R}^{10}$, and $t \in \mathbb{R}^3$ represent hand pose, hand shape, and wrist position, respectively. Given precomputed 3D hand keypoints and object meshes, we optimize MANO hand meshes by minimizing the following loss function:

$$\hat{\Theta}_h = \arg\min_{\Theta_h} (\lambda_{2D}\mathcal{L}_{2D} + \lambda_{3D}\mathcal{L}_{3D} + \lambda_{angle}\mathcal{L}_{angle} + \lambda_{tc}\mathcal{L}_{tc} + \lambda_p\mathcal{L}_p + \lambda_a\mathcal{L}_a)\,, \quad (1)$$

where $\mathcal{L}_{2D}$ and $\mathcal{L}_{3D}$ induce hand meshes to get close to the keypoints, $\mathcal{L}_{angle}$ provides hand pose priors, $\mathcal{L}_{tc}$ promotes temporal smoothness, $\mathcal{L}_a$ encourages contacts, and $\mathcal{L}_p$ prevents hand-object interpenetration. Details of these loss terms are provided in the supplementary material.

Hand-object segmentation. We obtain 2D hand-object masks by first rendering hand-object silhouettes from their 3D meshes. Since MANO meshes differ from real hands, we then apply the Segment Anything Model [32] to refine the masks, using sampled hand-object keypoints as prompts. To help detect hands in subsequent frames, we use Track-anything [70] to track the hands in these frames and localize them.

Marker removal. Despite the fact that attaching markers to objects can accurately capture object motion, markers may increase the appearance gap between captured objects and those in the wild. One direct defect is that an object pose tracking network could simply track markers rather than learn to track the target object. To improve the object's appearance and make it more realistic, after acquiring hand and object poses, we automatically remove the markers from RGB images. First, we obtain marker positions in the world coordinate system from the mocap system, and place spheres with a radius two times the marker at those positions in a simulation environment. We then render a 2D mask of the spheres for each real camera view. Benefiting from the accurate marker position estimates, such 2D masks can fully cover the markers in the original RGB images. Finally, we mask the spheres in the original RGB images and use a pre-trained LAMA [60] network to inpaint the sphere regions. Figure 4 shows a marker removal example.

3.3. Dataset Statistics

TACO contains 2.5K motion sequences describing extensive daily tool-using behaviors covering 20 object categories, 196 fine-grained object 3D models, 14 participants, and 15 actions. To support cross-category generalizable HOI studies, we explore the functional diversity of tools

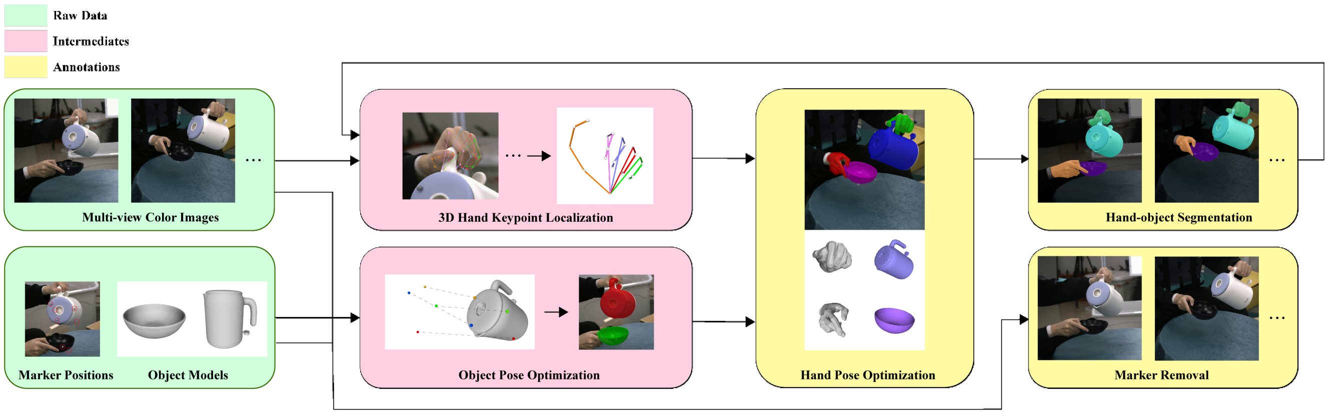

Figure 3. Automatic data annotating pipeline. The input consists of color frames from allocentric views, pre-scanned object models, and 3D positions of markers attached to object surfaces. We first separately localize 3D hand keypoints and obtain object poses, and then conduct contact-aware optimization to recover MANO [53] hand meshes. We finally segment hands and objects from images and automatically inpaint markers to acquire realistic object appearances.

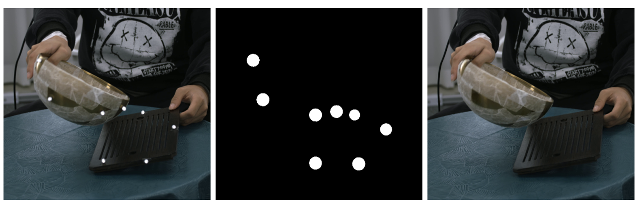

Figure 4. An example of marker removal. Three sub-figures respectively show the captured image patch, the automatically-computed marker mask, and the inpainted image patch.

and target objects, with each tool and target object being utilized to perform up to 7 different kinds of actions. We formulate a behavior as using a **tool category** to perform an **action** on a **target object category** and divide our data into 131 <tool category, action label, target object category> triplets. Some correlated triplets are exemplified in Figure 5. To enable various studies, TACO presents high-resolution color images from 12 allocentric views and RGBD images from one **egocentric-view**.

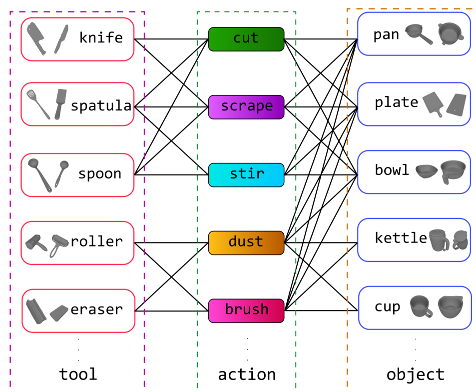

Figure 5. Examples of interaction triplets in TACO. The left, middle, and right columns exemplify categories of tool, action, and target object, respectively. Triplets in our dataset are represented by connected paths from tools to target objects.

4. Data Quality Evaluation

Qualitative contact optimization evaluation. Figure 6 depicts two frames of the TACO dataset. As shown in the figure, the attraction loss $\mathcal{L}_a$ promotes contact between the hands and the objects but may exacerbate penetration. The penetration loss $\mathcal{L}_p$ can mitigate penetration but does not address cases where the hands and the objects are not in contact. Our method combines both attraction loss and penetration loss, striking a balance between encouraging contact and preventing penetration.

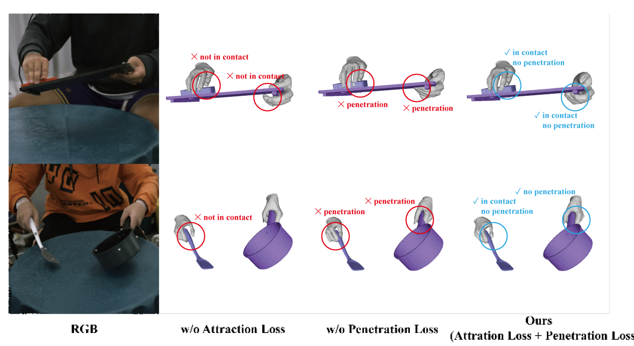

Figure 6. Qualitative contact optimization evaluation. From left to right: original RGB image, optimization without attraction loss, optimization without penetration loss, our method with attraction loss and penetration loss.

Quantitative hand pose evaluation. To quantitatively examine the quality and value of our hand poses, we conduct a cross-dataset evaluation between TACO and an existing high-quality HOI dataset DexYCB [4]. We select a multi-stage method, CMR [5], and a lightweight network, MobRecon [6], both of which can recover occluded hand

Method	Train / Test	TACO	DexYCB	TACO + DexYCB
CMR	TACO	10.73 / 30.87	13.53 / 39.54	**10.25 / 28.36**
	DexYCB	19.36 / 76.20	6.51 / 13.92	**6.44 / 13.62**
MobRecon	TACO	9.03 / 20.76	12.95 / 34.32	**8.69 / 20.14**
	DexYCB	15.17 / 44.73	6.51 / 13.61	**6.32 / 12.94**

Table 2. Cross-dataset evaluation with DexYCB [4] on 3D hand pose estimation. Results are in PA-MPJPE (mm, lower is better) and MPJPE (mm, lower is better), respectively.

meshes from a single-view RGB image captured in HOI scenarios, and then train and test their networks under different data combinations. For DexYCB, we follow its default train/test split (S0). For TACO, we randomly select 60 motion sequences for training and 40 for testing and crop the right hand from each image to align with the single-hand HO3D. Following existing hand mesh recovery methods [5, 6, 45, 67], we select Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE) and Mean Per Joint Position Error (MPJPE) as evaluation metrics. More details about the metrics can be found in the paper [6]. Table 2 shows the performance of the two methods under different settings. Due to the large gap in hand-object motion between the two datasets, the performances for the two methods would both drop rapidly when the training and the test sets are from different datasets. Nevertheless, combining training data from TACO and DexYCB achieves significant performance gains on both datasets for both the two methods, which demonstrates the accuracy of our hand pose annotations and indicates that TACO is a beneficial data complement for 3D hand pose estimation.

Marker removal quality evaluation. We conduct a novel experiment to examine the reality of our marker-removed images. The key idea is that a network can easily segment markers from the image while being hard to figure out the same regions when markers are replaced by the original object surface appearances. We thus use the originally captured image patches and the marker-removed ones to train a 2D marker segmentation network, respectively, and then apply the two networks to novel image patches with the same image processing setting as training. In practice, spheres larger than markers (as shown in Figure 4(b)) are required to be segmented, since the marker removal process exactly inpaints these regions. We select U-Net [54] as the network backbone and train the two networks with an L2-loss measuring differences between mask predictions and ground truths. When training and testing on raw image patches, U-Net achieves 63.8% mean Intersection over Union (mIoU), while its performance sharply drops to 11.1% mIoU after applying marker removal, indicating the reality and naturalness of the inpainted image regions. Figure 7 further compares heatmaps of network predictions under the two settings, exhibiting ambiguities of marker segmentation posed by realistic inpainted images. More details are provided in the supplementary material.

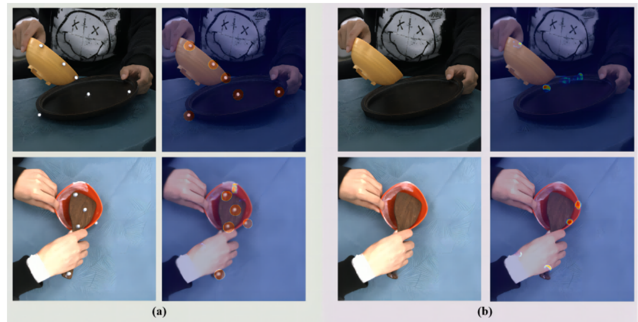

Figure 7. Marker removal quality evaluation. We train and test U-Net [54] on raw captured images (a) and marker-removed images (b), respectively. In each experiment, the left and right columns show the input image patches and the predicted segmentation heatmaps, respectively.

5. Experiments

In this section, we first introduce our data split (Section 5.1) supporting different research purposes about generalization. We then present three benchmarks on TACO: compositional action recognition (Section 5.2), generalizable hand-object motion forecasting (Section 5.3), and cooperative grasp synthesis (Section 5.4). Details of evaluation metrics are provided in supplementary materials.

5.1. Data Split

With diverse object geometries and interaction triplets, TACO focuses on supporting generalizable studies that apply to unseen object geometries or novel interaction triplets. Hence, we carefully split our test data into four subsets with different generalization purposes:

- Test set 1 (S1): *No generalization.* Tool geometries and interaction triplets are both involved in the training set.
- Test set 2 (S2): *Geometry-level generalization.* The tool geometry is novel, while the interaction triplet appears in the training set.
- Test set 3 (S3): *Interaction-level generalization.* The interaction triplet is novel, while the tool categories and geometries are included in the training set.
- Test set 4 (S4): *Compound generalization.* The tool category is novel, leading to unseen geometries and triplets.

The training set and the four test sets follow a data ratio of 4:1:1:1.5:2.5.

5.2. Compositional Action Recognition

Humans can recognize the action disentangled from their context and appearance biases of the objects, which is referred to as compositionality. To test a system's compositional generalization capability, with a diverse array of tool-action-object triplets in TACO, we benchmark compositional action recognition aiming at identifying actions

Test Set	Method	Top-1 (%, ↑)	Top-5 (%, ↑)
S1	AIM [73]	83.08	98.85
	CACNF [50]	**86.15**	**99.62**
S2	AIM [73]	**82.81**	**98.83**
	CACNF [50]	77.34	96.88
S3	AIM [73]	53.65	81.25
	CACNF [50]	**63.02**	**92.97**
S4	AIM [73]	39.33	73.67
	CACNF [50]	**44.00**	**80.50**

Table 3. Results on compositional action recognition. Methods are examined via Top-1 and Top-5 accuracy.

in HOI videos. In contrast to traditional action recognition tasks [11, 16, 56, 58], our focus is on scenarios where the tool-action-object triplets in the test set are unseen during training, including novel tool types or geometries. This poses additional challenges for evaluating model generalization and achieving human-like perception.

Problem formulation: In a bimanual HOI scenario, our goal is to identify action labels from first-person RGB video frames, with optional bounding boxes for hands and objects computed based on their poses.

Evaluation metrics: Following [50, 73], we use Top-1 Accuracy and Top-5 Accuracy to evaluate the efficacy of action recognition.

Baselines, results, and analysis: We compare two baselines to evaluate the compositional generalization capability: AIM [73], an adaptation of pre-trained image transformer models, representing state-of-the-art traditional action recognition approaches; and CACNF [50], which specializes in compositional and few-shot action recognition. Table 3 shows the results. A significant decrease in accuracy on the most difficult S4 set, compared to high accuracy on the non-generalizable S1 set, underscores compositional generalization challenges. Notably, the S2 set, with novel tool geometries within tool categories, performed similarly to S1, while the S3 set, featuring novel tool-action-object triplets, showed a significant accuracy drop, revealing that while action recognition models inherently possess basic generalizability to geometric changes, they struggle with novel triplets due to the unfamiliarity of interaction contexts. When comparing among baselines, CACNF outperformed AIM in challenging sets S3 and S4, benefiting from its bounding box-based approach that effectively disentangles actions from object and tool geometries, highlighting the advantages of a compositional approach in generalizable cases. All these results underscore the significant impact of compositional generalization on action recognition.

5.3. Generalizable Hand-object Motion Forecasting

With abundant 3D hand-object mesh annotations provided by TACO, we benchmark generalizable interaction forecasting aiming to predict the following hand-object motions from seen short motion clips. Unlike the prediction of low-frequency human motions [42, 46, 47], we observe that hands commonly achieve a complete manipulation behavior (e.g. pouring all the water out of a bowl) in a very short time, making forecasting both interesting and challenging.

Problem formulation: In a bimanual HOI scenario, given object point clouds and poses of two hands, the tool, and the target object in consecutive N frames, our goal is to forecast their poses in subsequent M frames. We select N=10 and M=10 in our dataset.

Evaluation metrics: Following human-object forecasting evaluations [1, 10, 63, 68], we use Mean Per Joint Position Error J_e to evaluate predicted left and right-hand skeletons, and leverage translation error T_e and rotation error R_e to measure object positions and orientations, respectively.

Baselines, results, and analysis: Due to the lack of hand-object interaction forecasting solutions, we transfer methodologies from human-object motion forecasting as baselines. To cover diverse method designs and possibilities, we select two generative models [62, 68] and two predictive models [10, 68] for comparison. Table 4 compares their performance under different generalization settings. We observed that the tool and the hand holding it play dominant roles during manipulation, making their motion forecasting more challenging compared to others. The generative models, in comparison to predictive ones, yield significantly larger errors for both hands and the tool. One possible reason is that their motion patterns are fast and complex, making the modeling of motion distribution challenging. Enhancing the accuracy of generative methods in hand-object manipulation scenarios is an interesting future direction. When comparing method performances between a non-generalizable set (S1) and sets involving generalization (S2-4), we note that methods consistently exhibit significant performance declines with the right hand and the tool, regardless of whether applied to novel tool geometries (S2, S4) or interaction triplets (S3, S4), meanwhile achieving close results on the others. This trend indicates substantial challenges in generalizing dominant interaction entities.

5.4. Cooperative Grasp Synthesis

Generating human-like hand grasps benefits various applications in VR/AR and dexterous manipulation. Existing grasp synthesis approaches [28, 30, 38, 81] mainly focus on creating stable grasps on static objects without considering interactive purposes. Taking advantage of extensive interaction behaviors from TACO, we benchmark a novel task that aims to synthesize realistic and physically plausible hand grasps in HOI scenarios. Besides achieving stable grasping for the object, the method must comprehend other interactive objects and human hands to generate cooperative motions and avoid conflicts. We evaluate this task on our four test sets, each targeting different aspects of generalization.

Problem formulation: Consider using the right hand to

Test Set	Method	J_e (mm, ↓) Right / Left	T_e (mm, ↓) Tool / Target	R_e (°, ↓) Tool / Target
S1	InterVAE [68]	54.9 / 48.9	55.0 / 15.8	66.94 / 6.63
	MDM [62]	61.2 / 53.7	49.4 / 14.9	52.81 / 5.85
	InterRNN [68]	29.3 / **20.4**	25.8 / **11.6**	**10.08** / 5.22
	CAHMP [10]	**28.8** / 22.1	**23.9** / 12.3	10.24 / **4.95**
S2	InterVAE [68]	58.8 / 50.8	54.3 / 12.2	88.19 / 5.24
	MDM [62]	65.8 / 55.8	48.0 / 10.1	75.78 / 3.59
	InterRNN [68]	35.2 / 23.1	31.0 / 8.9	11.09 / 4.03
	CAHMP [10]	**31.4 / 22.6**	**24.7 / 8.6**	**10.65 / 3.32**
S3	InterVAE [68]	56.5 / 50.1	56.7 / 12.7	68.03 / 5.69
	MDM [62]	66.4 / 58.0	46.2 / 11.5	61.90 / 4.14
	InterRNN [68]	32.5 / 22.0	28.2 / 9.7	11.42 / 4.32
	CAHMP [10]	**29.3 / 21.6**	**24.4 / 9.4**	**11.18 / 3.80**
S4	InterVAE [68]	63.9 / 54.2	63.0 / 14.5	71.38 / 5.01
	MDM [62]	70.5 / 57.9	50.2 / 12.2	70.25 / 3.55
	InterRNN [68]	36.6 / 22.3	30.9 / 10.6	11.92 / 3.86
	CAHMP [10]	**32.9 / 22.2**	**26.6 / 10.8**	**11.73 / 3.24**

Table 4. Results on generalizable interaction forecasting. We respectively examine predictions of the right hand, the left hand, the tool, and the target object.

manipulate a tool and the left hand to operate a target object. During manipulation, given the meshes of the two objects and the left hand, the goal is to generate a right-hand mesh grasping the tool in a manner conducive to the interaction.

Evaluation metrics: Following static grasp synthesis methods [28, 29, 38], we use interpenetration volume (*Pen. V*) and contact ratio (*Con. R*) to measure the physical plausibility of generated hands. To ensure the hand is appropriately interactive, we introduce a new metric named collision ratio (*Col. R*), which detects collisions between generated hands and the interaction environment. Nevertheless, to quantitatively examine whether the grasps are realistic, we design the FID score (*FID*) to measure distances between distributions of human grasps and synthesized ones.

Baselines, results, and analysis: We select two recent CVAE-based approaches ContactGen [38] and HALO-VAE [30], and modify their network structures to incorporate the left hand and the target object as additional VAE conditions. To assess the significance of interaction environments, we also evaluate HALO-VAE^{-}, which omits the use of environment inputs and directly generates from the tool mesh. Due to the strong correlation between hand grasps and tool geometries, we test these approaches with familiar (S1, S3) and unseen tool geometries (S2, S4). Table 5 summarizes quantitative evaluations of the two settings. Notably, HALO-VAE^{-} shows a significant decrease in performance on contact and collision ratios compared to HALO-VAE, underscoring the importance of perceiving and modeling interaction environments. When applied to novel object geometries and categories, all three approaches exhibit higher collision ratios and FID scores, and most of them show a marked decline in their ability to contact tools precisely and prevent penetrations. Figure 8 exemplifies

Test Set	Method	*Pen. V* (cm^3, ↓)	*Con. R* (%, ↑)	*Col. R* (%, ↓)	*FID* (10^{-2}, ↓)
S1 ⋃ S3	ContactGen [38]	0.27	80.12	10.40	12.18
	HALO-VAE^{-}	0.31	76.27	4.19	4.89
	HALO-VAE [30]	**0.26**	**82.86**	**2.39**	**2.02**
S2 ⋃ S4	ContactGen [38]	**0.26**	74.98	17.56	22.53
	HALO-VAE^{-}	0.29	71.69	14.25	5.09
	HALO-VAE [30]	0.32	**83.52**	**11.11**	**2.84**

Table 5. Results on cooperative grasp synthesis. We measure the physical plausibility of generated grasps by interpenetration volume and contact ratio, meanwhile using collision ratio to examine interaction faithfulness and FID score to assess reality.

failure cases of synthesized grasps, indicating a large room for improving physical feasibility and naturalness on complex and slender geometries.

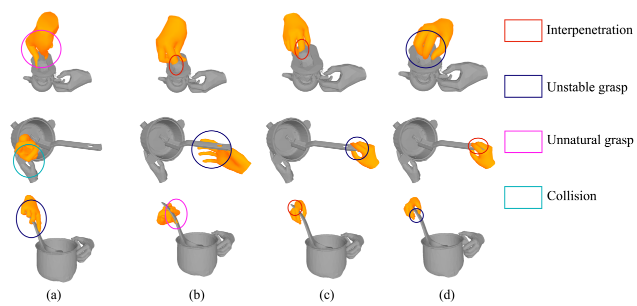

Figure 8. Failure cases of cooperative grasp synthesis. (a) and (b) are generated by ContactGen [38], while (c) and (d) are from HALO-VAE [30].

6. Limitations and Conclusion

We present TACO, the first large-scale, real-world 4D bimanual hand-object manipulation dataset. It encompasses a wide range of tool-action-object compositions and object geometries. TACO contains a total of 2.5K motion sequences and 5.2M video frames, captured from 12 third-person views and one egocentric view. We contribute an automatic data acquisition pipeline that accurately recovers hand-object meshes and segmentation, along with realistic hand-object appearances. Leveraging TACO's diverse data, we benchmark compositional action recognition, generalizable hand-object motion forecasting, and cooperative grasp synthesis. These benchmarks reveal new insights, challenges, and opportunities in the field of generalizable hand-object interaction studies.

There are three major limitations in TACO. Firstly, TACO currently does not cover articulated objects. Secondly, while TACO offers an extensive exploration of object geometries and HOI behaviors, it lacks scene diversities that are also crucial for understanding human manipulations. Thirdly, our solution to marker removal is an application of generative models, hence the original object appearances cannot be recovered perfectly.

References

[1] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13390–13400, 2021. 7

[2] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 361–378. Springer, 2020. 2, 3, 4

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 4

[4] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 2, 3, 4, 5, 6

[5] Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregationand adaptive 2d-1d registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5, 6

[6] Xingyu Chen, Yufeng Liu, Dong Yajiao, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5, 6

[7] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20577–20586, 2022. 2, 3

[8] Sammy Christen, Wei Yang, Claudia Pérez-D'Arpino, Otmar Hilliges, Dieter Fox, and Yu-Wei Chao. Learning human-to-robot handovers from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9654–9664, 2023. 2

[9] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020. 4

[10] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6992–7001, 2020. 7, 8

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 2, 7

[12] Christian RG Dreher, Mirko Wächter, and Tamim Asfour. Learning object-action relations from bimanual human demonstration using graph networks. *IEEE Robotics and Automation Letters*, 5(1):187–194, 2019. 2, 3

[13] 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 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 2, 3, 4

[14] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 4

[15] 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018. 2, 3, 4

[16] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 7

[17] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[18] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 2, 3, 4

[19] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11090–11100, 2022. 2, 3, 4

[20] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 2, 4

[21] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3148–3159, 2022. 3

[22] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 584–600. Springer, 2020. 3

[23] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504, 2021.

[24] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14646–14655, 2021. 3

[25] Zhi Hou, Baosheng Yu, and Dacheng Tao. Discovering human-object interaction concepts via self-compositional learning. In *European Conference on Computer Vision*, pages 461–478. Springer, 2022. 3

[26] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14713–14724, 2023. 2, 3, 4

[27] Changlong Jiang, Yang Xiao, Cunlin Wu, Mingyang Zhang, Jinghong Zheng, Zhiguo Cao, and Joey Tianyi Zhou. A2j-transformer: Anchor-to-joint transformer network for 3d interacting hand pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8846–8855, 2023. 4

[28] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 3, 7, 8

[29] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 8

[30] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *2021 International Conference on 3D Vision (3DV)*, pages 11–21. IEEE, 2021. 3, 7, 8

[31] Tae Soo Kim, Jonathan Jones, and Gregory D Hager. Motion guided attention fusion to recognize interactions from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13076–13086, 2021. 3

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 4

[33] Franziska Krebs, Andre Meixner, Isabel Patzer, and Tamim Asfour. The kit bimanual manipulation dataset. In *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, pages 499–506. IEEE, 2021. 2, 3, 4

[34] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021. 2, 3, 4

[35] Haoming Li, Xinzhuo Lin, Yang Zhou, Xiang Li, Yuchi Huo, Jiming Chen, and Qi Ye. Contact2grasp: 3d grasp synthesis via hand-object contact constraint. *arXiv preprint arXiv:2210.09245*, 2022. 3

[36] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2761–2770, 2022. 4

[37] Qingtao Liu, Yu Cui, Zhengnan Sun, Haoming Li, Gaofeng Li, Lin Shao, Jiming Chen, and Qi Ye. Dexrepnet: Learning dexterous robotic grasping network with geometric and spatial hand-object representations. *arXiv preprint arXiv:2303.09806*, 2023. 3

[38] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20609–20620, 2023. 3, 7, 8

[39] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 2, 3, 4

[40] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 4

[41] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 4

[42] Wei Mao, Richard I Hartley, Mathieu Salzmann, et al. Contact-aware human motion forecasting. *Advances in Neural Information Processing Systems*, 35:7356–7367, 2022. 7

[43] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2020. 3

[44] Eley Ng, Ziang Liu, and Monroe Kennedy. It takes two: Learning to plan for human-robot cooperative carrying. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7526–7532. IEEE, 2023. 2

[45] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1496–1505, 2022. 6

[46] Xiaogang Peng, Yaodi Shen, Haoran Wang, Binling Nie, Yigang Wang, and Zizhao Wu. Somoformer: Social-aware motion transformer for multi-person motion prediction. *arXiv preprint arXiv:2208.09224*, 2022. 7

[47] Xiaogang Peng, Siyuan Mao, and Zizhao Wu. Trajectory-aware body interaction transformer for multi-person pose forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17121–17130, 2023. 7

[48] Yuzhe Qin, Hao Su, and Xiaolong Wang. From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation. *IEEE Robotics and Automation Letters*, 7(4):10873–10881, 2022. 3

[49] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022. 2, 3

[50] Gorjan Radevski, Marie-Francine Moens, and Tinne Tuytelaars. Revisiting spatio-temporal layouts for compositional action recognition. *arXiv preprint arXiv:2111.01936*, 2021. 3, 7

[51] Ramanathan Rajendiran, Debaditya Roy, and Basura Fernando. Modelling spatio-temporal interactions for compositional action recognition. *arXiv preprint arXiv:2305.02673*, 2023. 3

[52] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 4

[53] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 4, 5

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 6

[55] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 2

[56] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 7

[57] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 3

[58] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 7

[59] Pengzhan Sun, Bo Wu, Xunsong Li, Wen Li, Lixin Duan, and Chuang Gan. Counterfactual debiasing inference for compositional action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3220–3228, 2021. 3

[60] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021. 4

[61] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 3, 4

[62] 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. 7, 8

[63] Weilin Wan, Lei Yang, Lingjie Liu, Zhuoying Zhang, Ruixing Jia, Yi-King Choi, Jia Pan, Christian Theobalt, Taku Komura, and Wenping Wang. Learn to predict how humans manipulate large-sized objects from interactive motions. *IEEE Robotics and Automation Letters*, 7(2):4702–4709, 2022. 7

[64] Haozhong Wang, Hua Yu, and Qiang Zhang. Detecting zero-shot human-object interaction with visual-text modeling. In *2023 9th International Conference on Virtual Reality (ICVR)*, pages 155–162. IEEE, 2023. 3

[65] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2022. 3

[66] Wei Xie, Zhipeng Yu, Zimeng Zhao, Binghui Zuo, and Yangang Wang. Hmdo: Markerless multi-view hand manipulation capture with deformable objects. *Graphical Models*, 127:101178, 2023. 2, 3, 4

[67] Hao Xu, Tianyu Wang, Xiao Tang, and Chi-Wing Fu. H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17048–17058, 2023. 6

[68] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023. 7, 8

[69] Rui Yan, Peng Huang, Xiangbo Shu, Junhao Zhang, Yonghua Pan, and Jinhui Tang. Look less think more: Rethinking compositional action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3666–3675, 2022. 3

[70] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023. 4

[71] 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 *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 11097–11106, 2021. 3

[72] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20953–20962, 2022. 2, 3, 4

[73] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. AIM: Adapting image models for efficient video action recognition. In *The Eleventh International Conference on Learning Representations*, 2023. 7

[74] Jianglong Ye, Jiashun Wang, Binghao Huang, Yuzhe Qin, and Xiaolong Wang. Learning continuous grasping function with a dexterous hand from human demonstrations. *IEEE Robotics and Automation Letters*, 8(5):2882–2889, 2023. 2

[75] Ruolin Ye, Wenqiang Xu, Zhendong Xue, Tutian Tang, Yanfeng Wang, and Cewu Lu. H2o: A benchmark for visual human-human object handover analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15762–15771, 2021. 2, 3, 4

[76] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Is an object-centric video representation beneficial for transfer? In *Proceedings of the Asian Conference on Computer Vision*, pages 1976–1994, 2022. 3

[77] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)*, 40(4):1–14, 2021. 3

[78] 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. 2, 3

[79] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 585–594, 2023. 3

[80] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19568–19577, 2022. 3

[81] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 3, 7

[82] Zehao Zhu, Jiashun Wang, Yuzhe Qin, Deqing Sun, Varun Jampani, and Xiaolong Wang. Contactart: Learning 3d interaction priors for category-level articulated object and hand poses estimation. *arXiv preprint arXiv:2305.01618*, 2023. 3